UNIVERSIDAD AUTÓNOMA DE MANIZALES

MAESTRÍA EN GESTIÓN Y DESARROLLO DE PROYECTOS DE SOFTWARE

PREDICCIÓN DEL TIEMPO DE LA ESTANCIA HOSPITALARIA DE PACIENTES CON TRASTORNO AFECTIVO BIPOLAR EN LA CLÍNICA SAN JUAN DE DIOS DE MANIZALES POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS

ING. CRISTIAN DANIEL ZULUAGA VALENCIA

MANIZALES, CALDAS, COLOMBIA

MARZO 2017

PREDICCIÓN DEL TIEMPO DE LA ESTANCIA HOSPITALARIA DE PACIENTES CON TRASTORNO AFECTIVO BIPOLAR EN LA CLÍNICA SAN JUAN DE DIOS DE MANIZALES POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS

ING. CRISTIAN DANIEL ZULUAGA VALENCIA

Informe final

Trabajo de Grado para optar al título de Magister en Gestión y Desarrollo de Proyectos de software

Director de tesis

Ph.D. MARIA HELENA MEJÍA SALAZAR

UNIVERSIDAD AUTÓNOMA DE MANIZALES

MAESTRÍA EN GESTIÓN Y DESARROLLO DE PROYECTOS DE SOFTWARE

MANIZALES

2017

Tabla de contenido

Índice de figuras 7
Índice de tablas 9
RESUMEN 12
ABSTRACT 13
INTRODUCCIÓN 14
DESCRIPCIÓN DEL ÁREA PROBLEMÁTICA 15
ANTECEDENTES 17
FORMULACIÓN DEL PROBLEMA 23
JUSTIFICACIÓN 24
OBJETIVOS 26
OBJETIVO GENERAL: 26
OBJETIVOS ESPECÍFICOS: 26
ALCANCE Y LIMITACIONES 27
RESULTADOS ESPERADOS 28
REFERENTE TEÓRICO 29
REFERENTE TEMÁTICO 29
Trastorno Afectivo Bipolar (TAB). 29
Clasificación del Trastorno Afectivo Bipolar 29
Generalidades en el tratamiento del TAB 33
Tratamiento farmacológico 34
Tratamiento no farmacológico 34
REFERENTE METODOLÓGICO 35
Sociedad de la información 35
Sociedad del conocimiento 35
Sociedad del aprendizaje 35
Sistemas de información hospitalarios y conceptos relacionados 36
Estándares 36
Datos – Información - Conocimiento 37
Explotación de información y sistemas inteligentes 38
Analítica de datos 39
Minería de datos 39

Almacenes de datos 40
Tipos de datos. 41
Extracción, transformación y carga (ETL) 42
Metodologías de Minería de Datos 43
Modelo KDD 50
Taxonomía de las técnicas de minería de datos 52
Algunas técnicas de minería de datos 54
Regresión 54
Regresión logística 54
Análisis discriminante. 55
Reducción de la dimensionalidad de conjuntos de datos 58
Análisis de componentes principales 58
Análisis de curvas ROC (Receiver Operating Characteristic Curve) 59
ESTRATEGIA METODOLÓGICA 62
Tipo de estudio. 63
Población. 63
Variables incluidas en el estudio y fuentes de información. 63
Control del sesgo de selección. 63
Viabilidad y factibilidad. 64
Aspectos éticos 64
Selección de la metodología de minería de datos 65
Procesamiento de los datos 67
Plan de análisis de los datos 67
PRESUPUESTO 69
CRONOGRAMA 70
RESULTADOS 71
Procesamiento de datos 71
Descripción de las Hospitalizaciones entre 2013 y 2014 71
Reingresos hospitalarios 76
Descripción de los pacientes hospitalizados entre 2013 y 2014 77
Distribución porcentual del TAB 78
Análisis de la edad 80

Análisis del género 82
Contraste edad vs género 83
Análisis sociodemográfico 86
Análisis para los diagnósticos 107
Consideraciones para la construcción del modelo multivariado 110
establece la ciencia estadística. 110
ANÁLISIS DE COMPONENTES PRINCIPALES (ACP) 111
MODELOS MULTIVARIADOS 116
MODELO DE REGRESIÓN LOGÍSTICA 116
MODELO DE ANÁLISIS DISCRIMINANTE 119
COMPARACIÓN Y EVALUACIÓN DE LOS MODELOS MULTIVARIADOS. 122
DISCUSIÓN 124
CONCLUSIONES 128
RECOMENDACIONES 130
REFERENTE BIBLIOGRÁFICO 131
ANEXOS 136
ANEXO 1: definición operacional y valores permitidos o categorías de las variables 136
ANEXO 2: Resumen Ejecutivo presentado al comité de Bioética de la Clínica San Juan de Dios de Manizales 139
ANEXO 3: Autorización de comité de bioética para el uso de los datos clínicos 143

Índice de figuras

Figura 1. Causas de hospitalización de pacientes entre 2010 – 2015 en la Clínica San Juan de Dios de Manizales (CJDMS). Fuente: propia 25
Figura 2. Extracción de registros de una base de datos transaccional para su análisis. Fuente: [Romeu & Pardo, 2010]. 41
Figura 3. Proceso de carga de un almacén de datos. Fuente: [Romeu & Pardo. 2010]. ... 41
Figura 4. Extracción, transformación y carga. Fuente: http://superhotmobile.com/etl/etl-tools-extract-transform-load-information-builders.html 42
Figura 5. Encuesta realizada por la KDnuggest en el año 2007. Fuente: [Moine, Haedo & Gordillo, 2012]. 43
Figura 6. Fases de la metodología SEMMA. Fuente: [Britos, 2008] 44
Figura 7. Dinámica de la metodología SEMMA. Fuente: [Britos, 2008]. 45
Figura 8. Dinámica de la metodología P3TQ. Fuente: [Britos, 2008]. 47
Figura 9. Dinámica de la metodología CRISP-DM. Fuente: [Chapman et al., 2000]. 48
Figura 10. Etapas que componen el proceso KDD. Fuente: [Fayyad, 1997]. 51
Figura 11. Distribución de resultados de una población que presenta una enfermedad y una que no presenta. Fuente: [MedCalc, 2015]. 59
Figura 12. Gráfico ejemplo de curcas ROC. Fuente [Cerda & Cifuentes, 2012]. 62
Figura 13. Cantidad de hospitalizaciones por TAB registradas entre 2013 y 2014 en la CSJDM. Fuente: propia. 72
Figura 14. Diagrama de caja para el número de hospitalización para las hospitalizaciones registradas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 72
Figura 15. Distribución de las hospitalizaciones registradas entre 2013 y 2014 por TAB según el tipo de diagnóstico y el año de ocurrencia en la CSJDM. Fuente: propia. 73
Figura 16. Diagrama de caja y valores atípicos extremos de los días de estancia de las hospitalizaciones registradas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.. 75
Figura 17. Pacientes hospitalizados entre 2013 y 2014 por TAB según reingresos en la CSJDM. Fuente: propia. 76
Figura 18. Pacientes hospitalizados entre 2013 y 2014 por TAB, según reingresos y año de ocurrencia del reingreso en la CSJDM. Fuente: Propia. 76
Figura 19. Cantidad de pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia 78
Figura 20. Distribución espacial en el departamento de Caldas de la proporción de pacientes hospitalizados con TAB entre 2013 y 2014 en la CSJDM. Fuente: propia. 80
Figura 21. Distribución de la edad de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 81
Figura 22. Distribución espacial en el departamento de Caldas de los grupos de edad de pacientes hospitalizados entre 2013 y 2014 TAB en la CSJDM. Fuente: propia. 82
Figura 23. Distribución espacial en el departamento de Caldas del género de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 82

Figura 24. Distribución porcentual del género de los pacientes hospitalizados por TABentre 2013 y 2014 en la CSJDM. Fuente: propia. 83
Figura 25. Distribución de la edad del género femenino en pacientes hospitalizadas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 85
Figura 26. Distribución de la edad del género masculino en pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 85
Figura 27. Diagrama de caja para la distribución de la edad por género de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 86
Figura 28. Distribución del régimen de aseguramiento a la seguridad social de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 86
Figura 29. Distribución del estrato socioeconómico de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 87
Figura 30. Distribución del estado civil de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 87
Figura 31. Distribución del nivel educativo de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 87
Figura 32. Distribución de los pacientes hospitalizados entre 2013 y 2014 por TAB según el tipo de diagnóstico en la CSJDM. Fuente: propia. 99
Figura 33. Distribución de los pacientes hospitalizados entre 2013 y 2014 por TAB según el tipo de diagnóstico y género en la CSJDM. Fuente: propia. 100
Figura 34. Distribución espacial en el departamento de Caldas del tipo del diagnóstico de los pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia. 101
Figura 35. Diagrama de caja y valores atípicos extremos de los días de estancia de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 102
Figura 36. Distribución espacial en el departamento de Caldas de los días estancia de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 103
Figura 37. Distribución del egreso hospitalario de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 108
Figura 38. Nube de puntos de la correlación de los días estancia y la edad de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 111
Figura 39. Nube de puntos de la correlación de los días estancia y el estrato socioeconómico de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 111
Figura 40. Curvas ROC para la evaluación de los modelos predictivos del tiempo de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 123

Índice de tablas

Tabla 1. Porcentaje de investigaciones (de un total de 72) que satisfacen los criterios de calidad. Fuente: [Dreiseitl & Ohno, 2002]. 21
Tabla 2. Criterios de diagnósticos CIE-10 y DSM-IV para TAB. Fuente: [Grupo de Trabajo de la Guía de Práctica Clínica sobre Trastorno Bipolar, 2012] 33
Tabla 3. Tareas de cada fase de la metodología CRISP-DM. Fuente: [Britos. 2008]. 50
Tabla 4. Tabla de contingencia. Fuente: [MedCalc, 2015]. 60
Tabla 5. Fases del proceso de minería de datos en cada modelo. Fuente: [Moine, Haedo & Gordillo, 2012]. 66
Tabla 6. Conceptos de inteligencia de negocios, técnicas y procesos de explotación de información abarcados por las metodologías. Fuente: [Britos, 2008]. 66
Tabla 7. Costos del personal. Fuente: propia. 69
Tabla 8. Costos de adquisiciones. Fuente: propia 69
Tabla 9. Costos de servicios. Fuente: propia 69
Tabla 10. Costo total del proyecto. Fuente: propia 69
Tabla 11. Cronograma de actividades. Fuente: propia 70
Tabla 12. Cantidad de hospitalizaciones registradas entre 2013 y 2014 por TAB según el tipo de diagnóstico en la CSJDM. Fuente: propia. 74
Tabla 13. Estadísticos descriptivos de los días estancia de las hospitalizaciones registradas entre 2013 y 2014 por TAB en la CSDJM. Fuente: propia. 75
Tabla 14. Resumen estadístico para cantidad de pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia 78
Tabla 15. Distribución porcentual y concentración del TAB en el departamento de Caldas según pacientes hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 79
Tabla 16. Estadísticos descriptivos de la edad de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 81
Tabla 17. Estadísticos descriptivos de la edad en el género femenino de pacientes hospitalizadas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 84
Tabla 18. Estadísticos descriptivos de la edad en el género Masculino de pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 84
Tabla 19. Estadísticos descriptivos de los días estancia según régimen contributivo de los pacientes hospitalizadas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 88
Tabla 20. Estadísticos descriptivos de los días estancia según régimen subsidiado de pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 89
Tabla 21. Prueba de Mann Whitney para muestras independientes de los días estancia y régimen de aseguramiento a la seguridad social en salud de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 89
Tabla 22. Anova para días de estancia por nivel educativo para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 90

Tabla 23. Detalle de resultados del método de diferencia máxima significativa por nivel educativo para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia..... 96
Tabla 24. Detalle de resultados del método de diferencia mínima significativa por nivel educativo para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia..... 97
Tabla 25. Anova para días de estancia por estado civil para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 97
Tabla 26. Múltiples rangos por estdo civil pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia 98
Tabla 27. Estadísticos descriptivos de los días estancia en pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 102
Tabla 28. Estadísticos descriptivos de los días estancia de pacientes de género femenino hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 104
Tabla 29. Estadísticos descriptivos de los días estancia de pacientes de género Masculino hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia..... 105
Tabla 30. Prueba de Mann Whitney para muestras independientes de los días de estancia y género de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia. 105
Tabla 31. Estadísticos descriptivos de los días estancia de pacientes hospitalizados en el año 2014 por TAB en la CSJDM. Fuente: propia. 106
Tabla 32. Estadísticos descriptivos de los días estancia de pacientes hospitalizados en el año 2013 por TAB en la CSJDM. Fuente: propia. 106
Tabla 33. Prueba de Mann Whitney para muestras independientes de los días de estancia y el año de hospitalización de pacientes con TAB en la CSJDM. Fuente: propia. 107
Tabla 34. Media de los días estancia según el diagnóstico de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia..... 107
Tabla 35. Listado de medicamentos suministrados a los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia..... 109
Tabla 366. Correlaciones de spearman para los días estancia y el estrato socioeconómico de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 110
Tabla 37. Prueba de Kaiser-Meyer-Olkin (KMO) y Bartlett para la reducción de variables en el estudio sobe los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia 112
Tabla 38. Tabla de comunalidades para la extracción de componentes principales en el estudio sobre los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014en la CSJDM. Fuente: propia. 113
Tabla 39. Varianza total explicada para la extracción de componentes principales en el estudio sobe los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 114

Tabla 40. Matriz de componentes para la extracción de componentes principales en el estudio sobe los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 115
Tabla 41. Pruebas omnibus sobre los coeficientes del modelo de regresión logística en predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 117
Tabla 42. Prueba de Hosmer y Lemeshow modelo de regresión logística en predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 117
Tabla 43. Variables en la ecuación modelo de regresión logística en predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 118
Tabla 44. Tabla de clasificación de variables modelo de regresión logística en predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 119
Tabla 45. Estadísticos descriptivos del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 119
Tabla 46. Autovalores del modelo de análisis discriminante en la predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 120
Tabla 47. Prueba de Lambda de Wilks del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 120
Tabla 48. Prueba de M Box del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM en la CSJDM. Fuente: propia 121
Tabla 49. Log determinante del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 121
Tabla 50. Funciones en centroides de grupo en el modelo de análisis discriminante para la predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM Fuente: propia. 122
Tabla 51. Coeficientes de la función discriminante canónica estandarizadas en el modelo de análisis discriminante para la predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 122
Tabla 52. Área bajo la curva curvas ROC para la evaluación de los modelos predictivos del tiempo de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia. 123

RESUMEN

El trastorno Afectivo Bipolar (TAB), es una enfermedad mental principalmente caracterizada por fluctuaciones entre la manía (exaltación) y la depresión (tristeza). Se estima una prevalencia entre el 1% y 2% de la población mundial. La Clínica Psiquiátrica San Juan de Dios ubicada en Manizales (CSJDM), capital del departamento de Caldas, es una clínica especializada en salud mental que presta servicios de consulta, tratamiento y hospitalización principalmente a la población del departamento de Caldas. De acuerdo a la información estadística de los últimos años, 4 de las 10 primeras causas de hospitalización están asociadas al TAB, representando un 45.7% de los casos.

Este fenómeno revela la importancia de estudiar a la población de la CSJDM que padece TAB, y para tal fin, se propuso una investigación relacionada con la Inteligencia Artificial, específicamente con el componente de búsqueda de conocimiento en bases de datos con la intención de construir un posible modelo de predicción de estancia hospitalaria para pacientes con TAB, en donde a partir de factores demográficos se permita conocer el tiempo que el paciente estará hospitalizado durante el tratamiento.

La investigación fue desarrollada bajo la metodología CRISP-DM, que es una metodología utilizada para los proyectos de minería de datos, para la aplicación de las técnicas multivariadas se utilizó el programa estadístico informático SPSS.

ABSTRACT

Bipolar Affective Disorder (BAD) is a mental illness primarily characterized by fluctuations between mania (excitement) and depression (sadness). It is estimated that 1% and 2% of the population suffer from this BAD. The Psychiatric Clinic San Juan de Dios located in Manizales (CSJDM), capital of the department of Caldas, is a specialized mental health clinic providing consultation services, treatment and hospitalization mainly to the population of the department of Caldas. According to the statistics on recent years, 4 of the 10 leading causes of hospitalization are associated with the BAD, representing 45.7% of all total cases.

This phenomenon reveals the importance of studying the population of CSJDM suffering BAD, and to this end, an investigation related to artificial intelligence, specifically with the search component of knowledge discovery in databases with the intention of building a possible prediction model for hospital stay for BAD patients, calculated from demographic factors that allow us to estimate the patien's hospitalization time during treatment.

The research was developed under the CRISP-DM methodology, which is a methodology used for data mining projects, for the application of multivariate techniques computer SPSS was used.

INTRODUCCIÓN

En la práctica médica, la toma de decisiones es un factor clave para la adecuada asignación de recursos, decisiones generalmente basadas en experiencias pasadas permiten a los profesionales en la salud dar respuesta a las necesidades que día a día demandan los pacientes. Sin embargo, las experiencias pasadas por si solas no permiten anticipar comportamientos en la población y se hace trascendental el uso de herramientas que faciliten la toma de decisiones a futuro. La aplicación de minería de datos resulta de gran interés en el campo médico y entre los campos mayor interés se encuentran los modelos de predicción. Estos modelos se pueden construir utilizando una variedad de técnicas, para convertir la información extraída de los datos en resultados significativos que generalmente aportan valor a la rama del conocimiento de la cual se deriva.

Debido al sorprendente crecimiento del desarrollo tecnológico, diariamente se generan y almacenen grandes volúmenes de datos al interior de las organizaciones, estos datos contextualizados se convierten en información que en un ámbito empresarial permiten controlar, optimizar, predecir y facilitar la toma de decisiones. El uso adecuado de la información disponible promueve el capital intelectual dentro de una organización, fortaleciendo uno de los activos más importantes de los recursos empresariales, el conocimiento. El análisis de los datos con los que cuentan los sistemas de gestión empresarial es realizado generalmente por medio de consultas con el lenguaje SQL (Structured Query Languaje) en bases de datos operacionales. Este análisis es poco flexible y poco escalable en grandes volúmenes de datos, pues genera información a partir de consultas previamente establecidas. Estas limitantes que se presentan a la hora de realizar la exploración de las bases de datos ha motivado el uso de técnicas de analítica de datos y minería de datos que posibiliten la extracción no trivial de conocimiento de la información almacenada, este conocimiento puede reportar grandes beneficios para las organizaciones.

La Clínica San Juan de Dios ubicada en Manizales, capital del departamento de Caldas, es una clínica especializada en la atención de pacientes con enfermedades mentales, en donde aproximadamente el 45% de los pacientes hospitalizados durante los últimos años corresponden a casos asociados al Trastorno Afectivo Bipolar. En este departamento, se presenta un fenómeno de alta prevalencia de enfermedades mentales asociadas al Trastorno Afectivo Bipolar. [Bedoya, et al., 2006]

Para la presente investigación, se postula el uso de técnicas de minería de datos bajo el uso de la metodología de CRISP-DM, para la construcción de dos modelos predictivos que permitan identificar los factores que determinan el tiempo de

estancia de los pacientes que padecen Trastorno Afectivo Bipolar, y que hayan sido hospitalizados en la Clínica San Juan de Dios de Manizales entre 2013 y 2014, aclarando que la investigación, resultados y hallazgos, corresponden únicamente los fenómenos estudiados entre dicho rango de fechas para la población mencionada.

DESCRIPCIÓN DEL ÁREA PROBLEMÁTICA

Las enfermedades mentales son fenómenos complejos en los cuales inciden aspectos culturales, sociales y ambientales, como también circunstancias simbólicas y biológicas. Estos fenómenos trascienden en la calidad de vida de las personas, pues interfieren en las relaciones del individuo con su entorno ocupacional, laboral, académico e interpersonal. En Colombia los trastornos mentales afectan especialmente a niños, adolescentes y jóvenes adultos, perturbando gravemente el rendimiento académico y laboral, que según estimaciones se genera un promedio entre 30 y 80 días anuales de pérdida de la productividad. [Arango, Rojas & Moreno, 2008].

En la sociedad aún existe un considerable estigma social que segrega a los pacientes con enfermedades mentales, por tal motivo, la población no solicita la ayuda disponible de manera oportuna, pues se tiene un desconocimiento de los beneficios de la atención psiquiátrica [Posada et al., 2004]. Se sabe que la primera puerta que tocan los pacientes que padecen de enfermedades mentales es la de los médicos generales, y en ocasiones existe un tiempo considerable entre la detección y tratamiento por parte de especialistas.

Proyecciones estadísticas indican que las condiciones psiquiátricas a nivel mundial se incrementarán de un 10.5% a un 15% en el 2020, este crecimiento supera proporcionalmente a las enfermedades cardiovasculares. [Arango, Rojas & Moreno, 2008]. En Colombia, según el Estudio Nacional de Salud Mental, financiado por el Ministerio de la Protección Social en 2003 y encabezado por el médico psiquiatra epidemiólogo José Posada [Ministerio Protección Social, 2005], el 40.1% de la población colombiana entre los 18 y 65 años habrá padecido alguna vez en su vida un trastorno psiquiátrico diagnosticado [Arango, Rojas & Moreno, 2008], además añade, que el 15% de la población ha padecido de algún trastorno asociado al estado del ánimo. En el informe de trastornos mentales en América Latina y el Caribe se estableció que 4.7 millones de personas mayores de 15 años procedentes de esta región padecen Trastorno Afectivo Bipolar (TAB), en donde la tasa de prevalencia para Colombia es del 2% [Rengifo et al., 2012].

El Trastorno Afectivo Bipolar (TAB), antes llamado enfermedad maníaco depresiva, es un trastorno asociado al estado de ánimo, caracterizado por

fluctuaciones entre la manía (fase de exaltación y grandeza) y la depresión (fase de tristeza, inhibición e ideas de muerte), acompañado en algunas oportunidades de síntomas psicóticos o alteraciones cognitivas [Judd, Akiskal, Schettler, Endicott, Maser, et al., 2002]. El TAB está asociado con cargas significativas en la salud de larga duración, cargas sociales y financieras, no sólo para los pacientes sino también para sus familias, otros cuidadores y la sociedad en general. Ésta enfermedad se debe a múltiples factores, en donde el componente genético es el más representativo según estudios, con una tasa entre el 79% y el 93% [Barnett & Smoller, 2009]. Por otra parte, se estima una prevalencia entre el 1% y el 2% de la población mundial, sin embargo, en estudios recientes, en algunas regiones se han observado prevalencias más elevadas, cercanas al 5% [Pardo, Fierro & Ibáñez, 2011].

La Organización Mundial de la Salud (OMS) clasificó el TAB como la sexta causa de discapacidad en toda la población entre los 15 y los 44 años y es catalogada la novena causa cuando se contemplan todas las edades [Ayuso, 2000]. A raíz de esto, supone una carga global para el paciente que afecta su calidad de vida, relaciones familiares, laborales, académicas y sociales. Además, los costos del tratamiento son elevados, pues ocupan el séptimo lugar en cuanto a tratamiento de enfermedades mundiales no mortales [Pardo, Fierro & Ibáñez, 2011]. Según estimaciones, una de cada cinco personas que padece este trastorno realizará un intento de suicidio, siendo treinta veces superior al registrado en la población en general [Lizcano et al, 2011].

Con una adecuada atención y tratamiento farmacológico, es posible alcanzar la remisión de los síntomas agudos y espaciar los periodos existentes entre las crisis causadas por la enfermedad. Los medicamentos tienen un rol fundamental en la fase aguda (tratamiento hospitalario) y en la fase de mantenimiento (tratamiento ambulatorio), pues tras la estabilización de los síntomas es necesario continuar con un tratamiento para prevenir las futuras recaídas [Pardo, Fierro & Ibáñez, 2011].

Aunque en la actualidad se cuenta de una amplia gama de medicamentos eficaces para el tratamiento del TAB, los resultados se ven afectados por la adherencia a lo prescrito y a factores que inciden en la evolución clínica del paciente, dadas estas condiciones, el tiempo de recuperación para una estancia clínica es considerada difícil de estimar. En este punto hay una discrepancia entre los hallazgos resultantes de la experimentación científica y la realidad de la aplicación en la práctica [Pardo, Fierro & Ibáñez, 2011].

ANTECEDENTES

Using data mining to describe long hospital stays (2009).

Gomez & Abasolo (2009), realizaron un estudio que permitiera describir la duración de la estancia hospitalaria a través de minería de datos, la cual puede proporcionar nuevo conocimiento para resolver problemas en el campo de la medicina mediante la detección de patrones desconocidos en datos existentes. En esta investigación, resaltan la complejidad que a veces genera abstraer los datos y su procesamiento cuando provienen de fuentes heterogéneas y no estructuradas, se requiere entonces de un experto médico para su interpretación, sin embargo, algunas veces los datos son insuficientes para describir detalladamente ciertos fenómenos a través de modelos de minería de datos.

Dicho proyecto fue desarrollado bajo la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), metodología que fue creada por un grupo de expertos en el descubrimiento de conocimiento, y es utilizada en el análisis y explotación de la información [Chapman et al, 2009]. CRISP-DM además, permite adaptarse a diferentes contextos, ya que no requiere de una herramienta específica. La recomendación de dicha metodología, es hacer un planteamiento del problema en términos del negocio y hacer la traducción al contexto de minería de datos; en donde se hace una descripción del problema, se procede a limpiar los datos, se ejecutan modelos de minería de datos para el procesamiento, se analizan los resultados, se evalúan con la ayuda de expertos y posteriormente de despliegan los resultados obtenidos. Esta metodología utiliza un ciclo de vida que permite separar el proyecto por fases [Gomez & Abasolo, 2009].

Esta investigación nos permite entender cómo se afrontan los proyectos de minería de datos en general, en especial para proyectos de salud, y nos expone una metodología que permite traducir problemas que inicialmente se encuentran en términos de negocio para llevarlo a un planteamiento que permita abordar los problemas desde la minería de datos.

Factors influencing hospital high length of stay outliers (2012).

Este estudio, realizado a partir de la información de hospitales públicos pertenecientes al Servicio Nacional de Salud Portugés entre 2000 y 2009, estudió los valores extremos de la duración de la estancia hospitalaria y su impacto en la gestión financiera a través del tiempo. Para tal fin, se utilizaron 9.253.087 casos de hospitalizaciones, y se aplicaron modelos de regresión logística para examinar la asociación de cada variable con los altos valores atípicos relacionados con el tiempo de estancia. Posteriormente, se aplicó regresión logística con las variables

para calcular los odds ratios ajustados, utilizando el software de análisis estadístico IBM SPSS Statistics versión 20.0 y SAS en su versión 9.1.

Los resultados revelaron que de los casos estudiados, se encontró una proporción de 3.9% de los valores atípicos altos relacionados con el tiempo de estancia, estos valores, representaron el 19.2% del total de los días de hospitalización. Con esta información se entiende que los casos atípicos tienen una influencia en los costos hospitalarios y deben ser contemplados en la financiación de los hospitales.

Este estudio resalta la importancia de la planificación y la política de los hospitales para proyectar adecuadamente los costos hospitalarios. Por otra parte, resalta la importancia del uso de técnicas como la regresión logística y su aporte en el campo de la salud para la construcción de modelos [Freitas et al, 2012].

La importancia de este antecedente de investigación y su aporte para el presente proyecto, se respalda en la búsqueda de factores que influencian en el tiempo de estancia hospitalaria. La regresión logística demuestra ser una técnica aplicable para este contexto que permite identificar satisfactoriamente como ciertos factores atípicos afectan el tiempo de estancia hospitalaria.

Modelo de regresión logística para la predicción de tratamiento intrahospitalario prolongado en pacientes de la Unidad de Salud Mental del Hospital San Juan de Dios de Santafé de Bogotá (1998).

En el Hospital San Juan de Dios de Santafé de Bogotá, se presentaba un fenómeno preocupante relacionado con los incrementos progresivos en los tiempos de estancia de los pacientes psiquiátricos, especialmente los pacientes que presentaban TAB refractarios al tratamiento con carbonato de litio. Como consecuencia de esto, se desarrolló un modelo para encontrar un grupo de factores de riesgo que permita predecir el desenlace del tratamiento intrahospitalario prolongado. En la literatura clínica se informa que, además de la refractariedad, existen otros factores asociados al diagnóstico de enfermedad afectiva que pueden afectar la prolongación del tiempo de tratamiento, como la presencia de psicosis, el abuso de sustancias y el antecedente de más de tres episodios de enfermedad.

El estudio hacia parte de una investigación de caracterización de pacientes hospitalizados en la Unidad de Salud Mental del Hospital San Juan de Dios de Santafé de Bogotá, para establecer factores y sus respectivas combinaciones asociadas con la utilización del modelo de regresión logística. En el modelo generado, se encontró que el género masculino, el diagnóstico de trastorno de

personalidad y una dosis alta inicial de benzodiacepinas son fuertes predictores del desenlace [Sánchez, 1998].

Este estudio, nos demuestra que el uso de técnicas como la regresión logística, permiten realizar un análisis exploratorio con resultados exitosos en el campo de la medicina, en donde el tiempo del tratamiento y de estancia hospitalaria pueden explicarse a través de la construcción de modelos multivariados de similares características.

Applying a BP Neural Network Model to Predict the Length of Hospital Stay (2013).

La duración de la estancia hospitalaria está estrechamente relacionada con el control de los gastos médicos y la gestión de los recursos hospitalarios. En esta investigación, se implementó un enfoque de minería de datos basado en Redes Neuronales Artificiales (RNA) con Back-Propagación, para construir un modelo de predicción del tiempo de la estancia hospitalaria.

Se analizaron los datos médicos de 921 pacientes diagnosticados con colecistitis y tratados en un hospital de China entre 2003 and 2007. El modelo de predicción alcanzó aproximadamente el 80% de precisión, y reveló 5 variables predictoras: días antes de la operación, grado de la herida, enfoque de la operación, tipo de carga y número de las admisiones.

La investigación encontró que, la minería de datos utilizando Redes Neuronales Artificiales con Back Propagation tiene una gran flexibilidad y robustez para la precisión del tiempo de estancia en pacientes con colecistitis, a diferencia de otros estudios basados en métodos estadísticos tradicionales como la regresión, facilitando el hallazgo de relaciones no lineales entre variables [Li et al, 2013].

El estudio realizado por este grupo de investigadores, permite contemplar la posibilidad del uso de técnicas de inteligencia artificial como las Redes Neuronales Artificiales para la predicción del tiempo de estancia hospitalaria, y nos permite tener una referencia con respecto al tamaño de la población en este tipo de investigaciones para una muestra significativa.

Comparison of Logistic Regression and Neural Network Analysis Applied to Predicting Living Setting after Hip Fracture (2004).

Esta investigación tuvo como objetivo describir y comparar las características de las redes neuronales artificiales y regresión logística para desarrollar modelos de predicción en la investigación epidemiológica. Los datos fueron tomados de 3708 muestras de personas con fracturas de cadera de 186 hospitales de 46 estados, y fueron analizados en la Universidad Buffalo de New York.

Los resultados de la investigación nos demuestran que el seguimiento a la terapia y la autonomía del uso de las funciones de la vejiga y el intestino, eran fuertes predictores para determinar la posibilidad de vivir luego de 6 meses de la cirugía de cadera. Tanto regresión logística, como los modelos de redes neuronales artificiales, hicieron un excelente trabajo de la predicción de las personas que vivían en la casa luego de la intervención quirúrgica, La comparación entre ambos modelos no encontró diferencias significativas entre la capacidad de predicción de la regresión logística y las redes neuronales artificiales para este estudio [Ottenbacher et al, 2004].

Logistic Regression and Artificial Neural Network classification models: a methodology review (2002).

En el campo de la medicina, la regresión Logística y las Redes Neuronales Artificiales son los modelos de elección en muchas clasificaciones de datos médicos. La investigación realizada por Dreiseitl & Ohno (2002), plantea diferencias y similitudes de ambos modelos desde un punto de vista técnico, y resume los hallazgos de los criterios de calidad para la regresión logística y las redes neuronales artificiales ajustados a trabajos de la literatura médica.

Ambos modelos se construyen a partir de la experiencia (datos de casos reales), en donde los datos pueden ser pre-procesados y expresados en conjuntos de reglas como los sistemas expertos basados en el conocimiento, o también pueden hacer las veces de conjuntos de entrenamiento para modelos de aprendizaje.

En el desarrollo de la investigación, se revisaron 72 artículos que comparan el rendimiento de la clasificación de las redes neuronales con los modelos de regresión logística. El objetivo fue determinar el nivel general de publicación de informes basado en los resultados de regresión logística y modelos de redes neuronales artificiales. El estudio en las investigaciones que utilizaban ambos métodos para ver cuál de los dos presentaba una mejor respuesta ante los objetivos planteados.

Los resultados muestran con mayor frecuencia la construcción de modelos basados en regresión logística que modelos basados en redes neuronales artificiales (Ver tabla 1). Esto se puede presentar debido a que las redes neuronales requieren mayor esfuerzo y consideraciones por parte del usuario para lograr el mismo nivel de satisfacción que con la regresión logística, además limita en gran medida a los lectores la reproducción de los resultados reportados durante la investigación. Los detalles de la construcción de los modelos también pueden ser considerados, debido a que no facilita a la evaluación de la calidad de los resultados obtenidos.

	Detallado (%)	No detallado (%)
Detalles del modelo de Regresión Logística	76	24
Detalles de modelo de Redes Neuronales	51	49
Estimación del error generalizado	89	11
Prueba estadística discriminante	61	39
Información de calibración	25	75

Tabla 1. Porcentaje de investigaciones (de un total de 72) que satisfacen los criterios de calidad. Fuente: [Dreiseitl & Ohno, 2002].

La evidencia encontrada en el estudio, indica que cuando se comparó estadísticamente el rendimiento de ambas técnicas, hubo una proporción de 5:2 en donde los casos redes neuronales no fueron significativamente mejores que los obtenidos bajo regresión logística [Dreiseitl & Ohno, 2002].

Esta investigación, permite establece criterios de selección de técnicas para la elaboración de modelos predictivos, el uso de regresión logística según la literatura es una técnica que se ajusta a la naturaleza del proyecto propuesto.

Análisis discriminante no métrico y regresión logística en el problema de Clasificación (2008).

En esta investigación se exponen los resultados de un proyecto de investigación donde se realizó un estudio de comparación entre análisis discriminante y regresión logística, la confrontación se realizó para el caso en el que se clasifican más de dos grupos que provienen de distribuciones normales y no normales, bajo diferentes tamaños de la muestra.

Los tamaños de la muestra y la naturaleza de los datos son factores importantes en la comparación, sin embargo el comportamiento de los dos procedimientos son similares. En los casos en los que la tasa de clasificación errónea para los dos procedimientos es similar, el análisis discriminante presenta mejores condiciones con respecto a la regresión logística en cuanto a la interpretación de los

resultados, esto debido a que la regresión logística no permite interpretación en términos lineales [Usunga & Patiño, 2008].

Debido a su similitud, Ambas técnicas permiten una interpretación de los resultados cuando se cuenta con una naturaleza de los datos como la que tiene el presente proyecto. Las curvas ROC (Receiver Operating Characteristic), son una opción de contraste que permite gráficamente visualizar la especificidad y la sensibilidad de las pruebas realizadas a los modelos resultantes.

Procesos cognitivos y emocionales predictores de la conducta prosocial y agresiva: La empatía como factor modulador (2002).

Esta investigación realizada en el Universidad de Valencia por Mestre, Samper & Frías (2002), busca revisar algunos procesos cognitivos y emocionales que regulan la conducta agresiva en la adolescencia, enfocándose especialmente en los procesos empáticos. Con esta intención se realizaron encuestas a 1.285 adolescentes hombres y mujeres entre los 13 y 18 años seleccionados aleatoriamente.

El uso de análisis discriminante, permitió establecer un perfil diferencial entre los sujetos que presentan agresividad y los sujetos prosociales. Además permitió identificar que los procesos emocionales alcanzaron una mayor correlación con la conducta agresiva, destacando la inestabilidad emocional como principal factor predictor de la agresividad. Los resultados indican que los sujetos más inestables emocionalmente tienen menor empatía y cuentan con menos recursos para frenar su impulsividad, mientras que, los mas empáticos son mas prosociales debido a su emocionalidad más controlada [Mestre, Samper & Frías, 2002].

Tras el uso del análisis discriminante en esta investigación, se puede concluir que es una técnica que puede ser utilizada para identificar factores que determinan el desenlace de un fenómeno determinado, con una interpretación de los resultados que facilite su entendimiento. Esta técnica ha demostrado que puede ser utilizada bajo características similares a la regresión logística, por tal motivo, y bajo las características de la investigación propuesta, parece ser una técnica candidata ideal para la comparación de resultados en ambos modelos (regresión logística y análisis discriminante)

FORMULACIÓN DEL PROBLEMA

Este proyecto busca dar respuesta a la siguiente pregunta de investigación:

¿Es posible la construcción de un modelo que permita la predicción de la estancia hospitalaria para pacientes diagnosticados y hospitalizados con Trastorno Afectivo Bipolar (TAB) en la Clínica San Juan de Dios de Manizales (CSJDM)?

Y de ser así:

¿Cuáles son los factores que influyen en mayor medida en la duración de la estancia hospitalaria?

JUSTIFICACIÓN

Según la Organización Mundial de La Salud [OMS, 2001], el progreso científico y tecnológico del mundo moderno ha tenido un gran impacto en las mejoras de la atención con respecto a la salud mental. Además, avances médicos han logrado reducir considerablemente el tiempo de la estancia hospitalaria y mejorar la calidad de vida de los pacientes que padecen de trastornos mentales, esto se debe a las nuevas alternativas para el tratamiento, resultado de avances en la farmacología e intervenciones exitosas en la psiquiatría [Posada et al., 2004]. Sin embargo, se espera un aumento en la incidencia y prevalencia de trastornos mentales, debido a problemas sociales como la violencia, abuso de sustancias psicoactivas, pobreza y otros factores relacionados [OMS, 2001].

Según información estadística obtenida en la Clínica San Juan de Dios de Manizales (CSJDM), institución que presta servicios de salud mental a la comunidad en general y en especial a la población caldense, de los últimos 5 años, 4 de las 10 primeras causas de hospitalización, se deben al TAB con una representación del 45.7% (5972 casos) (Ver figura 1). Este fenómeno se debe a la alta prevalencia de dicha patología en la región, especialmente al norte de Caldas. En estudios realizados por la Universidad de Antioquia [Bedoya, et al., 2006], el municipio de Aranzazu (Ubicado al norte del Departamento de Caldas) cuenta con condiciones hereditarias complejas, en donde las enfermedades genéticas como el TAB son muy frecuentes, *"Se postula que la incidencia de estas enfermedades son el resultado de un efecto fundador y de la práctica de matrimonios consanguíneos producida por el aislamiento poscolonial de la región"* [Bedoya, et al., 2006]. Debido a esta situación, y a otros efectos ambientales, se estima que la prevalencia de la enfermedad en Aranzazu es está por encima de la media Colombiana 2%[1].

Debido a los altos costos ocasionados por los tratamientos farmacológicos en el TAB y al alto impacto en la calidad de vida para los pacientes y sus familias, es necesario identificar cuáles son los factores que inciden en la disminución del tiempo de estancia de los pacientes que alcanzan los objetivos terapéuticos y farmacéuticos trazados por el médico especialista.

[1] La prevalencia de la población Colombiana para pacientes con TAB es cercana al 2%.

Figura 1. Causas de hospitalización de pacientes entre 2010 – 2015 en la Clínica San Juan de Dios de Manizales (CJDMS). Fuente: propia

A pesar de la existencia de estudios acerca de la incidencia de los medicamentos utilizados en los tratamientos del TAB en la evolución del paciente, para el entorno de la Clínica San Juan de Dios de Manizales no se han realizado estudios enfocados en la extracción de conocimiento de la fuente primaria de información que contiene datos de estancia como es el Sistema de Información Hospitalario. Además, en la región no se han adelantado investigaciones correspondientes al análisis de la estancia hospitalaria de pacientes que padecen TAB asociado a las características demográficas. Tampoco se han encontrado estudios que apliquen metodologías como las propuestas en este proyecto de investigación.

Nos encontramos entonces ante una necesidad del sector salud que debe ser estudiada para encontrar alternativas que permitan mejorar la adherencia al tratamiento durante le hospitalización. La investigación es importante desde diferentes perspectivas. Desde una perspectiva académica, la investigación proporciona una metodología que permite describir el comportamiento de las variables relacionados con la duración de la estancia hospitalaria desde ingreso, hasta su egreso respectivo. Desde un ámbito institucional, es importante caracterizar a los pacientes atendidos por este trastorno en la geografía caldense, con la intención de diseñar estrategias de atención más ajustadas a la población que genere impacto en la sociedad.

Se puede concluir argumentando la originalidad de la investigación, pues sería precursora en la región al trabajar con datos reales de pacientes que padecen este tipo de trastornos en una población golpeada fuertemente por estas patologías. Adicionalmente, los resultados propuestos podrían aportar elementos de valor para la construcción de guías de práctica clínica en la región.

OBJETIVOS

OBJETIVO GENERAL:

Construir un modelo predictivo de la estancia hospitalaria de los pacientes que padecen Trastorno Afectivo Bipolar (TAB) en la Clínica San Juan de Dios de Manizales (CSJDM), a partir de los resultados y mediante el uso de las técnicas de minería de datos regresión logística y análisis discriminante, usando la metodología para minería de datos CRISP-DM.

OBJETIVOS ESPECÍFICOS:

a) Construir una vista minable de datos a partir de la base de datos de la Clínica San Juan de Dios de Manizales (CSJDM) que contiene los datos de los pacientes que fueron hospitalizados entre Enero de 2013 y Diciembre de 2014 con el diagnóstico Trastorno Afectivo Bipolar (TAB); utilizando herramientas que faciliten el proceso de extracción, transformación y carga (ETL).

b) Reducir la dimensionalidad del conjunto de datos utilizando la técnica de Análisis de Componentes Principales (PCA).

c) Generar y validar dos modelos predictivos con los datos demográficos obtenidos de los pacientes, correspondiente a la aplicación de Regresión Logística y Análisis Discriminante.

d) Evaluar y confrontar los dos modelos predictivos con curvas ROC (Receiver Operating Characteristic o Característica Operativa del Receptor).

ALCANCE Y LIMITACIONES

Teniendo en cuenta que la minería de datos permite revelar información previamente desconocidos de manera automatizada e identificar tendencias y comportamientos de un conjunto de datos, se plantean una serie de restricciones para garantizar el éxito del proyecto.

La información utilizada para el estudio, corresponderá únicamente a la existente en las bases de datos de la Clínica San Juan de Dios de Manizales. Los datos con los que se realizará la investigación corresponden a los de los pacientes que fueron hospitalizados en la Clínica San Juan de Dios de Manizales con diagnóstico TAB durante el periodo comprendido entre Enero de 2013 y Diciembre de 2014. Por lo tanto el uso institucional que se le brinde a los resultados de la investigación está sujeto a actualizaciones necesarias que permitan incorporar la dinámica reciente de la problemática, de esta forma se permitirá ajustar el comportamiento del fenómeno de estudio a una situación reciente.

Para la construcción de los modelos predictivos se utilizará el programa estadístico informático IBM SPSS Statistics versión 23 de prueba.

RESULTADOS ESPERADOS

- Caracterizar la población atendida en la clínica San Juan de Dios de Manizales (Caldas) a partir del Análisis e interpretación de los datos de los pacientes que fueron hospitalizados con TAB durante enero de 2013 a diciembre de 2014. El proceso de caracterización no debe ser considerado en un fin por sí mismo, sino, un medio para abrir paso a futuros proyectos de investigación.

- Identificar los factores que inciden en el tiempo de estancia hospitalaria por medio de la reducción de la dimensionalidad de las variables asociadas a los pacientes que fueron hospitalizados con TAB.

- Obtención de dos modelos predictivos de minería de datos, que permitan predecir el tiempo de estancia de pacientes con TAB a partir de su información demográfica. Estos modelos pueden reducir el nivel de incertidumbre acerca del tiempo de estancia hospitalaria de los pacientes hospitalizados con TAB.

- Informe de proyecto en donde se detallen las fases desarrolladas durante la investigación.

REFERENTE TEÓRICO

REFERENTE TEMÁTICO

Trastorno Afectivo Bipolar (TAB).

El TAB es considerado un trastorno mental grave, hace parte de las patologías mentales que con mayor frecuencia requieren tratamiento intrahospitalario, se caracteriza por fluctuaciones entre la manía (exaltación y euforia) y la depresión (tristeza e ideas de muerte). Clínicamente se reconocen varias formas de la enfermedad según los episodios que predominen, esto, de acuerdo con los criterios de la Clasificación Internacional de Enfermedades en su décima versión (CIE-10), y los criterios del Manual Diagnóstico y Estadístico de los Trastorno Mentales en su cuarta versión (DSM-IV). Aproximadamente el 90% de los pacientes que han padecido un episodio maníaco, presentan un nuevo episodio afectivo y aquellos pacientes con TAB no tratado, presentan alrededor de 10 episodios maníacos o depresivos a lo largo de la vida. Entre un 10% y un 15% presentan más de tres episodios al año (cicladores rápidos). Los tratamientos del TAB requieren un tratamiento integral compuesto por dos pilares básicos de atención, una parte contempla los tratamientos psicofarmacológicos y la otra parte se centra en el tratamiento psicosocial [Grupo de Trabajo de la Guía de Práctica Clínica sobre Trastorno Bipolar, 2012].

Una respuesta al tratamiento, está definida como la reducción de al menos el 50% de la sintomatología inicial. Y una fase de remisión se considera en donde se logra controlar la intensidad de los síntomas hasta la reducción o ausencia de estos, esta remisión es el objetivo de los tratamientos considerados como agudos. El objetivo del tratamiento del TAB es lograr una remisión sostenida o recuperación de por lo menos 8 semanas.

Clasificación del Trastorno Afectivo Bipolar

Clínicamente son reconocidas varias formas de esta enfermedad según los episodios que predominen. De acuerdo a los criterios de la Clasificación Internacional de las Enfermedades en su décima versión (CIE-10) [Organización Mundial de la Salud, 1992] y criterios del Manual Diagnóstico y Estadístico de los Trastornos Mentales en su cuarta versión (DSM-IV), existen dos tipos de TAB, los pacientes con Trastorno Bipolar 1 (TB1) tiene al menos un episodio de manía o un episodio mixto (aquellos que combinan simultáneamente síntomas maníacos y depresivos) y pueden presentarse episodios depresivos antes o después. En el Trastorno Bipolar 2 (TB2), el paciente presenta síntomas maníacos menos graves

denominados fases hipomaníacas y episodios depresivos. Tanto la CIE-10 como el DSM-IV concluyen una serie criterios diagnósticos, sin embargo estos criterios diagnósticos no son idénticos, pues entre sus diferencias más significativas se encuentra la cantidad de episodios requeridos para determinar el diagnóstico y la distinción entre TB1 y TB2 [Grupo de Trabajo de la Guía de Práctica Clínica sobre Trastorno Bipolar, 2012].

CIE-10	DSM-IV-TR
Criterios CIE 10 para trastorno bipolar: Trastorno caracterizado por la presencia de episodios reiterados (es decir, al menos dos) en los que el estado de ánimo y los niveles de actividad del enfermo están profundamente alterados, de forma que en ocasiones la alteración consiste en una exaltación del estado de ánimo y un aumento de la vitalidad y del nivel de actividad (manía o hipomanía) y en otras, en una disminución del estado de ánimo y un descenso de la vitalidad y de la actividad (depresión). Lo característico es que se produzca una recuperación completa entre los episodios aislados. A diferencia de otros trastornos del humor (afectivos) la incidencia en ambos sexos es aproximadamente la misma. Dado que los enfermos que sufren únicamente episodios repetidos de manía son relativamente escasos y de características muy parecidas (antecedentes familiares, personalidad premórbida, edad de comienzo y pronóstico a largo plazo) al resto de los enfermos que tienen al menos episodios ocasionales de depresión, estos enfermos se clasifican como otro trastorno bipolar (F31.8). Los episodios de manía comienzan normalmente de manera brusca y se prolongan durante un período de tiempo que oscila entre dos semanas y cuatro a cinco meses (la duración mediana es de cuatro meses). Las depresiones tienden a durar más (su duración mediana es de seis meses), aunque rara vez se prolongan más de un año, excepto en personas de edad avanzada. Ambos tipos de episodios sobrevienen a menudo a raíz de acontecimientos estresantes u otros traumas psicológicos, aunque su presencia o ausencia	**Trastorno Bipolar Tipo I:** Presencia de al menos un episodio maníaco o mixto. Se describen a continuación los criterios diagnósticos según las características del episodio más reciente. Los síntomas de cada uno de los episodios se describen en la tabla siguiente. **Criterios para el diagnóstico de Trastorno bipolar I, episodio maníaco único (296.0x)** **a)** Presencia de un episodio maníaco único, sin episodios depresivos mayores anteriores. **b)** El episodio maníaco no se explica mejor por la presencia de un trastorno esquizoafectivo y no está superpuesto a una esquizofrenia, un trastorno esquizofreniforme, un trastorno delirante o un trastorno psicótico no especificado. **Criterios para el diagnóstico de F31.0 Trastorno bipolar I, episodio más reciente hipomaniaco (296.40**) **a)** Actualmente (o el más reciente) en un episodio hipomaníaco. **b)** Previamente se ha presentado al menos un episodio maníaco o un episodio mixto. **c)** Los síntomas afectivos provocan un malestar clínicamente significativo o un deterioro social, laboral o de otras áreas importantes de la actividad del individuo. **d)** Los episodios afectivos en los Criterios A y B no se explican mejor por la presencia de un trastorno esquizoafectivo y no están superpuestos a una esquizofrenia, un trastorno esquizofreniforme, un trastorno delirante o un trastorno psicótico no especificado.

no es esencial para el diagnóstico. El primer episodio puede presentarse a cualquier edad, desde la infancia hasta la senectud. La frecuencia de los episodios y la forma de las recaídas y remisiones pueden ser muy variables, aunque las remisiones tienden a ser más cortas y las depresiones más frecuentes y prolongadas al sobrepasar la edad media de la vida. **Incluye:** Trastorno maníaco-depresivo. Psicosis maníaco-depresiva. Reacción maníaco-depresiva. Pautas para el diagnóstico. **F31.0 Trastorno bipolar, episodio actual hipomaniaco** **a)** El episodio actual satisfaga las pautas de hipomanía (F30.0). **b)** Se haya presentado al menos otro episodio hipomaníaco, maníaco, depresivo o mixto en el pasado. **F31.1 Trastorno bipolar, episodio actual maníaco sin síntomas psicóticos** **a)** El episodio actual satisfaga las pautas de manía sin síntomas psicóticos (F30.1). **b)** Se haya presentado al menos otro episodio hipomaniaco, maníaco, depresivo o mixto en el pasado. **F31.2 Trastorno bipolar, episodio actual maníaco con síntomas psicóticos** **a)** El episodio actual satisfaga las pautas de manía con síntomas psicóticos (F30.2). **b)** Se haya presentado al menos otro episodio hipomaniaco, maníaco, depresivo o mixto en el pasado. **F31.3 Trastorno bipolar, episodio actual depresivo leve o moderado** **a)** El episodio actual satisfaga las pautas de episodio depresivo leve (F32.0) o moderado (F32.1). **b)** Se haya presentado al menos otro episodio hipomaniaco, maníaco, depresivo o mixto en el pasado.	*Especificar*: **Especificaciones de curso longitudinal** (con y sin recuperación interepisódica) **Con patrón estacional** (sólo es aplicable al patrón de los episodios depresivos mayores) **Con ciclos rápidos.** **Criterios para el diagnóstico de F31 Trastorno bipolar I, episodio más reciente maníaco (296.4x)** **a)** Actualmente (o el más reciente) en un episodio maníaco. **b)** Previamente se ha presentado al menos un episodio depresivo mayor, un episodio maníaco o un episodio mixto. **c)** Los episodios afectivos en los Criterios A y B no se explican mejor por la presencia de un trastorno esquizoafectivo y no están superpuestos a una esquizofrenia, un trastorno esquizofreniforme, un trastorno delirante o un trastorno psicótico no especificado. *Especificar* (para el episodio actual o el más reciente): **Con síntomas catatónicos** **De inicio en el posparto** *Especificar***:** **Especificaciones de curso longitudinal** (con o sin recuperación interepisódica) **Con patrón estacional** (sólo es aplicable al patrón de los episodios depresivos mayores) **Con ciclos rápido.** **Criterios para el diagnóstico de F31 Trastorno bipolar I, episodio más reciente depresivo (296.5)** **a)** Actualmente (o el más reciente) en un episodio depresivo mayor. **b)** Previamente se ha presentado al menos un episodio maníaco o un episodio mixto. **c)** Los episodios afectivos en los Criterios A y B no se explican mejor por la presencia de un trastorno esquizoafectivo y no están superpuestos a una esquizofrenia, un trastorno esquizofreniforme, un trastorno delirante o un trastorno psicótico no especificado.

Se puede utilizar un quinto carácter para especificar la presencia o ausencia de síntomas somáticos en el episodio depresivo actual: F31.30. Sin síndrome somático F31.31 Con síndrome somático. **F31.4 Trastorno bipolar, episodio actual depresivo grave sin síntomas psicóticos** **a)** El episodio actual satisfaga las pautas de episodio depresivo grave sin síntomas psicóticos (F32.2). **b)** Se haya presentado al menos otro episodio hipomaniaco, maníaco, depresivo o mixto en el pasado. **F31.5 Trastorno bipolar, episodio actual depresivo grave con síntomas psicóticos** **a)** El episodio actual satisfaga las pautas de episodio depresivo grave con síntomas psicóticos (F32.3). **b)** Se haya presentado al menos otro episodio hipomaniaco, maníaco, depresivo o mixto en el pasado. **F31.6 Trastorno bipolar, episodio actual mixto** El enfermo ha padecido en el pasado por lo menos un episodio hipomaniaco, maníaco o mixto y en la actualidad presenta una mezcla o una sucesión rápida de síntomas maníacos, hipomaniacos y depresivos. **Pautas para el diagnóstico** Alternancia de los episodios maníacos y depresivos, separados por períodos de estado de ánimo normal, aunque no es raro encontrar un estado de humor depresivo se acompañe durante días o semanas de hiperactividad y logorrea o que un humor maníaco e ideas de grandeza se acompañe de agitación y pérdida de la vitalidad y de la libido. Los síntomas maníacos y depresivos pueden también alternar rápidamente, de día en día o incluso de hora en hora. El diagnóstico de trastorno bipolar mixto sólo deberá hacerse si ambos tipos de síntomas, depresivos y maníacos, son igualmente destacados durante la mayor parte	*Especificar* (para el episodio actual o el más reciente): **Crónico** **Con síntomas catatónicos** **Con síntomas melancólicos** **Con síntomas atípicos** **De inicio en el posparto** *Especificar*: **Especificaciones de curso longitudinal** (con y sin recuperación interepisódica) **Con patrón estacional** (sólo es aplicable al patrón de los episodios depresivos mayores) **Con ciclos rápidos** **Criterios para el diagnóstico de F31 Trastorno bipolar I, episodio más reciente mixto (296.6x)** **a)** Actualmente (o el más reciente) en un episodio mixto. **b)** Previamente se ha presentado al menos un episodio depresivo mayor, un episodio maníaco o un episodio mixto. **c)** Los episodios afectivos en los Criterios A y B no se explican mejor por la presencia de un trastorno esquizoafectivo y no están superpuestos a una esquizofrenia, un trastorno esquizofreniforme, un trastorno delirante o un trastorno psicótico no especificado. **<u>Trastorno Bipolar Tipo II:</u>** Presencia de episodios depresivos mayores recidivantes con episodios hipomaníacos **Criterios para el diagnóstico de F31.8 Trastorno Bipolar Tipo II (296.89):** **a)** Aparición de uno o más episodios depresivos mayores **b)** Acompañados por al menos un episodio hipomaniaco **c)** Ausencia de un episodio maníaco o mixto

del episodio actual de enfermedad, que debe durar como mínimo dos semanas. **Excluye**: Episodio afectivo mixto aislado (F38.0) **F31.7 Trastorno bipolar, actualmente en remisión** El enfermo ha padecido al menos un episodio maníaco, hipomaniaco o mixto en el pasado y por lo menos otro episodio maníaco, hipomaniaco, depresivo o mixto, pero en la actualidad no sufre ninguna alteración significativa del estado de ánimo ni la ha sufrido en varios meses. No obstante, puede estar recibiendo tratamiento para reducir el riesgo de que se presenten futuros episodios. **F31.8 Otros trastornos bipolares** **Incluye**: Trastorno bipolar de tipo II. Episodios maníacos recurrentes **F31.9 Trastorno bipolar sin especificación**	

Tabla 2. Criterios de diagnósticos CIE-10 y DSM-IV para TAB. Fuente: [Grupo de Trabajo de la Guía de Práctica Clínica sobre Trastorno Bipolar, 2012]

Para efectos de la presente investigación, se emplearán los criterios de la CIE-10, los cuales son utilizados al interior de la Clínica San Juan de Dios de Manizales para la asignación de diagnósticos.

Generalidades en el tratamiento del TAB

El manejo clínico del paciente con TAB contempla dos dimensiones, la primera de estas es el estado afectivo actual (transversal), y la segunda es la evolución a largo plazo (longitudinal). Se debe evaluar la cantidad de episodios previos y sus características para elaborar un adecuado diseño del plan de tratamiento. El objetivo del plan terapéutico es mejorar el curso clínico, en donde se incluye la reducción de la frecuencia, gravedad y consecuencias de los episodios afectivos para brindar mayor autonomía al paciente y mejorar su calidad de vida y la de su familia [Jiménez, Martínez, Rosero & Bonilla, 2015].

Dentro de los tratamientos generalmente se reconocen tres estados o fases:

- Fase aguda: durante la descompensación afectiva, generalmente dura entre 6 y 12 semanas para las fases de manía, depresión o episodios mixtos.

- Fase de continuación: inicia a partir de la respuesta clínica y finaliza con la remisión del episodio, la duración de esta fase es variable según los episodios presentados.

- Fase de mantenimiento: esta fase tiene como objetivo la prevención de recaídas.

Tratamiento farmacológico

El tratamiento farmacológico se considera un elemento indispensable en todas las fases, sus objetivos son reducir la intensidad, frecuencia y consecuencias de los episodios agudos que mejoren el funcionamiento general y la calidad de vida del paciente. Antes de elegir el tratamiento farmacológico a utilizar, es preciso valorar cuidadosamente el tipo de TAB y complicaciones previas, al igual que la comorbilidad. Según la literatura, los medicamentos frecuentemente utilizados para el tratamiento son el Litio, anticonvulsivantes como el Ácido valpróico, Carbamazapina y Lamotrigina. También son usados algunos medicamentos antipsicóticos ó medicamentos de segunda generación como la Clozapina, Risperidona, Olanzapina, Ziprasidona, Aripiprazol, Paliperidona y Asenapina [Jiménez, Martínez, Rosero & Bonilla, 2015].

Tratamiento no farmacológico

La medicación por si sola no siempre asegura la ausencia de las recaidas, los factores ambientales, y la baja adherencia terapéutica tienen un papel fundamental en etapas sobre todo de mantenimiento. Por tal motivo se hace necesario realizar intervenciones no farmacológicas como son las intervenciones psicosociales, las cuales se agrupan en diferentes modalidades:

- Psicoeducación
- Terapia cognito-conductual
- Intervenciones familiares
- Terapia interpersonal

Este tipo de intervenciones pretenden contribuir a la prevención de recaidas, brindar información acerca de la enfermedad y su tratamiento, promover el cumplimiento del tratamiento farmacológico, sugerir conductas para alcanzar un estilo de vida que reduzca las posibilidades de nuevas crisis, y fomentar comportamientos saludables [Jiménez, Martínez, Rosero & Bonilla, 2015].

REFERENTE METODOLÓGICO

Sociedad de la información

La sociedad de la información, término utilizado desde los años 70, y profundizado en los 90 con el creciente desarrollo de Internet y de las Tecnologías de la Información y Comunicación (TIC), hace referencia a la sociedad resultante de la última revolución tecnológica, debido al avance de la ciencia, computadoras y telecomunicaciones, haciendo que este tipo de tecnologías se vuelva cotidiana y asequible; esta situación, hace que casi de la noche a la mañana se disponga de enormes cantidades de información y de la capacidad para el intercambio y procesamiento de la misma. La información y el conocimiento tienen un lugar privilegiado en la sociedad y en la cultura y la creación, distribución y manipulación y se convierten en parte estructural de las actividades culturales y económicas [Vila, 2012].

La digitalización de la información se considera el sustento de la nueva revolución informática. Su expresión hasta ahora más compleja, aunque sin duda seguirá desarrollándose para quizá asumir nuevos formatos en el mediano plazo [Vila, 2012].

Sociedad del conocimiento

El término sociedad del conocimiento surgió a finales de los años 90 y es empleada generalmente en medios académicos como alternativa a sociedad de la información. Se caracteriza por ser una sociedad en que se da la apropiación crítica y selectiva de la información por parte del ciudadano, una sociedad formada e informada, capaz de acceder a la información y de aprovecharla por el bien común, bien sea en el ámbito empresarial, educativo, de salud o personal [Burch, 2005].

Sociedad del aprendizaje

La expresión sociedad del aprendizaje surge en 1970 por Robert Hutchins, considerada fundamental para la filosofía educativa. Se caracteriza por considerar el aprendizaje como motor del desarrollo económico de una nación, en donde la

adquisición del conocimiento no puede estar limitada a las instituciones educativas, lo cual requiere adoptar nuevos enfoques provenientes de fuentes no tradicionales para fomentar una colaboración auténtica y abierta entre todos los sectores. La educación y la tecnología al ser parte importante de las nuevas tendencias en internet como catapultador para el aprendizaje, propicia cada vez más un entorno de colaboración interdisciplinaria en las competencias del siglo XXI, como lo son el pensamiento crítico y la resolución de problemas. [CISCO, 2010].

Sistemas de información hospitalarios y conceptos relacionados

Los Sistemas de Información nacen en una era industrial en donde las organizaciones buscan herramientas que apalanquen el crecimiento a través del manejo apropiado de la información y el conocimiento dentro del marco de la sociedad de la información. Un Sistema de Información de Salud, no es más que la aplicación de un Sistema de Información a un entorno sanitario. De modo genérico sería un sistema global e integrado, diseñado para gestionar la información que se genera en el funcionamiento clínico y administrativo de una red de establecimientos que conforman un sistema de salud o un hospital.

Los sistemas de información hospitalarios se encargan de la gestión de la información de ingreso y egreso de pacientes, facturación, finanzas, almacén, gestión de medicamentos, gestión de laboratorios, descripción de procedimientos, gestión de personal, y control de actividades entre otros. Se trata entonces de una solución de infraestructura tecnológica compuesta por hardware y software que soporta las operaciones hospitalarias; toda la información almacenada en las bases de datos de dicho sistema de información conforman lo que se conoce como la historia clínica del paciente. [Vila, 2012].

Estándares

La existencia de los Sistemas de Información de Salud para gestión integrada de sistemas sanitarios requiere del uso de estándares para el registro, codificación, almacenamiento, seguridad y envío de información. Estos estándares afectan tanto a la comunicación interna del sistema global, como para intercomunicación de sistemas aislados, pero no totalmente independientes.

Existe un estándar para el diseño de la arquitectura de historia clínica como lo es el ISO DIS 18308 (Requirements for an Electronic Health Record Reference Architecture) que define un conjunto de requisitos clínicos y técnicos, para una arquitectura de historia clínica que soporta el uso e intercambio de registros

electrónicos, entre y a través de diferentes sectores de salud, diferentes países y diferentes modelos de asistencia sanitaria. [ISO].

Existen también normas para la codificación estandarizada de enfermedades o resultados de pruebas clínicas ampliamente utilizado en formularios tanto en físico como electrónicos. El estándar SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) es una terminología clínica integral que se define como de las más importantes desarrolladas en el mundo y permite representar información clínica de forma precisa; la mantiene y distribuye la International Health Terminology Standards Development Organisation (IHTSDO). [SNOMED Clinical Terms].

En la codificación de enfermedades es el estándar CIE (Clasificación Internacional de Enfermedades), el que más fuerza tiene, siendo utilizado a nivel internacional. Fue desarrollado por la Organización Mundial para la Salud y se encuentra en su décima versión (CIE-10), la cual es una codificación frondosa que recoge enfermedades y una amplia variedad de signos, síntomas, hallazgos anormales, denuncias, circunstancias sociales y causas externas de daños y/o enfermedad [PSICOMED].

Para permitir la comunicación de información de la historia clínica electrónica del paciente entre sistemas y componentes que necesitan añadir, modificar, transferir o acceder a datos, el estándar más moderno y completo es el ISO/CEN 13606. Este estándar sigue una innovadora arquitectura de modelo dual que define una clara separación entre información y conocimiento. La interacción de un Modelo de Referencia, para el almacenamiento de los datos (información), y un Modelo de Arquetipos, para describir semánticamente esas estructuras de datos (conocimiento), proporciona una novedosa capacidad de evolución de los Sistemas de Información. El conocimiento (los arquetipos) pueden cambiar en el futuro, pero los datos permanecerán intactos [CIMM].

Datos – Información - Conocimiento

Hay una estrecha relación entre estos tres conceptos, pero no son lo mismo y es importante diferenciarlos bien. Son los elementos básicos en las etapas del ciclo de la generación de conocimiento y cada una de ellas necesita de la etapa anterior.

Datos: Los datos son la mínima unidad semántica que por si solos son irrelevantes pues no apoyan la toma de decisiones, sin embargo son la materia prima de la información. Pueden ser vistos como un conjunto discreto de valores que no dice nada del porqué de las cosas.

Información: La información es el resultado de un conjunto de datos que es sometido a un tratamiento, estructuración y puesta en contexto con relevancia y propósito, que tiene un significado y que por lo tanto brinda un apoyo para la toma de decisiones.

Conocimiento: El conocimiento es el resultado de integrar la información, experiencia y valores, esta información debe ser reconocida como útil y debe relacionarse con el acervo científico actual.

De una forma resumida podríamos definir la información como datos en contexto y conocimiento como información en contexto [Vila, 2012].

Explotación de información y sistemas inteligentes

La explotación de información se ha definido como la búsqueda de patrones interesantes en grandes volúmenes de datos. Ésta, hace referencia a la aplicación de métodos de sistemas inteligentes para descubrir patrones presentes en la información [Fayaad, 1997].

Los sistemas inteligentes constituyen el campo de la informática en el que se estudian y desarrollan algoritmos que implementan algún comportamiento inteligente y su aplicación a la resolución de problemas prácticos, en este campo surgen diversos estudios que permiten el descubrimiento de conocimiento a partir de conglomerados de información [Michalski, 1983]. Esto se convierte en una alternativa de solución a muchos problemas a los cuales probablemente no se les puede dar respuesta por medio de algoritmos y métodos estadísticos tradicionales, un claro ejemplo es el entorno clínico en donde múltiples variables afectan la evolución de ciertas patologías de diferente manera; sin embargo, el uso de algoritmos especiales, como los utilizados en las técnicas de aprendizaje automático, alcanzan resultados significativos para la elaboración de información.

Los métodos tradicionales de análisis de datos incluyen el estudio de variables por medio de la estadística (análisis de varianza, desviación estándar, regresión, análisis discriminante, series de tiempo, covarianza, análisis de componentes, análisis de clusters, análisis de factores, análisis de multivariable de la varianza y análisis de los discriminantes, etc.) [Moine, Haedo & Gordillo, 2012]. Estos métodos estadísticos son plenamente cuantitativos y se basan en muestras generalmente obtenidas a través de encuestas a una población determinada, mientras los métodos basados en sistemas inteligentes (algoritmos genéticos, inteligencia artificial, sistemas expertos, redes neuronales, etc.), tienen la particularidad de obtener resultados luego de analizar grandes volúmenes de datos y descubren información que los métodos convencionales no logran

encontrar. La búsqueda de información oculta en grandes volúmenes de datos ha contribuido significativamente en sectores financieros, educativos, tecnológicos, aseguradoras, y en ciencias como la salud y la biología entre otras.

Analítica de datos

Actualmente, empresas de diversos sectores económicos en Colombia han comenzado a descubrir los beneficios de la analítica de datos, y son aquellas de mayor éxito las que aplican este análisis para tomar decisiones más inteligentes, que permitan actuar rápidamente y optimicen sus resultados. El propósito de la analítica de datos es facilitar la elaboración de acciones estratégicas que proporcionen servicios que permitan satisfacer las necesidades de un público objetivo. Su estructura, ayuda a las organizaciones a visualizar y explorar sus datos para crear estrategias que permitan llegar de mejor forma a la población y potenciar sus negocios.

Para alcanzar estándares de alta competitividad, es necesario integrar, clasificar y analizar la información, con la intención de solución problemas en las diferentes áreas de negocio. Cabe resaltar que, los resultados del análisis de los datos de forma eficiente generan mayor productividad y permiten establecer acciones de mejora en los procesos.

La analítica de datos, permite a las organizaciones plantear mejoras a sus procesos de negocio, aumentar los niveles de satisfacción de los usuarios y predecir tendencias futuras para mejorar la toma de decisiones. Este tema es crucial para las organizaciones, ya que el éxito del modelo de negocio se fundamenta cada día más en la información que provee del análisis de los datos internos confrontados con el entorno [Fernández, 2015]

Minería de datos

Según [Hernandez, 2004], minería de datos es el conjunto de técnicas que se encaminan a la extracción de conocimiento no trivial a partir de grandes volúmenes de datos. Dicho conocimiento previamente desconocido puede resultar potencialmente útil para los fenómenos estudiados [Stefanovic, Majstorovic & Stefanovic. 2006]. Considerando lo anterior, debe señalarse que la minería de datos utiliza técnicas que buscan dar solución a problemas de predicción, clasificación y segmentación [Umapathy, 2007].

Debido al crecimiento irrefutable durante los últimos años de la sociedad de la información, las organizaciones han experimentado un aumento considerable en el cantidad de los datos que se encuentra alojados en bases de datos relacionales y

otros tipos de fuentes; En ocasiones las empresas no saben que los grandes volúmenes de datos que residen en sus sistemas pueden ser analizados y explotados para obtener nuevo conocimiento valioso y estratégico, sin embargo, este conocimiento no puede ser descubierta bajo el uso de métodos estadísticos convencionales. [Moine, Haedo & Gordillo, 2012].

Desde una perspectiva general, el objetivo de la minería de datos es el análisis automático o semi-automático de grandes volúmenes de datos para extraer patrones interesantes desconocidos previamente. Para tal fin es indispensable contar con almacenes de datos que permitan acceder a información histórica, ya que esta puede resultar útil para explicar el pasado y poder predecir el futuro; este comportamiento en el ámbito organizacional es conocido como valor añadido de una empresa (know-how). La minería de datos pretende automatizar estas tareas de forma cuantitativa incluyendo toda la información disponible [Romeu & Pardo, 2010].

Almacenes de datos

Anteriormente los análisis de los datos se realizaban con el uso de herramientas de consulta que trabajaban sobre las bases de datos transaccionales (SQL) que soportan la operación al interior de las organizaciones. Estas herramientas son poco escalables para el manejo de altos volúmenes de información, por tal motivo es necesario el uso de tecnologías basadas en nuevas arquitecturas conocidas como Data Warehousing (Almacen de Datos). Según [Kimball, Reeves, Ross & Thornthwaite. 1998] el Data Warehousing es el proceso por el cual se organizan grandes cantidades de datos heterogéneos, de forma que permita la recuperación de información para adelantar procesos analíticos. Estos almacenes generan bases de datos centralizadas con una perspectiva histórica, fusionando los datos de múltiples fuentes para permitir que se realicen consultas multidimensionales las cuales son vitales para las organizaciones competitivas [Romeu & Pardo, 2010]. La construcción del almacén de datos se hace a través de la selección de diferentes fuentes de datos tanto internas como externas, y se debe filtrar puesto que el almacén debe contener solo datos necesarios para el procesamiento. (Ver Figura 2)

Las bases de datos transaccionales son las encargadas de hacer la interacción con los sistemas de información, en estas bases de datos se presentan operaciones de lectura, inserción, actualización y borrado de datos, mientras que en un almacén de datos únicamente se realizan operaciones de lectura (Ver Figura 3).

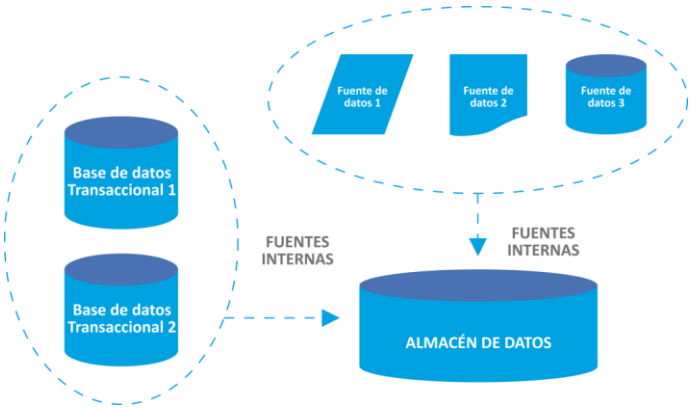


Figura 2. Extracción de registros de una base de datos transaccional para su análisis. Fuente: [Romeu & Pardo, 2010].

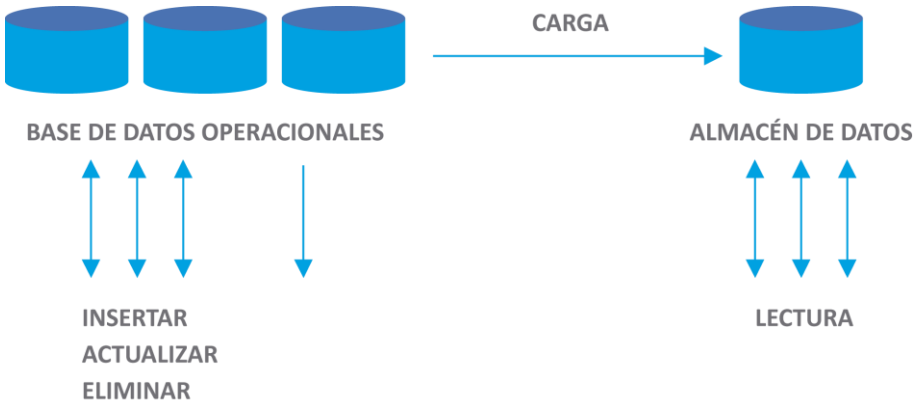


Figura 3. Proceso de carga de un almacén de datos. Fuente: [Romeu & Pardo. 2010].

Tipos de datos.

Los almacenes de datos contienen una serie de registros los cuales se componen en diferentes tipos:

- Numéricos: Cualquier tipo de número (enteros o reales)
- Categóricos: Toman el valor entre conjuntos finitos de categorías y existen dos tipos:
    - Ordenados: cuando existe un orden lógico entre las posibles variables (Bajo, Medio, Alto)
    - No Ordenados: Cuando no existe un orden lógico entre las diferentes categorías (Psicología, Psiquiatría, Cardiología, Pediatría, etc.)

Extracción, transformación y carga (ETL)

El proceso de ETL consta de tres fases en donde cada una de estas es necesaria en el proceso de integración de datos; este proceso consiste en la recuperación datos desde fuentes de origen internas o externas, para ser procesados y transformados según una serie de criterios definidos, y posteriormente ser cargados a una estructura objetivo, generalmente en un sistema foráneo [Kimball & Caserta, 2004]. (Ver Figura 4)

Aunque este proceso no es mandatorio en exploraciones de minería de datos, es necesario cuando se cuenta con grandes volúmenes de datos con el fin de homogenizar los registros o cuando los algoritmos exigen ciertos parámetros para la entrada [Gonzalez, 2011]. Durante este proceso se busca cumplir con los siguientes criterios para darle valor a los datos:

- Remueven errores y corrigen datos vacíos.
- Proveen medidas de la confiabilidad de los datos.
- Almacenan el flujo de datos transaccionales de manera segura.
- Ajustan datos de múltiples fuentes para ser usados en conjunto.
- Estructuran datos para ser utilizados por herramientas de usuario final

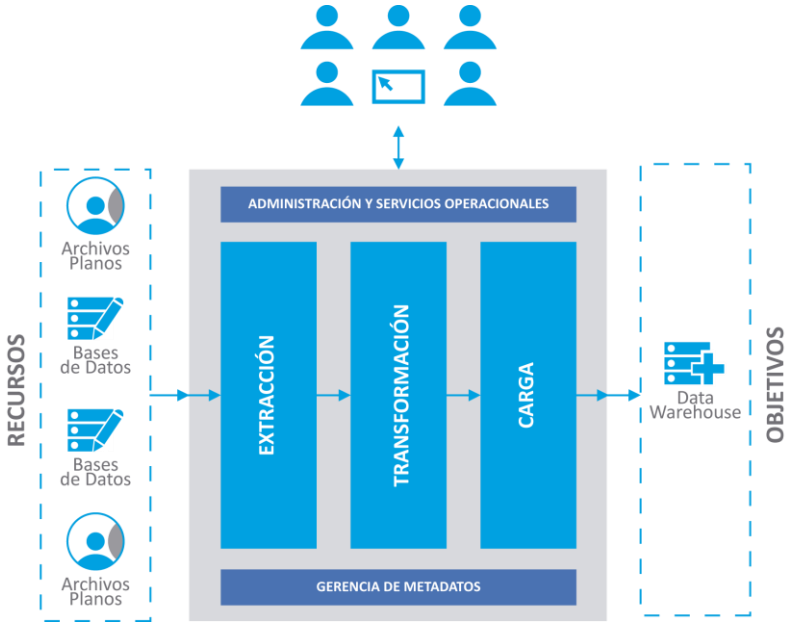


Figura 4. Extracción, transformación y carga. Fuente: http://superhotmobile.com/etl/etl-tools-extract-transform-load-information-builders.html

Metodologías de Minería de Datos

Un proceso de minería de datos involucra fases de comprensión del negocio, identificación de necesidades, selección y limpieza de los datos, generación de modelos matemáticos, ejecución, validación de modelos y finalmente consolidación del nuevo conocimiento adquirido y utilizarlo para resolver el problema planteado. La relación entre todas estas fases tiene una complejidad que se traduce en una jerarquía de subfases [Britos, 2008].

Algunos modelos conocidos como metodologías son en realidad modelos de proceso, en donde se realiza un conjunto de actividades para llevar a cabo una tarea. La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer, y la metodología especifica cómo hacerlo. [Moine, 2012]. Hasta el año 2000 se utilizaba comúnmente el modelo KDD para proyectos de minería de datos, sin embargo, el crecimiento de dicha área en el siglo XXI permitió la incursión de tres nuevos modelos que bosquejan un enfoque sistémico para llevar a cabo el proceso de búsqueda de información, en la comunidad científica se evidencia el uso frecuente de las metodologías CRISP-DM, SEMMA y P3TQ, [Britos, 2008]

Según el estudio publicado en el año 2007 por la comunidad KDnuggest (Data Mining Community's Top Resource), CRISP-DM se ha convertido en la metodología más utilizada para tal fin como se puede observar en la figura 5.

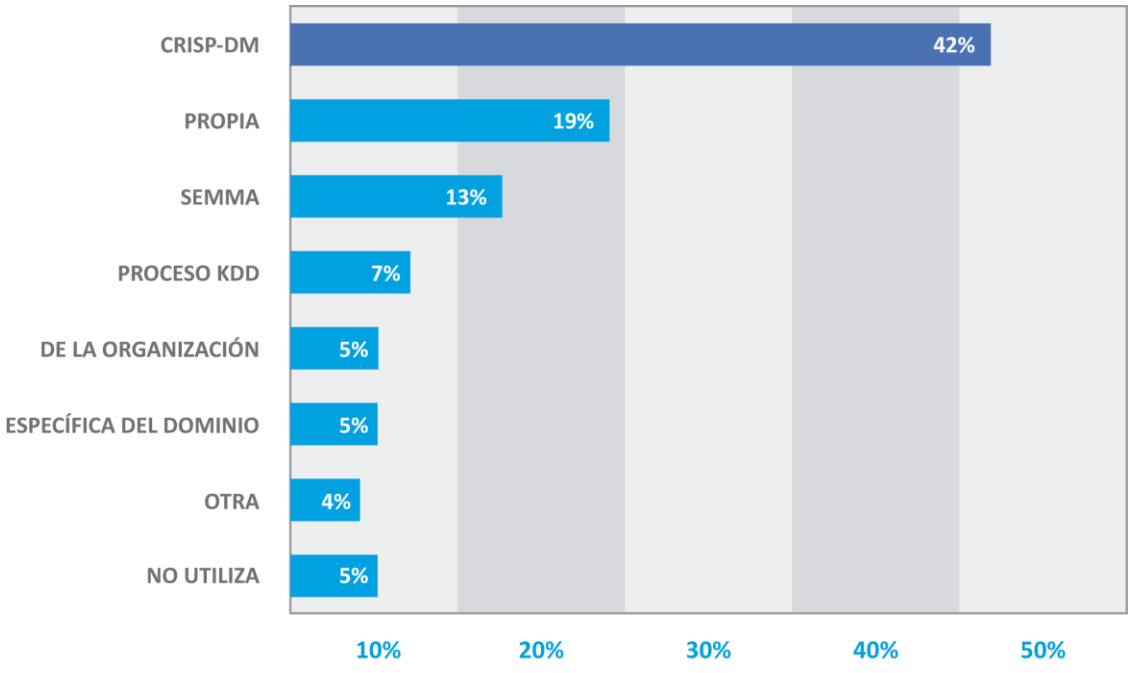


Figura 5. Encuesta realizada por la KDnuggest en el año 2007. Fuente: [Moine, Haedo & Gordillo, 2012].

Algunos modelos profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o SEMMA). [Moine, Haedo & Gordillo, 2012].

Metodología SEMMA

La metodología SEMMA creada por SAS Institute, se define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos [SAS Institute]. Su nombre viene del acrónimo Sample, Explore, Modify, Model, Assess y se centra en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema. Fue propuesta especialmente para trabajar con el software de minería de datos de la compañía SAS. (Ver Figura 6)

Figura 6. Fases de la metodología SEMMA. Fuente: [Britos, 2008]

El proceso da inicio con la selección de una muestra representativa del caso de estudio para dar validez al modelo y aportar valor a la investigación. La forma más común de obtener una muestra es por el método de muestreo aleatorio simple, en donde cada individuo perteneciente a una población tiene la misma posibilidad de ser seleccionado.

La segunda etapa de la metodología se da inicio una vez se tenga delimitada la muestra de la población objetivo, se debe buscar la eficiencia del modelo por medio de una exploración de la información en búsqueda de la simplificación del problema para optimizar la eficiencia del modelo. Para tal fin, la metodología propone el uso de técnicas estadísticas o herramientas que faciliten la visualización para promover la identificación de correlaciones entre las variables. De esta manera se procura establecer cuáles son las entradas del modelo.

Teniendo identificadas las entradas del modelo se comienza con la manipulación de los datos, definiendo que datos cuentan con la pertinencia necesaria según el objeto de estudio para ser introducidos al modelo. Una vez se logre la identificación del conjunto de datos se procede al análisis y modelado de los datos.

La etapa de modelado se encarga de establecer relaciones entre las variables explicativas o independientes y la variable que es objeto de estudio, que posibiliten inferir el valor de las mismas con el nivel de confianza asociado. Generalmente se utilizan técnicas para el modelado basados en la estadística tradicional (métodos de agrupamiento, análisis discriminante, análisis de regresión), y técnicas basadas en sistemas inteligentes (redes neuronales, lógica difusa, árboles de decisión, reglas de asociación, computación evolutiva).

En la última fase se realiza la valoración de los resultados, contrastándolos con otros métodos estadísticos o diferentes muestras de otro tipo de población con niveles de confianza diferentes. En la figura 7 se ilustra el esquema general de la metodología [Britos, 2008].

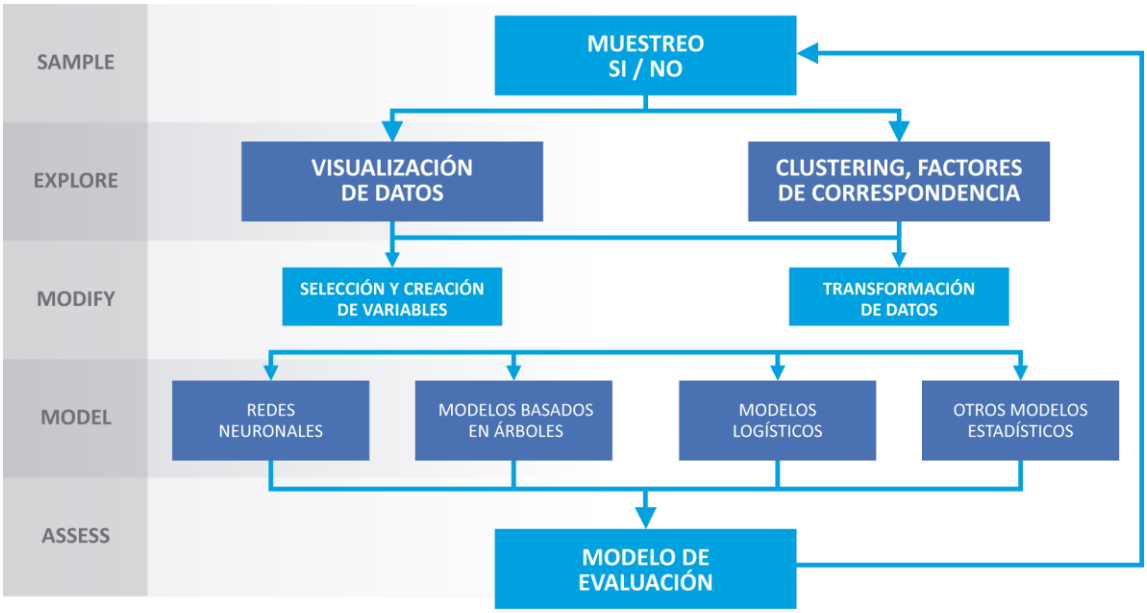


Figura 7. Dinámica de la metodología SEMMA. Fuente: [Britos, 2008].

La fase de modelado de datos en la metodología SEMMA requiere el uso de un software que permita ejecutar una serie de algoritmos de minería de datos para buscar automáticamente una combinación de datos.

Metodología P3TQ

Dorian Pyle propone la metodología Catalyst, conocida como P3TQ (Product, Place, Price, Time, Quantity) en 2003, y plantea la formulación de dos modelos: Modelo de Negocio y Modelo de Explotación de Información o Minería de Datos, los cuales se cuentan con un foco en la cadena de valor de la organización [Pyle, 2003].

El Modelo de Negocio proporciona una guía de pasos para identificar un problema de negocio y la oportunidad que este genera. Cuando el proyecto no tiene bien definido el problema o la oportunidad de negocio es recomendable el uso de las relaciones P3TQ que existen en la cadena de valor de la organización (producto, lugar, precio, tiempo, cantidad).

El Modelo de Explotación de Información o Minería de Datos proporciona una guía para la ejecución de los modelos de minería de datos a partir del Modelo de Negocio [Britos, 2008].

Esta metodología, cuenta con una serie de pasos o fases llamadas “boxes” para ambos modelos, la dinámica consiste en que cuándo se finalice cada fase, producto de cada uno de los boxes, se debe evaluar el resultado y determinar cuál es el próximo paso a seguir. Este modelo permite contar con cierta flexibilidad a la hora de ajustar los caminos posibles dentro de una investigación.

Los boxes que se utilizan en la metodología P3TQ son:

- Actividades, que indican una serie de pasos a realizar.
- Descubrimientos, que proveen acciones de exploración que se necesitan para poder decidir qué hacer en el próximo paso.
- Técnicas, que proporcionan información suplementaria sobre los pasos recomendados en las cajas de descubrimiento o de acción.
- Ejemplos, que dan una descripción detallada de cómo usar una técnica específica.

La metodología P3TQ, plantea 5 escenarios diferentes en los cuales se aplican según las circunstancias (dato, oportunidad, prospectiva, definido, estratégico) [Britos, 2008]. En la figura 8 se puede apreciar la interacción de los diferentes modelos y sus componentes.

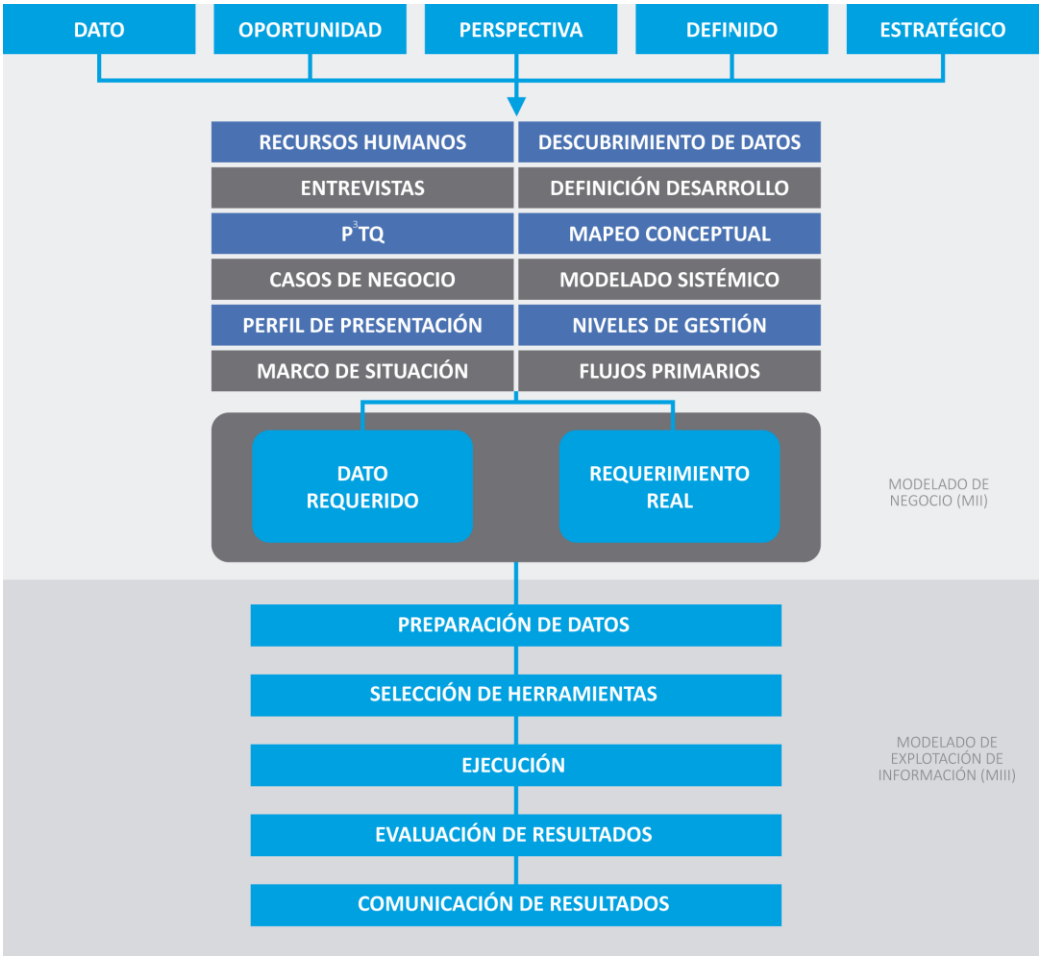


Figura 8. Dinámica de la metodología P3TQ. Fuente: [Britos, 2008].

Metodología CRISP-DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) fué creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, siendo actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos. Se encarga de analizar el proceso de explotación de la información en seis fases diferentes [Chapman et al., 2000].

Cada una de las fases está descompuesta en tareas de segundo nivel que no necesariamente se deben ejecutar sucesivamente, esta metodología establece un conjunto de reglas que se deben tener en cuenta y plantea unas actividades para cada fase del proyecto pero no especifica cómo llevarlas a cabo [Moine , Haedo & Gordillo, 2012].

En la figura 9 se puede observar la interacción propuesta por la metodología entre las fases.

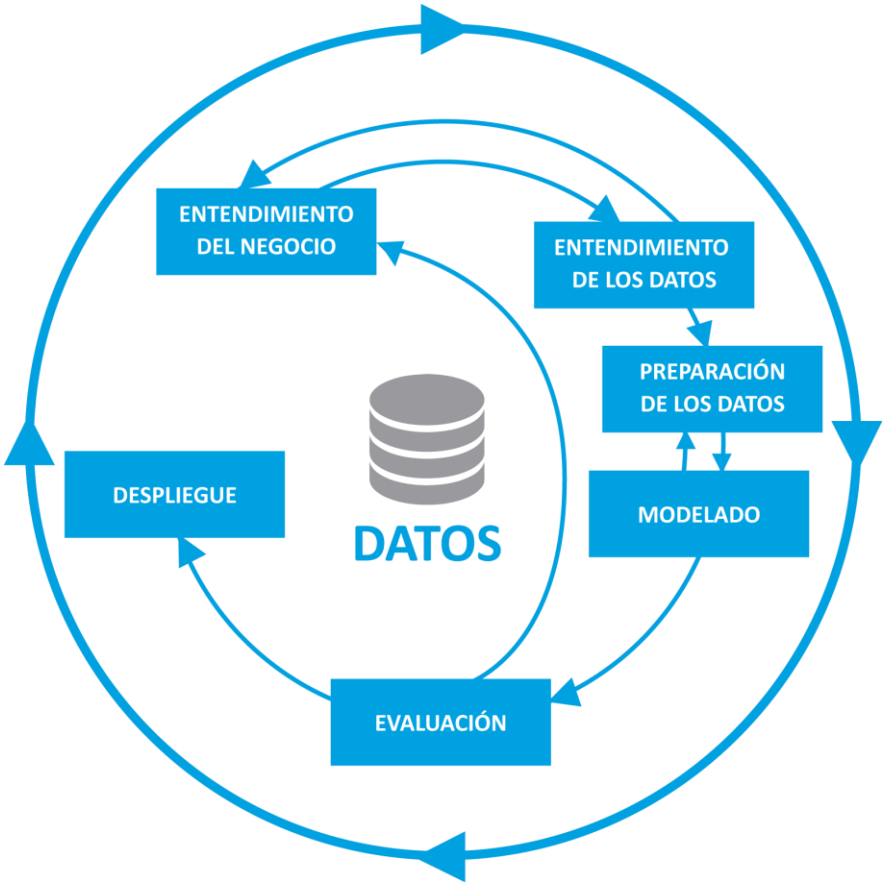


Figura 9. Dinámica de la metodología CRISP-DM. Fuente: [Chapman et al., 2000].

A continuación se describen las fases en las [Chapman et al., 2000] dividen CRISP-DM. (Ver Tabla 3)

- **Entendimiento del negocio**: Esta fase inicial tiene como propósito la compresión de los objetivos del negocio para transformarlos en objetivos técnicos trazables que contrasten con la visión organizacional. Si esta fase no se desarrolla satisfactoriamente no se lograrán obtener resultados fiables y no aportarán valor al negocio. Por este motivo es de vital importancia traducir el conocimiento adquirido en el negocio en un problema de minería de datos que pueda transformarse en un plan preliminar para alcanzar los objetivos.

- **Entendimiento de los datos:** En esta fase se da inicio a la recolección de los datos, el objetivo principal es identificar la calidad de los datos y establecer una serie de relaciones evidentes que permitan definir una serie de hipótesis.

- **Preparación de los datos:** En esta fase se adelanta la preparación para adaptar los repositorios seleccionados a las técnicas de minería de datos que se utilicen posteriormente. Es durante esta etapa en donde se realiza la limpieza de datos y generación de nuevas variables adicionales que permitan adaptarse a la técnica de modelado que será utilizada.

- **Modelado:** Durante esta fase se seleccionan las técnicas de modelado que se ajusten más al objetivo del proyecto de minería de datos, al momento de seleccionar la técnica se deben satisfacer los siguientes criterios: la técnica debe ser apropiada para el tema, disposición de datos adecuados, cumplir con los requisitos del problema, tiempos adecuados para obtención del modelo y conocimiento de la técnica.

- **Evaluación:** Durante esta fase se realiza la evaluación del modelo tomando como punto de referencia los criterios de éxito del problema planteado. Si el modelo es válido en función de los criterios de éxito establecidos se puede realizar la implantación del modelo.

- **Despliegue:** En esta fase y una vez que el modelo haya sido construido y validado, se transforma el conocimiento obtenido en oportunidades de mejora dentro del negocio; el analista realiza una serie de recomendaciones basadas en la evidencia del proceso realizado, de lo que pueden concluir estrategias para la implementación.

Fase	Tareas Componentes	Actividades asociadas
Entendimiento del negocio	Determinar los objetivos del negocio	• Background • Objetivos del proyecto • Criterios de éxito del negocio
	Evaluar la situación	• Inventarios de recursos • Requisitos, supuestos y requerimientos • Riesgos y contingencias • Terminología • Costos y beneficios
	Determinar objetivos del proyecto de explotación de información	• Metas del Proyecto de Explotación de Información • Criterios de éxito del Proyecto de Explotación de Información
	Realizar el plan de proyecto	• Plan de proyecto • Valoración inicial de herramientas
Entendimiento de los datos	Recolectar los datos iniciales	• Reporte de recolección de datos iniciales
	Descubrir los datos	• Reporte de descripción de los datos
	Explorar los datos	• Reporte de exploración de datos
	Verificar la calidad de los datos	• Reporte de calidad de datos

Preparación de los datos	Caracterizar el conjunto de datos	• Conjunto de Datos • Descripción del Conjunto de Datos
	Seleccionar los datos	• Inclusión / exclusión de datos
	Limpiar los datos	• Reporte de calidad de datos limpios
	Estructurar los datos	• Derivación de atributos • Generación de registros
	Integrar los datos	• Unificación de datos
	Caracterizar el formato de los datos	• Reporte de calidad de los datos
Modelado	Seleccionar una técnica de modelado	• La técnica modelada • Supuestos del modelo
	Generar plan de pruebas	• Plan de pruebas
	Construir el modelo	• Configuración de parámetros • Modelo • Descripción del modelo
	Evaluar el modelo	• Evaluar el modelo • Revisión de la configuración de parámetros
Evaluación	Evaluar el resultado	• Valoración de resultados mineros con respecto al éxito del negocio • Modelos aprobados
	Revisar	• Revisión del proceso
	Determinar próximos pasos	• Listar posibles acciones
Despliegue	Realizar plan de implementación	• Plan de Implementación
	Realizar plan de monitoreo y mantenimiento	• Plan de monitoreo y mantenimiento
	Realizar el informe final	• Informe final • Presentación Final
	Realizar la revisión del proyecto	• Documentación de la experiencia

Tabla 3. Tareas de cada fase de la metodología CRISP-DM. Fuente: [Britos. 2008].

Modelo KDD

En la literatura actual podemos encontrar una gran cantidad de definiciones sobre el descubrimiento de conocimiento en bases de datos (KDD, Knowledge Discovery in Databases). Una de las más completas es la que aporta [Fayyad, 1997]: *"El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos)."*

A inicios de 1996 el modelo de KDD constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas fundamentales en proyectos de explotación de datos, en este modelo la minería de datos es una de las etapas dentro del proceso, en donde se extraen los patrones objetivo de la investigación, Sin embargo en la comunidad científica actualmente se usa el término KDD y

minería de datos indistintamente para referirse al proceso de descubrimiento de conocimiento [Moine, Haedo & Gordillo, 2012].

Uno de los componentes más utilizados en el proceso KDD es la minería de datos que integra técnicas de análisis de datos y extracción de modelos, ésta, analiza conjuntos de datos para el descubrimientos de patrones de forma automática o semiautomática.

Este proceso de descubrimiento de información potencialmente útil dentro de las bases de datos se viene utilizando ampliamente ya como una disciplina con un cuerpo teórico muy estructurado. El KDD no es un producto de software, sino un proceso compuesto de varias etapas, es iterativo y explora grandes volúmenes de datos para identificar patrones y determinar relaciones [Fayyad, 1997].

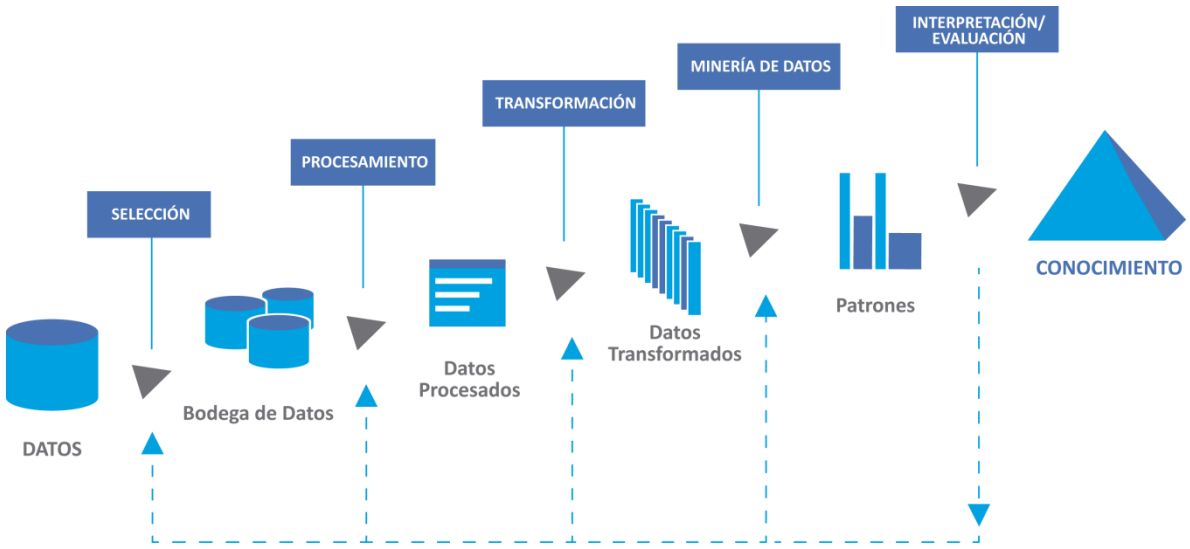


Figura 10. Etapas que componen el proceso KDD. Fuente: [Fayyad, 1997].

En la figura 10 se aprecian las etapas con las que cuenta el proceso de KDD, las cuales son:

1. **Selección de datos:** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Acá los datos relevantes para el análisis son extraídos desde la fuente de datos.

2. **Preprocesamiento:** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos

inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

3. **Transformación:** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.

4. **Minería de datos:** En esta fase es en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos". Este proceso se puede realizar de forma automática o semiautomática según la técnica seleccionada.

5. **Interpretación y Evaluación:** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos. El análisis de expertos en el tema de estudio es vital en esta etapa, pues es quien determina el aporte y valor de la investigación al campo de estudio aplicado [WebMining, 2011].

El KDD al ser un proceso asistido es altamente centrado en el usuario, pues es éste quien interactúa directamente con las herramientas disponibles para dar la inteligencia necesaria a un sistema determinado en la obtención de conocimiento, lo que se convierte en un importante desafío a la hora de determinar las herramientas apropiadas para determinado tipo de problema y cuando usarlas. Particularmente en KDD esto es un problema aun para investigadores, pues el uso de las herramientas adecuadas no garantiza la creación de modelos exitosos [Nigro, 2004].

Taxonomía de las técnicas de minería de datos

Las tareas de las que se encarga la minería de datos son básicamente de dos tipos:

Predictivas: Donde se utilizan variables para predecir valores futuros en función de los datos con los que se disponga inicialmente

Descriptivas: Donde se realiza un proceso de análisis de los datos que se disponen en búsqueda de patrones que puedan describir los datos, permitiendo obtener información posiblemente desconocida de ellos.

[Cabena, et al. 1998] propone para la minería de datos cuatro grandes segmentos que son soportadas por diversas técnicas:

- Modelos predictivos con técnicas de:
    - Clasificación.
    - Predicción de valores.
- Segmentación de bases de datos con técnicas de:
    - Clustering poblacional.
    - Clustering con redes neuronales.
- Análisis de relaciones con técnicas de:
    - Descubrimiento de asociaciones.
    - Descubrimiento de secuencias de patrones.
    - Descubrimiento de secuencias temporales similares.
- Detección de desviaciones con técnicas:
    - Estadísticas.
    - Visualización.

Los métodos inductivos parten de un conjunto de datos iniciales y del conocimiento que estos generan para lograr la construcción de modelos que generen resultados. Este tipo de método hace uso de dos tipos de técnicas [Romeu & Pardo. 2010].

Técnicas predictivas:

- Interpolación: Genera funciones continuas sobre varias dimensiones.
- Predicción secuencial: Se realizan observaciones con el fin de determinar cuál es el siguiente valor de una secuencia.
- Aprendizaje supervisado: Cada observación se compone de muchos factores, se aprende un clasificador a partir de la información que se proporcione. En estos casos la función genera un valor discreto en lugar de continuo. El experto proporciona las clases y los ejemplos iniciales.

Técnicas descriptivas:

- Aprendizaje no supervisado: Es el conjunto de observaciones que no tiene clases asociadas y cuyo objetivo es detectar regularidades en los datos. El experto proporciona los ejemplos mas no las clases, los criterios los define el algoritmo.

- Abducción o aprendizaje analítico: Partiendo de reglas de origen se busca explicar la evidencia respecto a los hechos que se han producido en el contexto. A partir de consecuencias se buscan los hechos.

Algunas técnicas de minería de datos

Regresión

Los métodos de regresión se han convertido en un componente fundamental en el análisis de datos para el descubrimiento de dependencias entre una variable objetivo y una o más variables que explican determinado fenómeno. La regresión utiliza técnicas estadísticas estándar como lo es la regresión lineal, sin embargo muchos de los problemas del mundo real no se resuelven con proyecciones lineales, pues existen múltiples factores que afectan la predicción, en vista de esto, existen otras técnicas de regresión con la capacidad de pronosticar con mayor grado de precisión los valores futuros como lo es la regresión logística [Edelstein, 1999].

Regresión logística

La regresión logística es un procedimiento cuantitativo de gran utilidad para darle explicación a problemas en donde las variables dependientes toman valores en un conjunto finito, es usada para explicar una variable categórica binaria como el éxito (1) o fracaso (0). El objetivo principal de la regresión logística es determinar el modelo más ajustado que permita describir las relaciones de un resultado con las variables independientes (explicativas o predictoras) [Alderete, 2006].

Supongamos que la respuesta dicotómica o variable dependiente representa la ocurrencia o no de un suceso, por ejemplo:

- Un paciente egresa o no antes de determinado tiempo.
- Un paciente muere o no antes del alta

Y son las variables independientes las que pueden ser de cualquier naturaleza, cuantitativas y cualitativas. El proceso de la regresión logística es binomial ya que solo tiene dos posibles resultados. Este proceso se caracteriza por la probabilidad de éxito representada por $\boldsymbol{p}$. La ecuación general de la Regresión Logística se expresa como:

$$p\,(\,y = 1\,|\theta) = \frac{e^{z}}{1 + e^{z}}$$

Donde $p(y = 1|\theta)$ es la probabilidad de acierto dado un nivel de atributo θ, y $\boldsymbol{Z}$ es la combinación lineal de las variables predictoras de esa probabilidad de acierto. [Hidalgo, Gómez, Padilla, 2005]. En donde Z puede expresarse como:

$$Z = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g$$

Donde θ representa el nivel de habilidad o atributo del sujeto de prueba, y g es el grupo al que pertenece el individuo, siendo θg la interacción entre el atributo según el grupo en donde se encuentra. Mientras que β_0, β_1, β_2 y β_3 representan los coeficientes para las interacciones entre los grupos.

La estimación de los parámetros se hace por medio del método de máxima verosimilitud, en donde las estimaciones de los parámetros del modelo son los valores que maximizan la función log- verosimilitud [Santana, 2009].

$$L(Y,X) = \sum_{k=1}^{n} \{Y_k \log[\pi(X_k)] + (1 - Y_k) \log[1 - \pi(X_k)]\}$$

En esta ecuación, Y_k representa la variable respuesta dicotómica (1 para el éxito, 0 para el fallo) y X_k representa el conjunto de variables que predicen a Y_k.

Una de las dificultades en la interpretación de los modelos de variables dicotómicas es que la respuesta dada por la probabilidad de determinado evento no es lineal. Para este caso, la regresión logística emplea una transformación logit, que se entiende como el logaritmo natural de un odds ratio que se define como la razón entre la probabilidad de ocurrencia de p y la probabilidad de fracaso (1 - p). Esta transformación se aplica a la variable dependiente con la intención de expresar una relación lineal entre los resultados de la variable dependiente y sus variables independientes. Dicho esto, el modelo de regresión logística puede representarse de la siguiente manera [Santana, 2009].

$$\log(odds) = LOGIT(P) = \ln\left[\frac{P}{1-P}\right] = Z$$

Análisis discriminante.

El análisis discriminante tiene sus orígenes en las formulaciones del cálculo de distancias formulada inicialmente por Karl Pearson (1920), más tarde por Mahalanobis (1930) y posteriormente el término discriminación se presenta debido a los estudios realizados por Ronald Fisher (1936). Como técnica de análisis de

clasificación, busca predecir la pertenencia de los sujetos a una categoría dentro de los valores posibles de las variables predictivas. Esta técnica nos permite comprobar hasta qué punto las variables explicativas o independientes, explican correctamente a los individuos o sujetos objeto del estudio según su desenlace. Además, es la prueba apropiada para la selección de variables predictivas que permiten identificar los grupos con sus respectivas variables que permiten alcanzar una mejor clasificación [Torrado & Berlanga, 2013].

Esta técnica se considera más que una prueba de clasificación, una prueba de dependencia, cuyo propósito es similar al análisis de regresión logística, la principal diferencia radica en que solo permite variables cuantitativas. La variable dependiente debe ser categórica , y las variables independientes son continuas y establecen a qué grupo pertenecen los sujetos objeto de estudio. Una vez se cuente con los atributos, se forma una combinación lineal de variables predictivas para maximizar las diferencias entre los grupos que construyen el modelo predictivo.

La aplicación de la técnica de análisis discriminante se obtiene del resultado de una ecuación llamada función discriminante, que expresa la combinación lineal de las variables predictivas. Para estos casos, el máximo número de funciones discriminantes obtenidas es equivalente al mínimo entre el número de variables y el de los grupos menos 1, para q grupos, (q -1).

El análisis discriminante se aplica a individuos en el caso en donde se puedan asignar solamente a dos grupos (salida dicotómica), a partir de K variables discriminadoras.

Fisher, resuelve este problema mediante la función discriminante $D = W_1 X_1 + W_2 X_2 + \ldots + W_k X_k$.

Las puntuaciones discriminantes son los valores que se obtienen al asignar valores de $(X_1, X_2, \ldots, X_k)$ en la siguiente ecuación.

Se trata de obtener los coeficientes de ponderación W_j

Si se consideran N observaciones entonces la función discriminante $Di = W_1 X_{1i} + W_2 X_{2i} + \ldots + W_k X_{ki} \quad \forall\, i=1,\cdots,N.$

(Di) es la puntuación discriminante correspondiente a la observación i-ésima [De la Fuente, 2011]. La función discriminante en la forma matricial se representa de la siguiente manera:

$$\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \\ X_{1N} & X_{2N} & \cdots & X_{kN} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{pmatrix}$$

Expresada en función de las desviaciones a la media:

$$\begin{pmatrix} D_1 - \overline{d}_1 \\ D_2 - \overline{d}_2 \\ \vdots \\ D_N - \overline{d}_N \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \\ X_{1N} & X_{2N} & \cdots & X_{kN} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{pmatrix}$$

La función discriminante en diferencias es entonces: d = X w

La variabilidad de la función discriminante se expresa por la suma de los cuadrados de las desviaciones de variables discriminantes:

$$d' d = w' X' X w$$

Donde X' X es una matriz simétrica que expresa desviaciones cuadráticas respecto a la media de las variables. Se puede entonces descomponer en suma de cuadrados entre grupos F y suma de cuadrados dentro de los grupos V.

T = X' X es la matriz de suma de cuadrados y productos cruzados (varianzas covarianza) para un conjunto de observaciones T = X' X = F + V, en donde: d d' = w' X' X w = w' (F + V) w = w' F w + w´V w.

Los ejes discriminantes provienen de los vectores asociados a los valores propios de la matriz (V^{-1} F) ordenados de mayor a menor [De la Fuente, 2011].

Las puntuaciones discriminantes corresponden a los valores obtenidos a la proyección de cada punto del espacio k-dimensional de las variables originales sobre el eje discriminante. Con los coeficientes w se obtienen:

$$Max\ \lambda = \frac{w' F w}{w' V w} = \frac{separación\ entre\ grupos}{separación\ dentro\ del\ grupo}$$

Reducción de la dimensionalidad de conjuntos de datos

En muchas ocasiones las investigaciones se ven enfrentadas a grandes cantidades de registros y atributos que describen determinados fenómenos. Debido a esto se ha buscado por medio de procedimientos reducir la dimensionalidad del conjunto de datos por medio de la reducción de atributos relevantes para la investigación, dejando los más representativos para el objeto de estudio. Algunas veces los grandes volúmenes de datos pueden producir análisis de datos fuera de los rangos deseados [Toro, Pérez & Bernal, 2007]. Lo que buscan las técnicas de reducción de la dimensionalidad es reducir la cantidad de columnas o atributos y perder la menor cantidad de información posible al mismo tiempo

Análisis de componentes principales

El Análisis de Componentes Principales (PCA) es una técnica estadística multivariable. Esta, permite reducir la dimensionalidad del conjunto de datos a costa de una pequeña pérdida de la información, transformando un conjunto de **(*p*)** variables iniciales en un nuevo conjunto de **(*q*)** variables **(*q* <= *p*)** a los que se conoce como componentes principales. Las **(*q*)** nuevas variables se obtienen de la combinación lineal de las variables originales. Los componentes se ordenan en función del porcentaje de la varianza explicada, dado esto, el primer componente será el más relevante por ser el que explica mayor porcentaje de la varianza de los datos [Martín, Díaz, Torres & Garnica, 1994].

En la práctica es frecuente que se disponga de información adicional que amplía la matriz de los datos originales. Se puede disponer de nuevas medidas de nuevos individuos para los que se conozcan sus variables analizadas, estos datos se conocen como datos suplementarios debido a que no forman parte de la creación de los componentes principales. La técnica del Análisis de Componentes Principales fue desarrollada por Hotelling en 1933 y está fundamentada en ajustes ortogonales por mínimos cuadrados desarrollada por K. Pearson en 1901.

Los componentes principales son nuevas variables que cuentan con las siguientes propiedades:

- Cada componente debe ser la combinación lineal de las variables originales.
- La suma de la varianza generalizada de los componentes es igual a la original.

- La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector proprio que lo define y la suma de los vectores propios de la matriz

Análisis de curvas ROC (Receiver Operating Characteristic Curve)

El análisis de curvas ROC es un método estadístico utilizado para establecer la sensibilidad (capacidad de clasificar casos realmente positivos en casos positivos) y la especificidad (capacidad de clasificar los casos realmente negativos en casos negativos) de una prueba diagnóstica de un modelo, que produce resultados continuos en función de los falsos positivos. Las curvas ROC tienen tres propósitos específicos [Cerda & Cifuentes, 2012]:

- Determinar el punto de corte de una escala continua en el que se alcanza la sensibilidad y especificidad más alta,
- Evaluar la capacidad discriminativa de una prueba diagnóstica, es decir, su capacidad de diferenciar cuando un paciente permanece hospitalizado o egresa del servicio.
- Comparar la capacidad discriminativa de dos o más pruebas diagnósticas que expresan sus resultados a través de escalas continuas.

Las curvas ROC son conformadas por pares ordenados, construidas con verdaderos positivos y falsos positivos. Un par formado por una tasa de verdaderos positivos y una de falsos positivos es graficado para cada configuración. De esta manera, es comúnmente definida como el gráfico de la tasa de verdaderos positivos para todos los posibles casos de la prueba.

Como ejemplo, teniendo en cuenta los resultados de una prueba en particular en dos poblaciones, una población con una enfermedad, la otra sin la enfermedad, la distribución de los resultados de la prueba se superponen de la siguiente forma (Ver Figura 11):

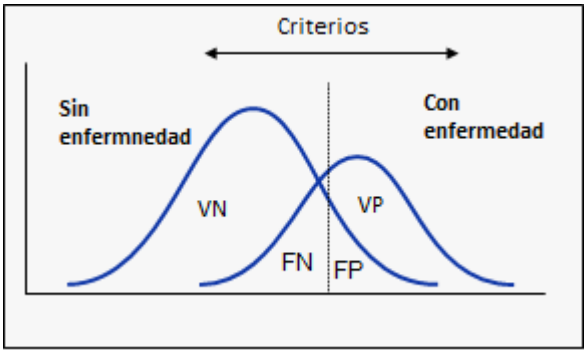


Figura 11. Distribución de resultados de una población que presenta una enfermedad y una que no presenta. Fuente: [MedCalc, 2015].

Para la población que será clasificada entre enfermos y no enfermos, existirán 4 posibilidades según los criterios de evaluación:

- Si el caso es positivo y es clasificado como positivo se cuenta como un verdadero positivo (VP).
- Si el caso es positivo y es clasificado como negativo, cuenta como un falso negativo (FN).
- Si el caso es negativo y es clasificado como negativo se cuenta como un verdadero negativo (VN).
- Si el caso es negativo y es clasificado como positivo se cuenta como un falso positivo (FP).

En la tabla de contingencia (Ver Tabla 4), se muestra la relación existente entre estas clasificaciones [Noguera, 2010].

	Muestra	
	Estándar de oro positivo	**Estándar de oro negativo**
Positivo	Verdadero positivo **(VP)**	Falso Positivo **(FP)**
Negativo	Falso Negativo **(FN)**	Verdadero Negativo **(VN)**

Tabla 4. Tabla de contingencia. Fuente: [MedCalc, 2015].

En este contexto, se utilizan dos índices para evaluar la calidad en la prueba diagnóstica:

- Sensibilidad, hace referencia a la probabilidad de que el resultado de la prueba sea positiva para aquellos casos en donde la enfermedad está presente (verdadera tasa positiva):

$$Sensibilidad = \frac{VP}{VP + FN}$$

- Especificidad, hace referencia a la probabilidad de que el resultado de la prueba sea negativo cuando la enfermedad no está presente (verdadera tasa negativa):

$$Especificidad = \frac{VN}{FP + VN}$$

A raíz de estos índices de calidad de la prueba, se desprenden las siguientes ecuaciones que buscan responder casos más puntuales [Noguera, 2010]:

$$Cociente\ de\ probabilidad\ positiva = \frac{Sensibilidad}{1 - Especificidad}$$

$$Cociente\ de\ probabilidad\ negativa = \frac{1 - Sensibilidad}{Especificidad}$$

$$Valor\ predictivo\ positivo = \frac{VP}{VP + FP}$$

$$Valor\ predictivo\ negativo = \frac{VN}{FN + VN}$$

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

Probablemente la propiedad más ampliamente utilizada en las curvas ROC es el área bajo la curva, AUC (Area Under Curve). Esta, es una porción del área de un cuadrado, cuyos valores estarán entre 1 y 0, sin embargo, al ser una prueba no informativa, produce una diagonal entre (0,0) y (1,1), la cual ocupa un área de (0.5) en la gráfica por defecto, por lo cual ningún clasificador al referirse a variables dicotómicas puede ser menor a este valor [Noguera, 2010].

$$AUC = \int_0^1 ROC(t)dt$$

A continuación, un ejemplo (Ver Figura 12) en donde se puede ilustrar el uso de las curvas ROC para dos test diagnósticos hipotéticos (A y B), y la línea de no-discriminación (Línea ND). Para cada curva ROC, las flechas indican el punto de corte que determina la sensibilidad y la especificidad conjuntas más alta.

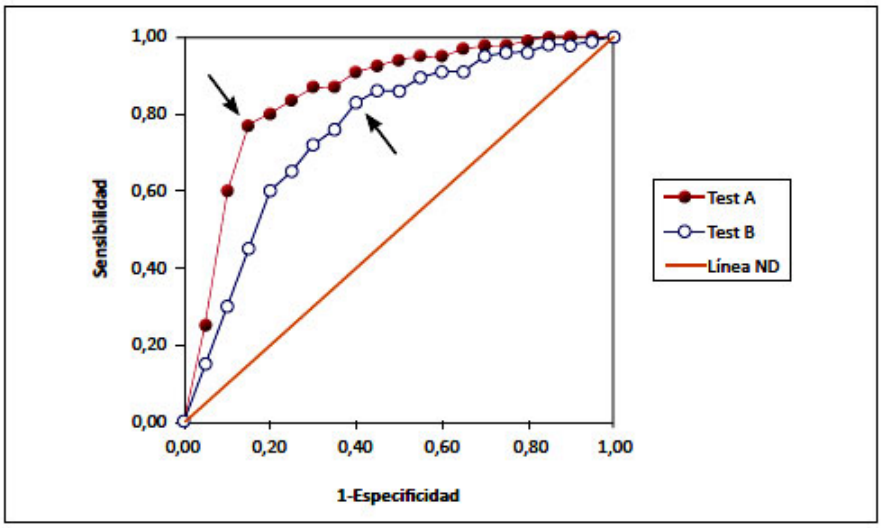


Figura 12. Gráfico ejemplo de curcas ROC. Fuente [Cerda & Cifuentes, 2012].

ESTRATEGIA METODOLÓGICA

Buscando predecir el tiempo de estancia de pacientes que padecen TAB en la CSJDM, se postula el uso de técnicas de minería de datos que permitan determinar la probabilidad de ocurrencia de un suceso determinado. Para la presente investigación, el evento que determinará el tiempo de estancia hospitalaria es el tiempo transcurrido (días) entre el ingreso y el egreso de un paciente en un servicio de hospitalización.

Los días de estancia serán pues la variable dependiente que se buscará predecir en esta investigación. Para la predicción del tiempo de estancia, se buscará determinar cuáles son las variables y características (variables independiente) que inciden en la cantidad de días transcurridos durante el ingreso y el alta de hospitalización.

Con el fin de predecir el tiempo como variable dependiente dentro del estudio y dadas las técnicas de minería de datos a utilizar dentro, según médicos especialistas en psiquiatría de la CSJDM, la población que será utilizada para la investigación de pacientes con TAB, los días de estancia se pueden categorizar en dos grupos principalmente, los que egresen entre los 0 y 15 días y los que egresen pasados los 15 días luego de su ingreso.

Suministradas estas condiciones por criterio experto de los médicos psiquiatras de la CSJDM, se buscará identificar a través del modelo predictivo, cuáles son los factores que determinan el tiempo de estancia para pacientes que padecen TAB

en la CSJDM. Además, dada la naturaleza cuantitativa de la variable dependiente (días de estancia), se recodificará la variable en dos categorías: estancias de 15 días o menos, y más de 15 días de estancia, con el objetivo de presentar un análisis comparativo entre regresión logística y análisis discriminante, dos modelos multivariados que presentan similitudes técnicas.

Tipo de estudio.

Se realizará un estudio descriptivo, en el cual se describan algunas características de pacientes que padecen TAB, hospitalizados entre los años 2013 y 2014 en la CSJDM, este estudio utilizará la información histórica suministrada por fuentes secundarias como lo es la base de datos institucional en donde se encuentra la historia clínica.

El tipo de investigación utilizada será de consulta empírica, puesto que el propósito es determinar cómo unas variables influyen en otras, este tipo de investigación frecuentemente es utilizado para calibrar modelos empíricos que permiten la predicción de un resultado (variables dependientes) conociendo ciertos valores (variables independientes).

Población.

Se tomará como población todos los casos de pacientes que hayan sido diagnosticados con TAB y posteriormente hospitalizados en la CSJDM, entre Enero 1 del 2013 y Diciembre 31 de 2014, por lo tanto, no se realizará un muestreo debido a que la clínica dispone de una base de datos e historia clínica digital en la cual están registrados los pacientes objeto del estudio. El recuento de casos para esta investigación es de 1084 hospitalizaciones con TAB.

Variables incluidas en el estudio y fuentes de información.

La principal fuente de información para esta investigación serán las bases de datos institucionales, y por tratarse de información sensible para la institución, únicamente se tomarán las variables que allí se encuentren relacionadas con el objeto de estudio. La cantidad de estas variables, su naturaleza, el nivel de medición, su definición operacional y valores permitidos o categorías se encuentran detallados en el [Anexo 1].

Control del sesgo de selección.

Para realizar una correcta selección de los pacientes objeto de estudio se procederá de la siguiente manera:

Una vez sean identificados los pacientes con el diagnóstico TAB en las bases de datos de la institución, se verificarán las características de haber sido hospitalizado entre el 1 de Enero de 2013 y Diciembre 31 de 2014, en caso de alguna inconsistencia, se procederá a ubicar el caso particular en la historia clínica correspondiente y se debe registrar el dato que allí corresponda. Ya dispuestos todos los integrantes de la población que cumplan con las características anteriormente señaladas, se organizarán en hojas de cálculo incluyendo todas las variables disponibles para el análisis.

Viabilidad y factibilidad.

El comité de bioética de la Clínica San Juan de Dios se encuentra interesado en la realización de este trabajo y están dispuestos a proporcionar el acceso a las bases de datos, a la historia clínica y a los formularios secundarios, además, la institución podrá proporcionar información estadística de todas las admisiones y hospitalizaciones que se han presentado entre 2013 y 2014.

Aspectos éticos

De acuerdo con los principios establecidos en la Resolución Colombiana número 008430 del Ministerio de Salud en Octubre 4 de 1993, por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud. Título II: **DE LA INVESTIGACIÓN EN SERES HUMANOS**, Capítulo 1: **DE LOS ASPECTOS ÉTICOS DE LA INVESTIGACIÓN EN SERES HUMANOS.**

Artículo 8: "*En las investigaciones en seres humanos se protegerá la privacidad del individuo, sujeto de investigación, identificándolo solo cuando los resultados lo requieran y éste lo autorice*" [República de Colombia Ministerio de Salud, 1993].

En vista de esto, el presente proyecto solo utilizará información epidemiológica, estadísticas de estancias y su respectiva información demográfica. Cabe resaltar que no se utilizará ni divulgará ningún tipo de información que pueda ir en contra de las políticas de confidencialidad establecidas por la institución. Se garantiza que no se identificará a ningún paciente objeto de estudio y se mantendrá la confidencialidad de la información relacionada con su privacidad.

Por tratarse de datos contenidos en la historia clínica y sus formularios secundarios se tendrá como referente la resolución 1995 de 1999, por la cual se establecen normas para el manejo de la Historia. Para todos los efectos considerados en la ley de habeas data, la información será tratada de manera impersonal. [República de Colombia Ministerio de Salud, 1999]

Para tal fin, se presentó resumen ejecutivo del proyecto [Anexo 2] al comité de bioética de la Clínica San Juan de Dios de Manizales, quién aprobó la investigación [Anexo 3].

Selección de la metodología de minería de datos

Durante el análisis comparativo de metodologías para evaluar la gestión de proyectos de minería de datos [Moine, Haedo & Gordillo, 2012], en cuanto a punto de partida la metodología P3TQ plantea más herramientas para una adecuada identificación de los escenarios iniciales del proyecto, SEMMA inicia desde el conjunto de datos de SAS y tanto CRISP-DM como KDD comienzan con un análisis previo del negocio que proporciona elementos importantes para la identificación de puntos de partida.

La estructuración de fases del proceso favorece a KDD, CRISP-DM y P3TQ, puesto que contemplan un análisis y comprensión del problema, caso opuesto a SEMMA quien inicia desde el propio conjunto de datos. La preparación de los datos se ve implícita en todas las metodologías al igual que la evaluación de los patrones encontrados [Moine, Haedo & Gordillo, 2012].

SEMMA carece de una implementación de los resultados obtenidos, mientras que CRISP-DM además de tener dicha fase, propone una planificación para análisis futuro (análisis postmortem) en donde se busca plasmar información importante que aporte a futuras investigaciones y proporcione lecciones aprendidas en cuanto a decisiones tomadas y tecnologías utilizadas.

Los modelos CRISP-DM y P3TQ especifican con mayor detalle las tareas que se deben llevar durante el proceso de desarrollo del proyecto. Al puntualizar las actividades específicas a realizar dentro de cada fase, ambos modelos pueden ser considerados metodologías que aportan elementos de gran valor para alcanzar el éxito del proyecto debido al nivel de detalle con la que se describen las actividades dentro de cada fase. Las metodologías CRISP-DM y P3TQ proponen actividades para la gestión del proyecto como lo son el tiempo, el costo y el riesgo, sin embargo no explican tareas de control y monitoreo, por su parte KDD y SEMMA no incluye ninguna de estas actividades dentro de su proceso [Moine, Haedo & Gordillo, 2012].

Bajo estas premisas se puede concluir que las metodologías para gestión de proyectos de minería de datos más adecuadas son CRISP-DM y P3TQ, los cuales cuentan con un mayor nivel de completitud (Ver Tabla 5).

Fases	KDD	CRISP-DM	SEMMA	P3TQ
Análisis y comprensión del negocio	• Comprensión del dominio de la aplicación	• Comprensión del negocio		• Modelado de negocio
Selección y preparación de los datos	• Crear el conjunto de datos • Limpieza y pre-procesamiento de los datos • Reducción y proyección de los datos	• Entendimiento de los datos • Preparación de los datos	• Muestreo • Comprensión • Modificación	• Preparación de los datos
Modelado	• Determinar la tarea de minería • Determinar el algoritmo de minería • Minería de datos	• Modelado	• Modelado	• Selección de herramientas de modelado inicial
Evaluación	• Interpretación	• Evaluación	• Valoración	• Refinamiento del modelo
Implementación	• Utilización del nuevo conocimiento	• Despliegue		• Comunicación

Tabla 5. Fases del proceso de minería de datos en cada modelo. Fuente: [Moine, Haedo & Gordillo, 2012].

Según [Britos, 2008], la comparación entre las dos metodologías (CRISP-DM, P3TQ) se define por el aporte desde la inteligencia de negocios al proyecto (Ver Tabla 6).

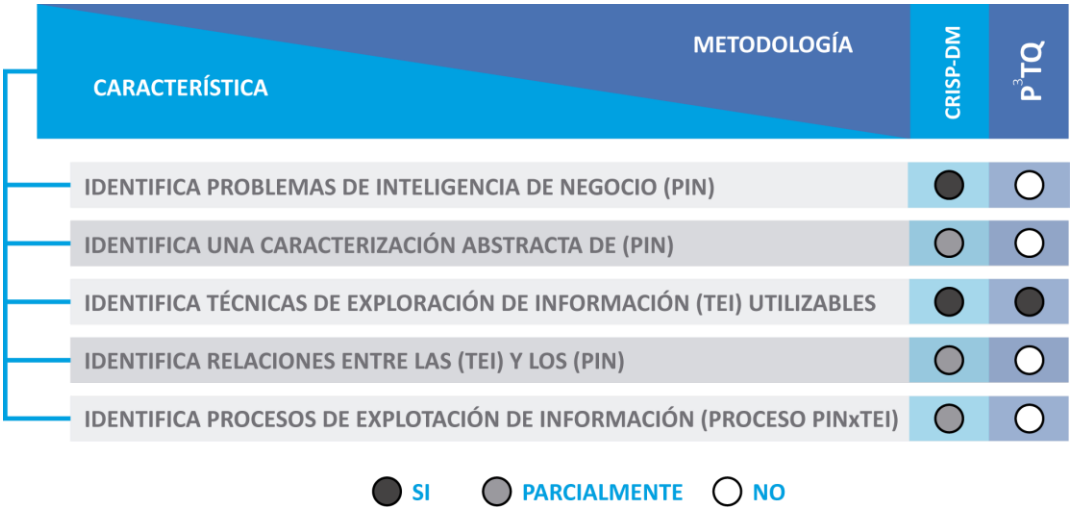


Tabla 6. Conceptos de inteligencia de negocios, técnicas y procesos de explotación de información abarcados por las metodologías. Fuente: [Britos, 2008].

Procesamiento de los datos

Una vez obtenida la hoja de cálculo con los pacientes objeto de estudio, ésta se subirá a una versión de prueba de la plataforma informática de SPSS 23 para la aplicación del proceso de minería de datos bajo el modelo Cross Industry Standard Process for Data Mining (CRISP-DM) para las siguientes fases:

- Comprensión de los datos.
- Preparación de los datos.
- Modelado.
- Evaluación.

Plan de análisis de los datos

Una vez realizada la fase 2 (Comprensión de los datos) y la fase 3 (Preparación de los datos) de CRISP-DM, se procederá de la siguiente manera:

1. **Análisis univariado.**
   Se realizará un análisis univariado en el cual se describirá de manera independiente cada una de las variables que conforman la base de datos de pacientes con TAB. Específicamente, se calcularán distribuciones de frecuencias para las variables cualitativas (tablas y gráficos), medidas de tendencia central y medidas de dispersión para las variables cuantitativas (tablas y gráficos).

2. **Análisis bivariado.**
   Se realizará un análisis bivariado con la intención de determinar una relación o asociación entre las variables del conjunto de datos que permitan describir los datos (para las variables cualitativas será asociación, mientras para las variables cuantitativas será la relación o correlación). Las técnicas estadísticas aplicadas en este análisis dependerán de la naturaleza, nivel de medición y valores posibles que contienen las variables de estudio. Este análisis estará acompañado por tablas y gráficos que ilustren los hallazgos dentro del conjunto de datos.

3. **Reducción de dimensionalidad del conjunto de datos**

   Se aplicará la técnica de Análisis de Componentes Principales (PCA) para descubrir la estructura latente (dimensiones) del conjunto de variables y en lo posible reducir el espacio de atributos, es decir, a partir de la totalidad de las variables que se tienen, se tratará de reducir a un número menor de las mismas, como tal, este procedimiento es "no dependiente" (es decir, no se asume una variable dependiente especifica). Este análisis de componentes

principales será utilizado en esta investigación para estos dos fines, primero reafirmar la correlación existente entre los factores, y segundo permitir la reducción de factores que componen la totalidad de los atributos, identificando los factores más relevante que deberían ser seleccionado a la hora de realizar un análisis multivariado.

4. **Análisis multivariado (Modelado)**
   Para el análisis multivariado se requieren como mínimo 50 registros o una relación de 5 registros por variable, en esta investigación la cantidad de registros mínima recomendada serian aproximadamente 400. Para la construcción del modelo se cuenta con 1084 registros, más del doble de lo esperado, ahora bien, para la evaluación del modelo generalmente se usa un 5% de los registros, pero para esta investigación que presenta una buena cantidad de registros, se decidió disponer de mayor cantidad de registros para el conjunto de pruebas con un 9%, utilizando el 91% como conjunto de aprendizaje.

   Las técnicas multivariadas que se utilizaran serán:

   a. Regresión logística:
   b. Análisis Discriminante.

5. **Evaluación de los modelos (fase Evaluación)**
   Para este análisis se evaluarán los dos modelos predictivos de los días de hospitalización de los pacientes con TAB y serán comparados a través del análisis de curvas ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor), el cual proporciona elementos para seleccionar un modelo óptimo.

Durante la última etapa de la investigación, se consignarán las conclusiones del proyecto y se generará un reporte de la investigación en donde se podrá encontrar una discusión del tema con el apoyo de médicos psiquiatras, quienes desde su conocimiento pueden darle sentido a los resultados encontrados y darles contexto en el entorno social; de igual manera se documentará la experiencia adquirida durante el desarrollo del proyecto en donde se incluirán recomendaciones para futuros proyectos con características similares.

PRESUPUESTO

COSTOS DEL PERSONAL						
	PARTICIPANTE	**ROL**	**HORAS / SEMANA**	**VALOR HORA**	**HORAS ESTIMADAS**	**COSTO TOTAL**
1	Ing. Cristian Daniel Zuluaga	Investigador	18	$ 15.000	1260	$ 18.900.000
2	Ph.D. María Helena Mejía	Directora	3	$ 50.000	210	$ 10.500.000
					SUB TOTAL	$ 29.400.000

Tabla 7. Costos del personal. Fuente: propia.

COSTO DE ADQUISICIONES				
	DESCRIPCIÓN	**CANTIDAD**	**VALOR UNIARIO**	**COSTO TOTAL**
1	Equipo de cómputo	1	$ 1.900.000	$ 1.900.000
2	Papel (resma)	1	$ 12.000	$ 12.000
			SUB TOTAL	$ 1.912.000

Tabla 8. Costos de adquisiciones. Fuente: propia

COSTO DE SERVICIOS				
	DESCRIPCIÓN	**CANTIDAD**	**VALOR UNIARIO**	**COSTO TOTAL**
1	Impresiones	600	$ 150	$ 90.000
2	Diseños gráficos	1	$ 200.000	$ 200.000
3	Desplazamientos	48	$ 4.000	$ 192.000
			SUB TOTAL	$ 482.000

Tabla 9. Costos de servicios. Fuente: propia

COSTO TOTAL DEL PROYECTO		
	DESCRIPCIÓN	**VALOR**
1	Costos del personal	$ 29.400.000
2	Costos de adquisiciones	$ 1.912.000
3	Costos de servicios	$ 482.000
	COSTO TOTAL	$ 31.794.000

Tabla 10. Costo total del proyecto. Fuente: propia

CRONOGRAMA

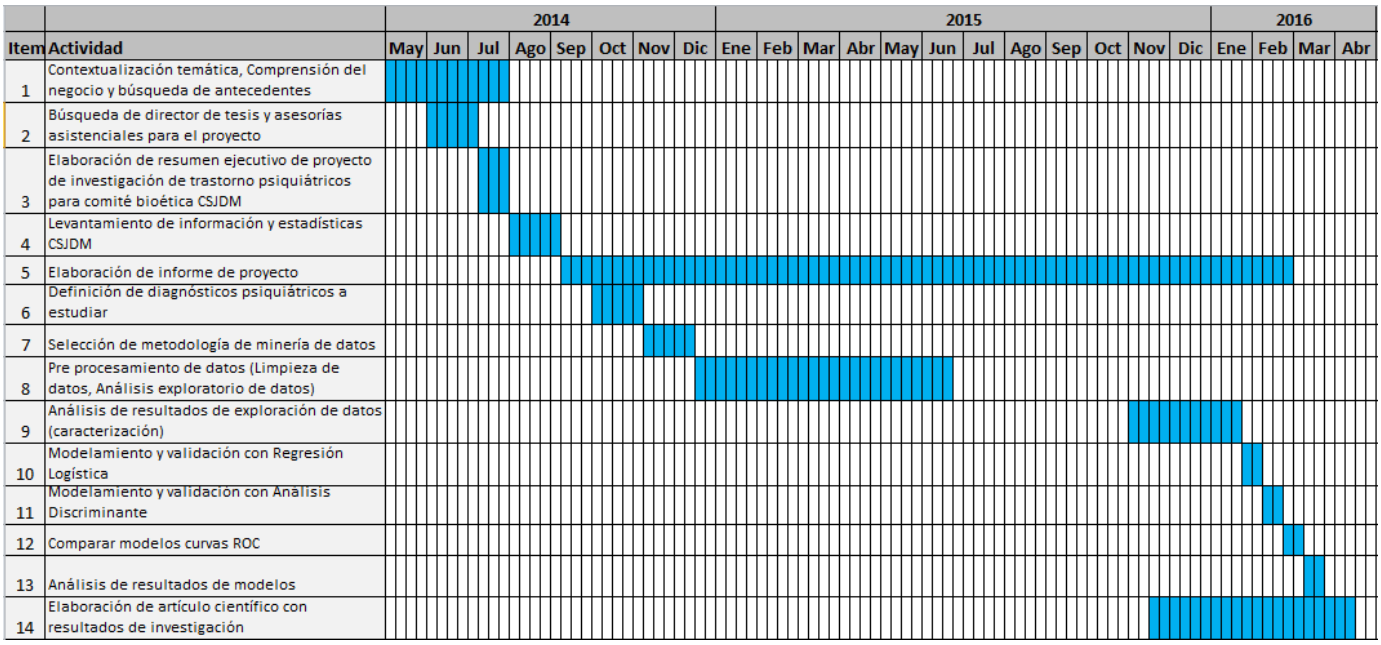


Tabla 11. Cronograma de actividades. Fuente: propia

RESULTADOS

Procesamiento de datos

Descripción de las Hospitalizaciones entre 2013 y 2014

En la presente investigación se tomaron 1753 eventos de Trastorno Afectivo Bipolar (TAB) que requirieron hospitalización en la Clínica San Juan de Dios de Manizales (CSJDM), de las cuales 639 equivalentes al 36% fueron en el 2013 y 1114 que representan el 64% en el 2014. Es importante aclarar que para este análisis se incluyeron todas las hospitalizaciones registradas entre 2013 y 2014, independientemente si el paciente reingresaba o no en el transcurso de este periodo. En la figura 13 se evidencia que en promedio se presentaron 73 hospitalizaciones de esta enfermedad por mes; además, teniendo en cuenta que los datos sugieren un valor de 0,22 en el coeficiente de asimetría, existe una mayor concentración de valores a la izquierda del promedio, lo que indica que hay menor número de meses que superan la cantidad de hospitalizaciones promedio, siendo estos concernientes al periodo comprendido entre marzo y noviembre del año 2014; destacando que en los meses de octubre y noviembre el número de hospitalizaciones ascendió a 123, siendo la cantidad de casos por mes más elevada dentro del periodo de tiempo considerado para el estudio.

El número de hospitalizaciones mensuales para el año 2013 presentó un valor promedio de 53 con una desviación estándar de 17, mientras que para el año 2014 su valor medio fue de 93 con una desviación estándar de 22, lo cual permite identificar un coeficiente de variación de Pearson (CV) de 31,6% y 23,9% para los años 2013 y 2014 respectivamente. Dicho coeficiente indica que existe una mayor variabilidad de los datos con respecto a la media para el número de hospitalizaciones del año 2013, sin embargo, ambos coeficientes sugieren una dispersión baja de los resultados con respecto a su promedio.

La prueba de hipótesis que busca contrastar la hipótesis nula de diferencia entre medias igual a cero, frente a la hipótesis alternativa de diferencia entre medias diferente a cero, sugiere que con un 95% de confianza, dado que el estadístico t calculado es -4,91, se rechaza la hipótesis nula puesto que el valor-p es menor que 0,05, indicando que existen diferencias entre las medias del número de hospitalizaciones mensuales de los años 2013 y 2014; dicha diferencia se hace evidencia gráficamente en el diagrama de caja (Figura 14)

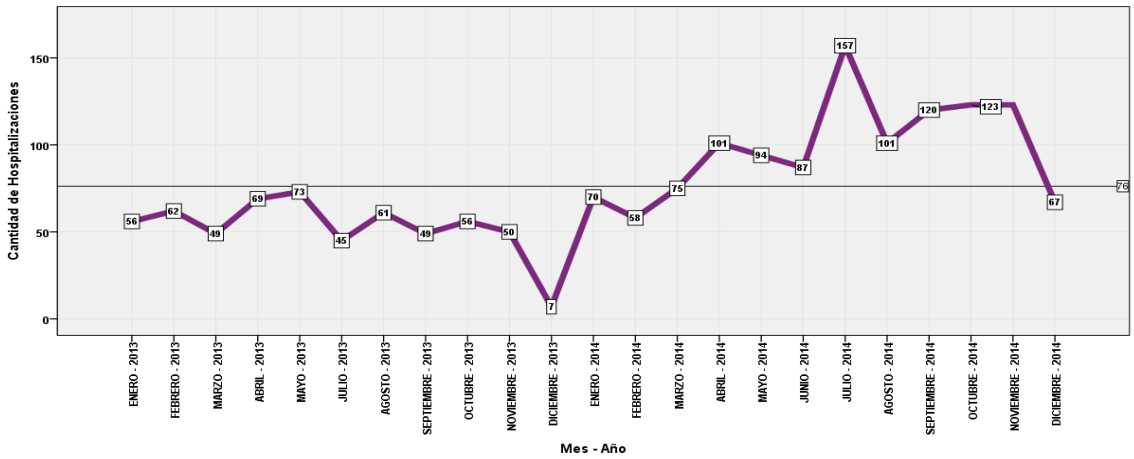


Figura 13. Cantidad de hospitalizaciones por TAB registradas entre 2013 y 2014 en la CSJDM. Fuente: propia.

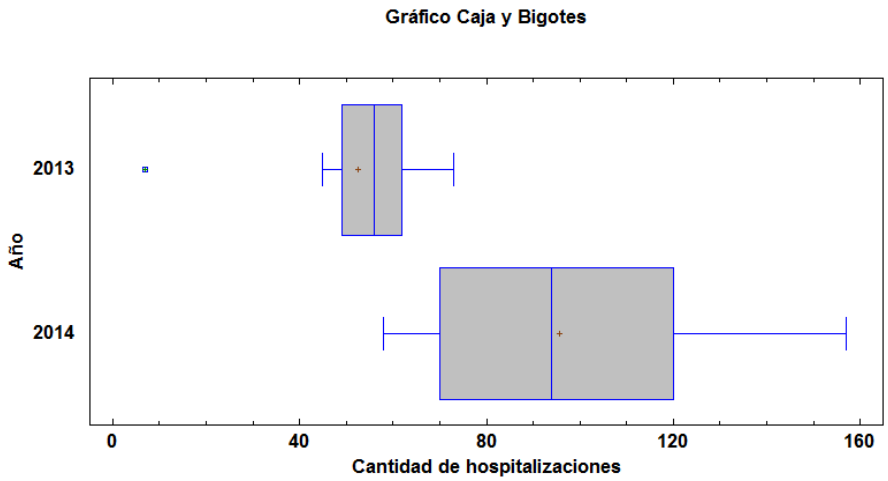


Figura 14. Diagrama de caja para el número de hospitalización para las hospitalizaciones registradas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

De acuerdo con el tipo de diagnóstico asociado al TAB, el mayor porcentaje de hospitalizaciones se registraron por Trastorno Bipolar, episodios maníacos con síntomas psicóticos, el cual se clasifica con el código F31.2 según la décima versión de la Clasificación Internacional de Enfermedades (CIE-10). Este diagnóstico para el 2013 reporta un 12.6% (221) del total de hospitalizaciones relacionadas con el TAB, mientras que para el año 2014 el porcentaje de hospitalizaciones para este diagnóstico se incrementó a un 17.5% (306).

En general todos los diagnósticos asociados al TAB presentaron mayores porcentajes de hospitalizaciones en el año 2014, a excepción del Trastorno Bipolar, actualmente en remisión (F31.7 según CIE-10), que presentó el mismo porcentaje de hospitalizaciones del año 2013; y el Trastorno Bipolar, episodio

actual hipomaníaco (F31.0 según CIE-10), que pasó del 0.8% (14) de las hospitalizaciones en el 2013 al 0.4% (7) en el 2014. Se destaca que el diagnóstico Otros Trastornos Bipolares (F31.8 según CIE-10), pasó del 1.5% (27) de las hospitalizaciones en el 2013 al 9.6% (169) en el 2014. (Ver Figura 15 y Tabla 12.)

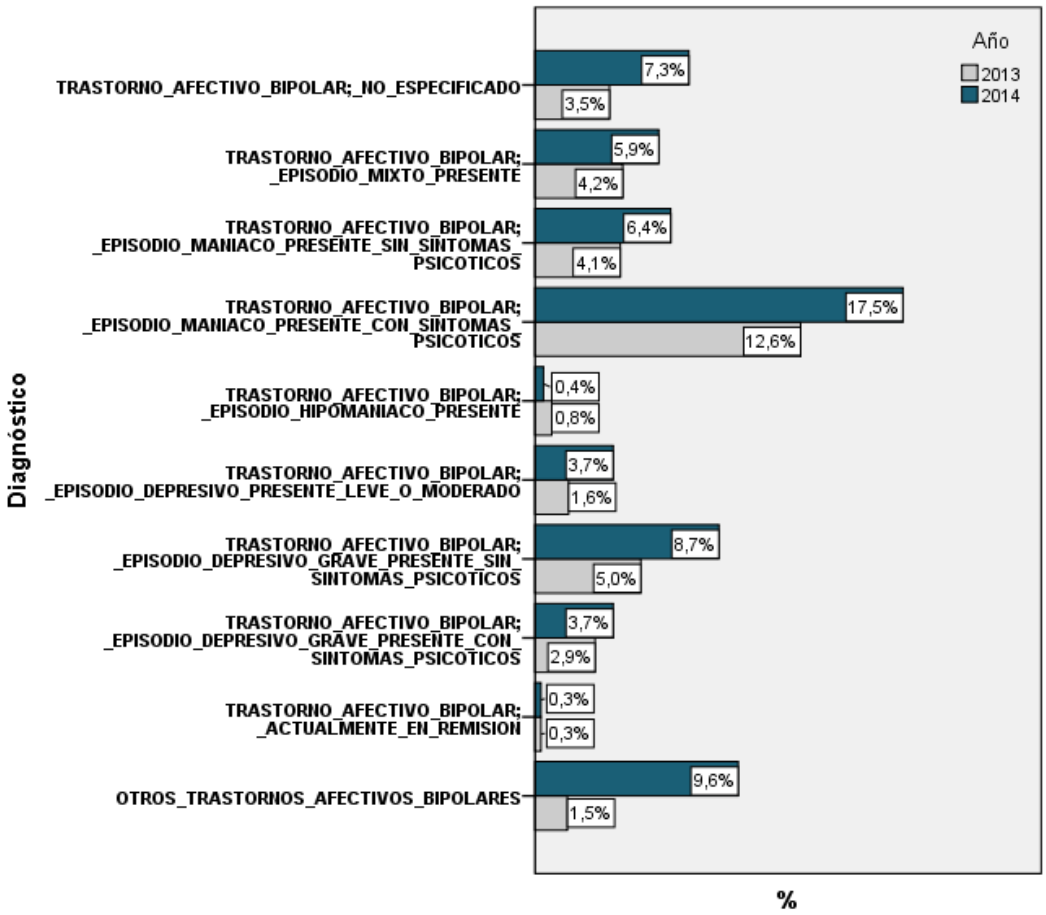


Figura 15. Distribución de las hospitalizaciones registradas entre 2013 y 2014 por TAB según el tipo de diagnóstico y el año de ocurrencia en la CSJDM. Fuente: propia.

DIAGNÓSTICO	Año		
	2013	2014	Total
F31.8 OTROS TRASTORNOS AFECTIVOS BIPOLARES	27	169	196
F31.7 TRASTORNO AFECTIVO BIPOLAR; ACTUALMENTE EN REMISIÓN	5	5	10
F31.5 TRASTORNO AFECTIVO BIPOLAR; EPISODIO DEPRESIVO GRAVE PRESENTE CON SÍNTOMAS PSICÓTICOS	50	65	115
F31.4 TRASTORNO AFECTIVO BIPOLAR; EPISODIO DEPRESIVO GRAVE PRESENTE SIN SÍNTOMAS PSICÓTICOS	88	153	241
F31.3 TRASTORNO AFECTIVO BIPOLAR; EPISODIO DEPRESIVO PRESENTE LEVE O MODERADO	28	65	93
F31.0 TRASTORNO AFECTIVO BIPOLAR; EPISODIO HIPOMANÍACO PRESENTE	14	7	21
F31.2 TRASTORNO AFECTIVO BIPOLAR; EPISODIO MANÍACO PRESENTE CON SÍNTOMAS PSICÓTICOS	221	306	527
F31.1 TRASTORNO AFECTIVO BIPOLAR; EPISODIO MANÍACO PRESENTE SIN SÍNTOMAS PSICÓTICOS	71	113	184
F31.6 TRASTORNO AFECTIVO BIPOLAR; EPISODIO MIXTO PRESENTE	73	103	176
F31.9 TRASTORNO AFECTIVO BIPOLAR; NO ESPECIFICADO	62	128	190
TOTAL	639	1114	1753

Tabla 12. Cantidad de hospitalizaciones registradas entre 2013 y 2014 por TAB según el tipo de diagnóstico en la CSJDM. Fuente: propia.

El análisis de la variable correspondiente al número de días de estancia en hospitalización durante el periodo enero 2013 y diciembre 2014, es relevante al contemplar que es denominada la variable de interés para la investigación. En la tabla que se presenta a continuación (Ver Tabla 13) se destacan las estadísticas descriptivas obtenidas con todos los datos de la población, observándose que en promedio los pacientes permanecen 14 días aproximadamente en hospitalización; sin embargo, la desviación estándar para dichos resultados arroja un valor de 10,828, de tal forma que el coeficiente de variación de Pearson indica que existe heterogeneidad en los datos, al presentar un valor de 76,63%; de tal forma que, estadísticamente existe una alta dispersión entre los días de estancia en hospitalización. Los días de hospitalización presentan datos extremos entre 1 y 122; sin embargo, existe una concentración elevada de los resultados alrededor de la media, la cual se describe como leptocúrtica, debido a que el valor de la curtosis es de 11,47; además, los datos presentan una asimetría a la derecha, indicando que los mismos se concentran a la izquierda del promedio. Por otra parte, se observa como la media recortada al 5% se encuentra más cerca a la mediana que a la media aritmética, demostrando la presencia de valores extremos. En esta relación, en el diagrama de caja (Ver figura 16) podemos observar como los valores extremos se encuentran a partir de 38 días de estancia.

ITEM	ESTADÍSTICO DESCRIPTIVO	ERROR ESTANDAR
Media	14,13	0,259
Media recortada al 5%	13,18	
Mediana	13	
Varianza	117,242	
Desviación estándar	10,828	
Mínimo	1	
Máximo	122	
Rango	121	
Rango intercuartil	12	
Asimetría	2,192	0,058
Curtosis	11,47	0,117

Tabla 13. Estadísticos descriptivos de los días estancia de las hospitalizaciones registradas entre 2013 y 2014 por TAB en la CSDJM. Fuente: propia.

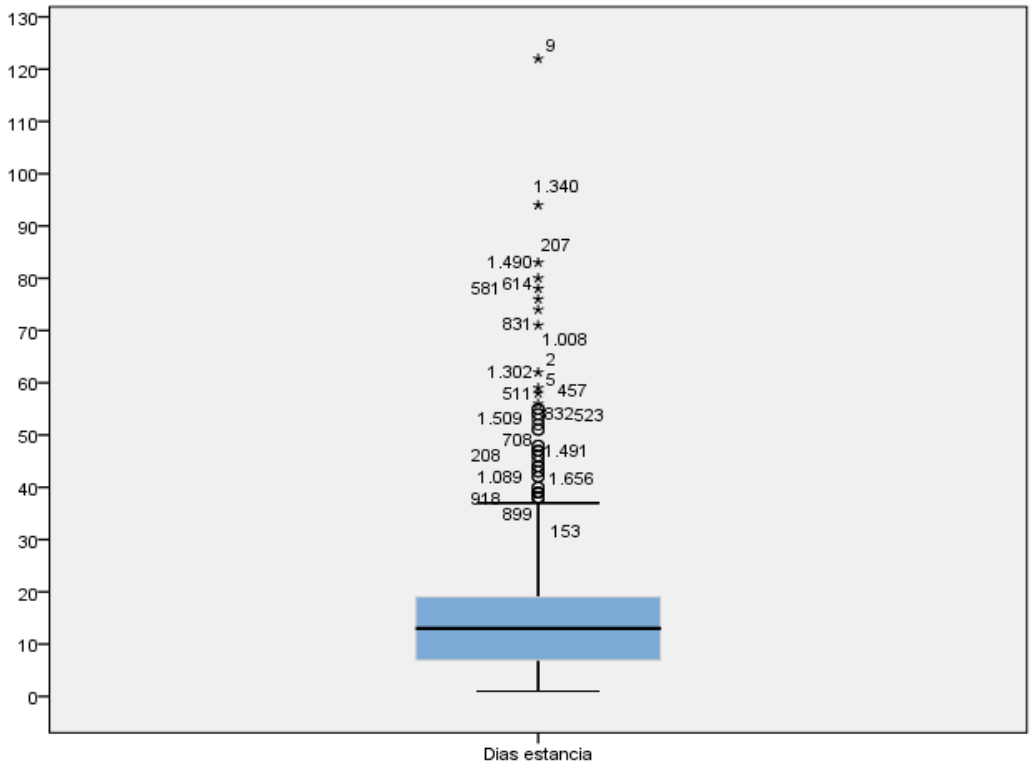


Figura 16. Diagrama de caja y valores atípicos extremos de los días de estancia de las hospitalizaciones registradas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

Reingresos hospitalarios

Desde otra perspectiva, las 1753 hospitalizaciones se presentaron en 1191 pacientes, lo que equivale a decir que un 67.9% del total de las hospitalizaciones fueron realizadas en pacientes que no requirieron reingresos durante el tiempo de estudio, mientras que el 32.1% de las hospitalizaciones corresponden a reingresos de pacientes previamente hospitalizados, de los cuales registraron principalmente un reingreso. Al desagregar estos datos por año se evidenció que para el año 2013 el 82% (524) de los pacientes hospitalizados presentaron reingresos, mientras un 18% (115) presentó principalmente un reingreso. Para el año 2014 los reingresos ascendieron a un 40.2% (447) y los pacientes hospitalizados sin reingreso alcanzaron el 59.8% (667) (Ver Figuras 17 y 18).

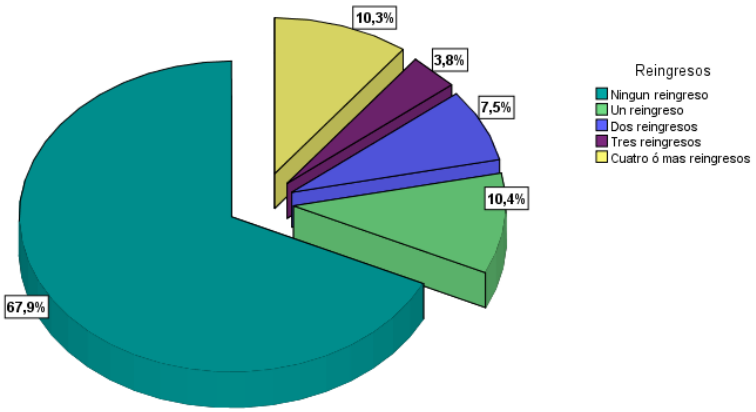


Figura 17. Pacientes hospitalizados entre 2013 y 2014 por TAB según reingresos en la CSJDM. Fuente: propia.

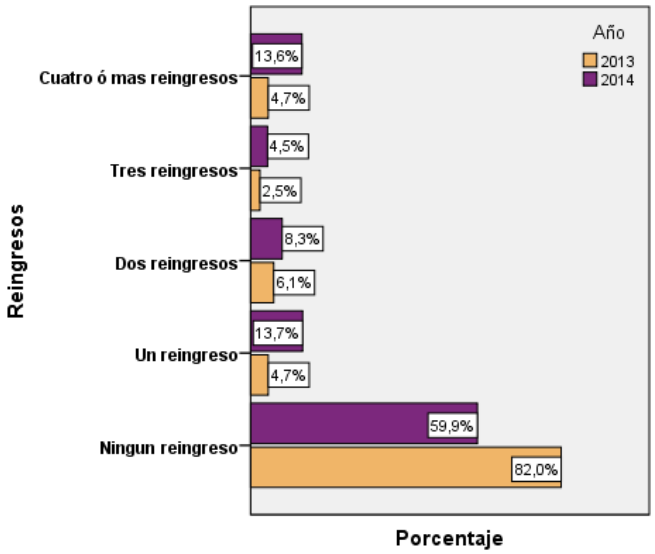


Figura 18. Pacientes hospitalizados entre 2013 y 2014 por TAB, según reingresos y año de ocurrencia del reingreso en la CSJDM. Fuente: Propia.

Descripción de los pacientes hospitalizados entre 2013 y 2014

A diferencia del análisis anterior que describe la cantidad de eventos (1753) de hospitalizaciones durante el periodo comprendido por el estudio, esta sección se centra en la caracterización de los pacientes atendidos durante el periodo de estudio, es decir, la descripción de los pacientes que fueron hospitalizados (1191), en donde cada uno de ellos debe ser tomado como un individuo único, sin reapariciones en las hospitalizaciones, en este sentido, de los 1191 pacientes hospitalizados, el 44% (524) fueron hospitalizados en el 2013 y el 56% (667) durante el 2014, presentándose un incremento de 27.3% de pacientes nuevos hospitalizados en el 2014.

En la figura 19 se puede apreciar el comportamiento del número de pacientes hospitalizados por mes; evidenciándose que 52 corresponde al dato promedio obtenido para los dos años de estudio, dicho valor fue afectado principalmente por la cantidad de ingresos presentados en el año 2014, donde la mayor cantidad de pacientes hospitalizados se evidenció en los periodos comprendidos entre marzo y mayo y julio a noviembre del mismo año; se destaca que en el mes de noviembre de 2014 el número de casos ascendió a 71, siendo la cifra más elevada en los dos años que se consideraron para el presente estudio. Con el fin de determinar la existencia de diferencias entre las medias de la cantidad de pacientes hospitalizados por TAB (Hipótesis alternativa) en los años 2013 y 2014, se contrarrestó dicho supuesto frente a la hipótesis nula de diferencia entre medias igual a cero, encontrándose que, con un nivel de confianza del 95%, el estadístico t toma el valor de -2,08, lo que sugiere un valor-p de 0,049, el cual al ser inferior a 0,05 permite el rechazo de la hipótesis nula y con esto, la determinación de que las medias son diferentes. La variación de los datos con respecto al promedio para los dos años analizados, es comparada mediante el coeficiente de variación de Pearson (CV), el cual arroja un valor de 34,9% y 22,7% para los años 2013 y 2014, respectivamente; dichos valores obtenidos a partir de la medias 43,67 y 55,58 y las desviaciones estándar 15,24 y 12,63 para cada conjunto de datos; los resultados refieren que la variabilidad de los datos con respecto al promedio es mayor para los resultados del año 2013; es decir, que el número de pacientes hospitalizados presenta mayor fluctuación con relación a la media.

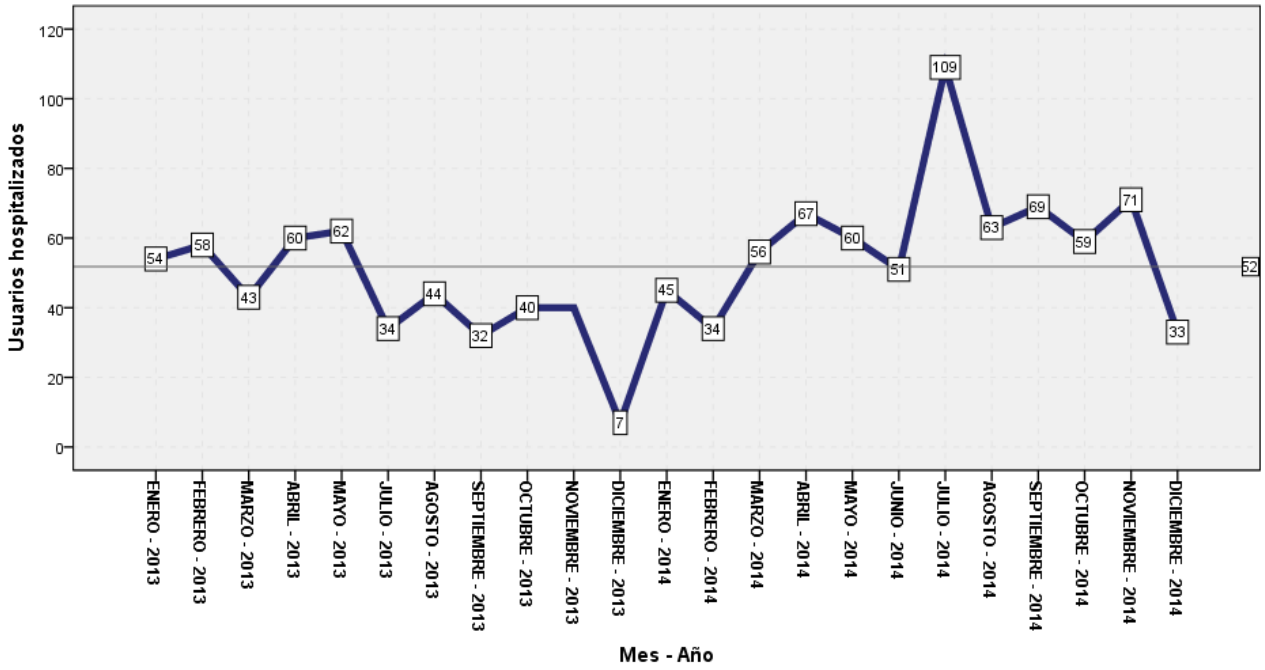


Figura 19. Cantidad de pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia

	Recuento	Promedio	Desviación Estándar	Coeficiente de Variación	Mínimo	Máximo	Rango
2013	12	43,6667	15,2455	0,349133	7	62	55
2014	12	55,5833	12,6308	0,227241	33	71	38

Tabla 14. Resumen estadístico para cantidad de pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia

Distribución porcentual del TAB

Estos pacientes hospitalizados en la CSJDM provienen en el 99% de los casos del departamento de Caldas, y más de la mitad, un 59% de ellos provienen del municipio de Manizales, capital del departamento de Caldas. Los demás municipios en su totalidad aportaron 41% de los casos (Ver Tabla 15). Vistos estos datos, desde la perspectiva de la población de cada municipio el panorama fue muy diferente, el municipio de Aranzazu presentó la mayor proporción de pacientes hospitalizados con 29 por cada 10000 habitantes, mientras Manizales y Filadelfia alcanzaron 17 hospitalizados por cada 10000 habitantes, cada uno. Se destaca que en el municipio de San José no se presentaron casos de pacientes hospitalizados con algún diagnóstico asociado al TAB. (Ver Figura 20.)

MUNICIPIO	Frecuencia de Casos	% De Casos	Población Proyectada*	Proporción de pacientes**
ARANZAZU	33	3	11566	29
MANIZALES	706	59	394627	18
FILADELFIA	19	2	11200	17
SALAMINA	25	2	16968	15
MARULANDA	5	0	3410	15
NEIRA	38	3	30285	13
LA MERCED	7	1	5623	12
VILLAMARIA	68	6	55228	12
CHINCHINÁ	53	4	51696	10
ANSERMA	34	3	33920	10
PÁCORA	12	1	12244	10
BELALCAZAR	10	1	10960	9
MARQUETALIA	13	1	14982	9
RISARALDA	8	1	9693	8
MANZANARES	19	2	23447	8
SUPIA	20	2	26542	8
VITERBO	8	1	12506	6
AGUADAS	14	1	22293	6
PENSILVANIA	16	1	26360	6
VICTORIA	4	0	8505	5
MARMATO	4	0	9026	4
SAMANÁ	11	1	25769	4
PALESTINA	7	1	17795	4
RIOSUCIO	23	2	60798	4
NORCASIA	2	0	6430	3
LA DORADA	19	2	76574	2
FUERA DEL DEPTO CALDAS	13			
TOTAL	1191		978447	12

*** Estimaciones de población 1985 - 2005 y proyecciones de población 2005 - 2020 total departamental por área [DANE, 2015]**

**** Proporción de pacientes cada 10,000 habitantes**

Tabla 15. Distribución porcentual y concentración del TAB en el departamento de Caldas según pacientes hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

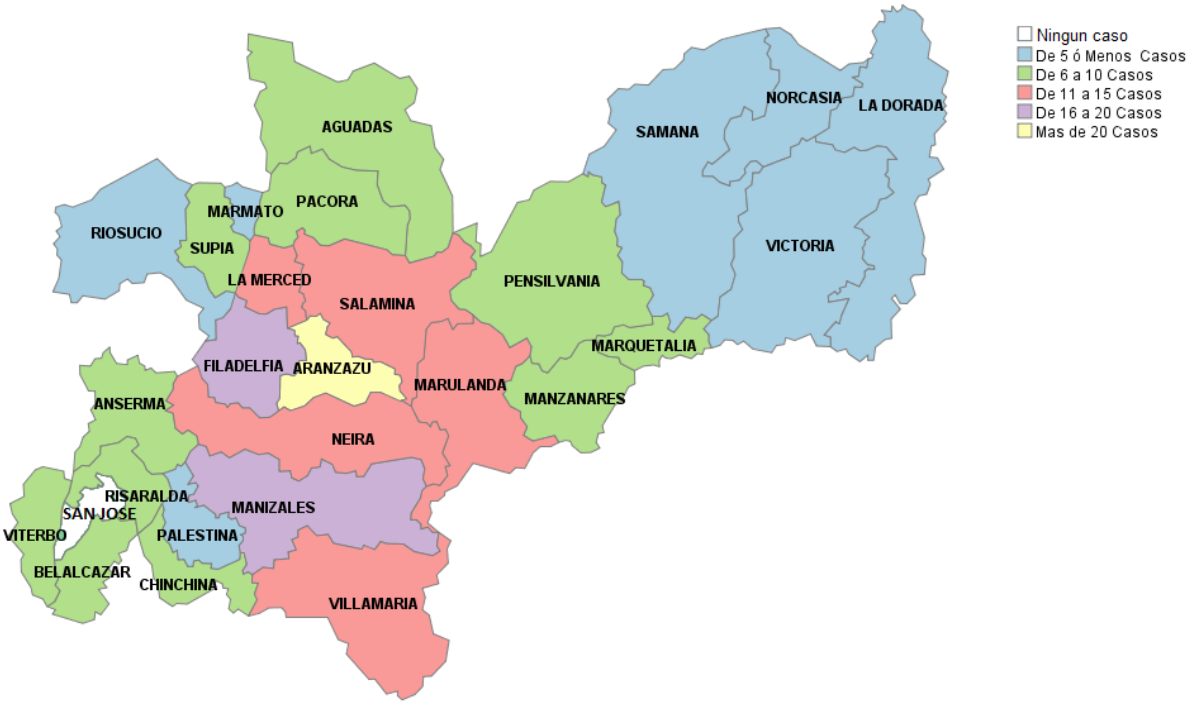


Figura 20. Distribución espacial en el departamento de Caldas de la proporción de pacientes hospitalizados con TAB entre 2013 y 2014 en la CSJDM. Fuente: propia.

Análisis de la edad

En cuanto a la edad de los pacientes hospitalizados se encontró una media aritmética de 41.43 años, con una desviación estándar de 18.51, lo que sugiere una baja concentración de los datos, al presentar un coeficiente de variación de Pearson de 44,69%, destacando que los datos extremos de la población se sitúan entre 4 y 87 años; en la tabla 16 y en el figura 21 se puede apreciar como la distribución de la edad presenta una leve inclinación de los datos hacia la izquierda del promedio; sin embargo debido a su proximidad al cero y al error de la asimetría, su comportamiento permite que tanto la mediana como la media se ubiquen en un punto medio muy próximo. Entre tanto, se observa como la media recortada al 5% se encuentra cerca de la mediana y la media aritmética, demostrando la no presencia de valores extremos.

Cabe resaltar, que al describir espacialmente la edad en grupos quinquenales, se observó como en los municipios de Anserma, Palestina y Villamaria predominaron los pacientes en el grupo de edad de 11 a 20 años, mientras en el municipio de Risaralda predominó el grupo de pacientes con una edad superior a 70 años (Ver Figura 22).

EDAD		
ITEM	**ESTADÍSTICO DESCRIPTIVO**	**ERROR ESTÁNDAR**
Media	41,43	0,537
Media recortada al 5%	41,06	
Mediana	42	
Varianza	342,948	
Desviación estándar	18,519	
Mínimo	4	
Máximo	87	
Rango	83	
Rango intercuartil	31	
Asimetría	0,12	0,071
Curtosis	-0,881	0,142

Tabla 16. Estadísticos descriptivos de la edad de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

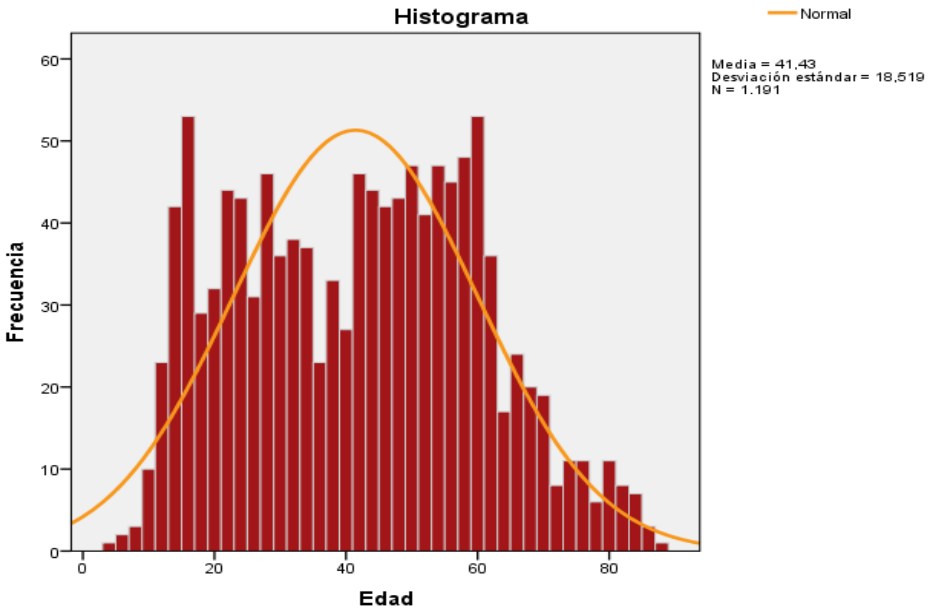


Figura 21. Distribución de la edad de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

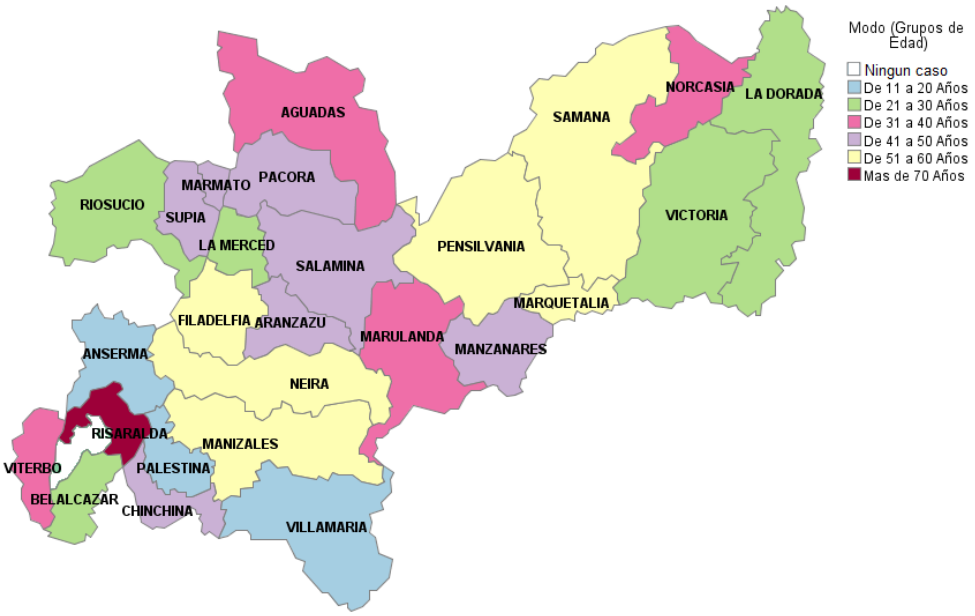


Figura 22. Distribución espacial en el departamento de Caldas de los grupos de edad de pacientes hospitalizados entre 2013 y 2014 TAB en la CSJDM. Fuente: propia.

Análisis del género

En cuanto al género de estos pacientes hospitalizados, se registró que el 64.23% eran del género femenino, mientras que un 35.77% eran del género masculino, Nótese que en la totalidad de los municipios del departamento de Caldas, el género predominante fue el femenino, a excepción del municipio de Belalcazar en donde predominó el género masculino, así mismo, es de recordar que el municipio de San José no registro pacientes hospitalizados por TAB (Ver Figuras 23 y 24).

Figura 23. Distribución espacial en el departamento de Caldas del género de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

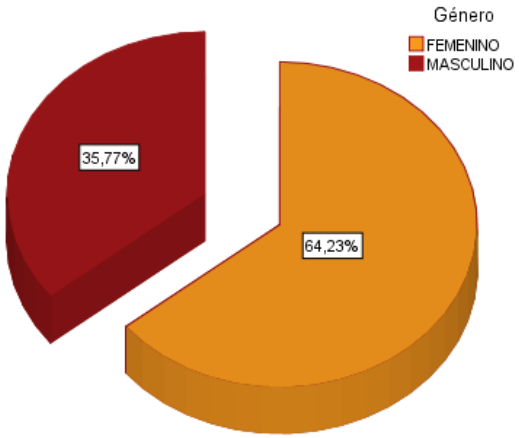


Figura 24. Distribución porcentual del género de los pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia.

Contraste edad vs género

Al contrastar la edad con el género de los pacientes hospitalizados por TAB, se encontró que la edad media del género femenino fue de 43.05 años con una desviación estándar de 17.3 años; por su parte, el género masculino registró una edad media de 40.91 años con una desviación estándar de 18.6 años. Al analizar los resultados individuales provenientes del coeficiente de variación de Pearson se encontró que para ambos sexos, la dispersión de los valores fue elevada, teniendo en cuenta que los resultados fueron 40,24% y 45,47% para los géneros femenino y masculino, respectivamente. Con el fin de determinar si las medias de los dos géneros eran iguales (Ho) o diferentes (Ha) a un nivel de confianza del 95%, se realizó una prueba de hipótesis basada en el estadístico t, el cual para dichos datos tomó un valor de 2,42 y un valor-p de 0,015, quien al ser inferior al 0,05 permitió rechazar la hipótesis nula, estableciendo que con un 5% de significancia la edad media de las mujeres difirió de la edad media de los hombres.

Los histogramas que se presentan posteriormente evidencian que para ambos géneros existe una leve concentración de los resultados a la izquierda del promedio, por lo tanto el valor de la asimetría toma los valores positivos 0,115 (Género femenino) y 0,165 (Género masculino).

EDAD EN GÉNERO FEMENINO		
ITEM	ESTADÍSTICO DESCRIPTIVO	ERROR ESTÁNDAR
Media	43,05	,512
Media recortada al 5%	41,85	
Mediana	44,00	
Varianza	300,087	
Desviación estándar	17,32	
Mínimo	6	
Máximo	87	
Rango	81	
Rango intercuartil	29	
Asimetría	,115	,088
Curtosis	-,819	,177

Tabla 17. Estadísticos descriptivos de la edad en el género femenino de pacientes hospitalizadas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

EDAD EN GÉNERO MASCULINO		
ITEM	ESTADÍSTICO DESCRIPTIVO	ERROR ESTÁNDAR
Media	40,908	0
Media recortada al 5%	39,54	
Mediana	42,00	
Varianza	346,06	
Desviación estándar	18,6	
Mínimo	4	
Máximo	85	
Rango	81	
Rango intercuartil	33	
Asimetría	,165	,118
Curtosis	-,985	,236

Tabla 18. Estadísticos descriptivos de la edad en el género Masculino de pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

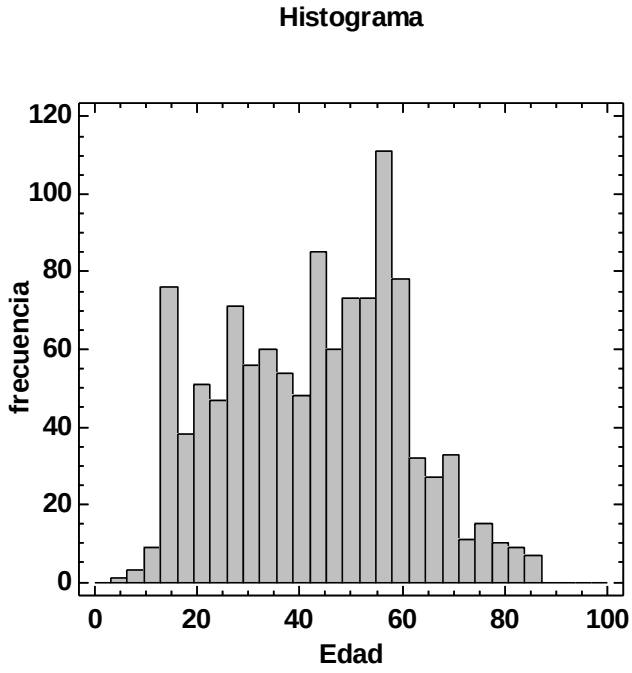


Figura 25. Distribución de la edad del género femenino en pacientes hospitalizadas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

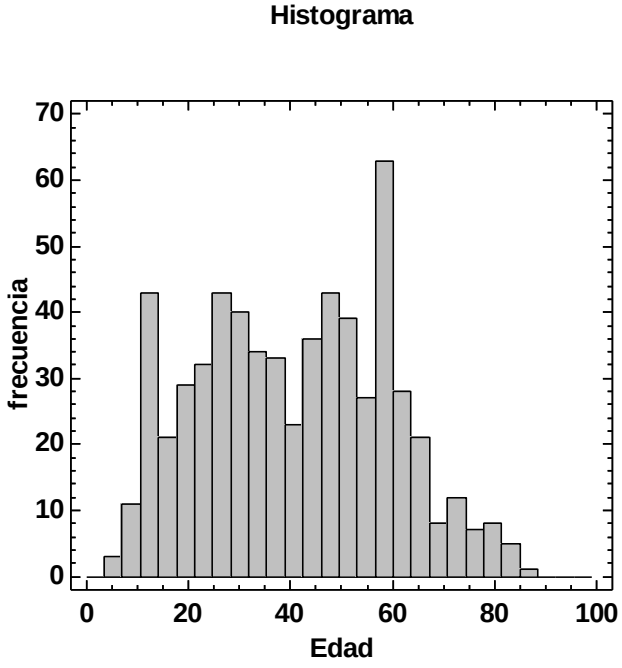


Figura 26. Distribución de la edad del género masculino en pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

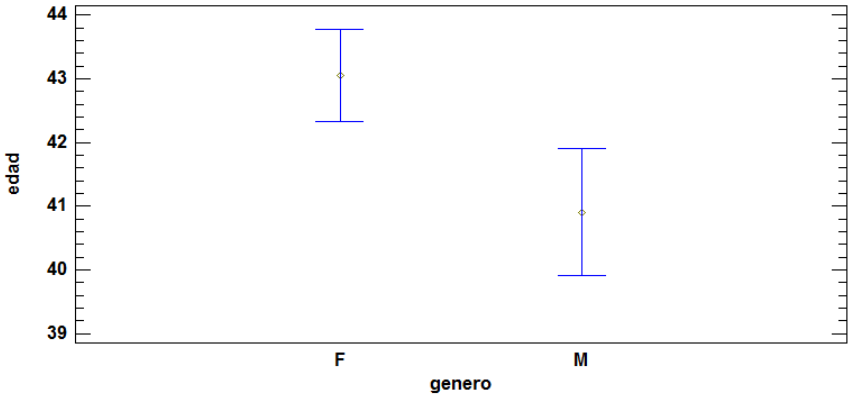


Figura 27. Gráfico de medias para la distribución de la edad por género de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

Análisis sociodemográfico

Al revisar algunas características sociodemográficas de esta población de pacientes hospitalizados por TAB, los resultados de la figura 28 presentaron en el 98.4% de los casos tenía afiliación a algún tipo de seguridad social en salud, siendo el régimen contributivo el de mayor importancia. En cuanto al estrato socioeconómico, un 11% de los pacientes pertenecían al estrato 4 o superiores, mientras el 89% pertenecen al estrato 3 o inferiores, siendo los estratos 2 y 3 los más predominantes (Ver Figura 29). Por su parte, el estado civil de estas personas, presenta a una población principalmente soltera, con el 39%, casados con el 24% y unión libre con el 13% (Ver Figura 30). Finalmente, en cuanto al nivel educativo los datos presentaron a una población que en el 25.3% de los casos no habían terminado el bachillerato, mientras un 13.2% eran profesionales universitarios (Ver Figura 31).

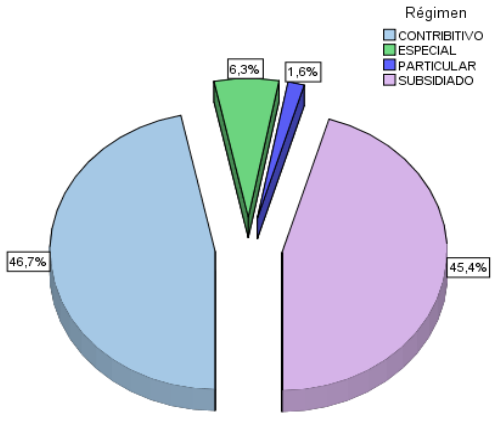


Figura 28. Distribución del régimen de aseguramiento a la seguridad social de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

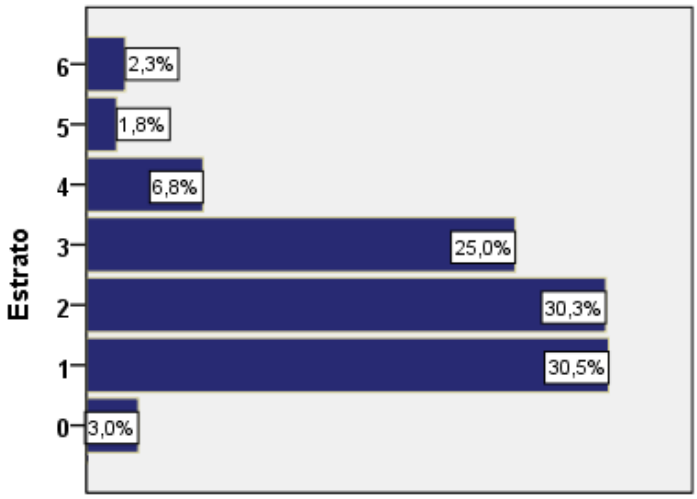


Figura 29. Distribución del estrato socioeconómico de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

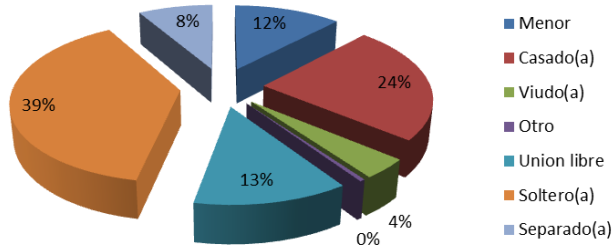


Figura 30. Distribución del estado civil de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

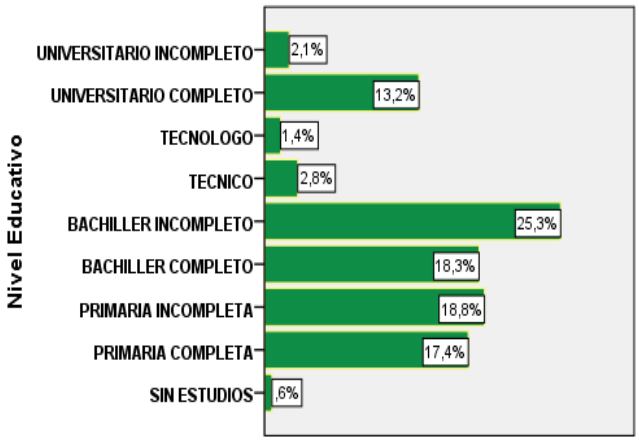


Figura 31. Distribución del nivel educativo de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

Al determinar los días de estancia de acuerdo a estas variables sociodemográficas, se encontró en el caso del régimen de aseguramiento al sistema de seguridad social en salud, que los días estancia del régimen contributivo fueron de 12.41 días con una desviación estándar de 10.3 días; por su parte el régimen subsidiado registró una media de 14.74 días con una desviación estándar de 8.8 días (Ver Tablas 19 y 20); por lo tanto, el régimen contributivo presenta una alta variación de los resultados al obtener un valor en el coeficiente de variación de Pearson de 82,99%, mientras que el régimen subsidiado manifiesta un menor valor, 60,04%, que representa igualmente una baja concentración de los datos; en cualquiera de los casos, los valores obtenidos tienden a concentrarse levemente a la izquierda del promedio, tal como lo sugiere el coeficiente de asimetría, el cual toma valores positivos en los dos régimen analizados.

Ahora bien, para determinar si los días estancia del régimen subsidiado eran mayores que los del régimen contributivo, se realizó la prueba no paramétrica de Mann Whitney para muestras independientes y los resultados presentados en la tabla 21 sugieren que los días estancia del régimen subsidiado fueron mayores que el régimen contributivo, siendo esta diferencia estadísticamente significativa a un valor p de 0.000.

RÉGIMEN CONTRIBUTIVO	ESTADÍSTICO DESCRIPTIVO	ERROR ESTÁNDAR
Media	12,41	,437
Media recortada al 5%	11,46	
Mediana	11,00	
Varianza	106,119	
Desviación estándar	10,301	
Mínimo	1	
Máximo	122	
Rango	121	
Rango intercuartil	10	
Asimetría	3,662	,104
Curtosis	29,072	,207

Tabla 19. Estadísticos descriptivos de los días estancia según régimen contributivo de los pacientes hospitalizadas entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

RÉGIMEN SUBSIDIADO	ESTADÍSTICO DESCRIPTIVO	ERROR ESTÁNDAR
Media	14,74	,381
Media recortada al 5%	14,17	
Mediana	14,00	
Varianza	78,251	
Desviación estándar	8,846	
Mínimo	1	
Máximo	58	
Rango	57	
Rango intercuartil	11	
Asimetría	1,221	,105
Curtosis	2,983	,210

Tabla 20. Estadísticos descriptivos de los días estancia según régimen subsidiado de pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

	Días de estancia
U de Mann-Whitney	121105,000
W de Wilcoxon	275395,000
Z	-5,500
Sig. asintótica (bilateral)	,000

Tabla 21. Prueba de Mann Whitney para muestras independientes de los días estancia y régimen de aseguramiento a la seguridad social en salud de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

La variable nivel educativo fue analizada individualmente con respecto a los días estancia de los pacientes hospitalizados, encontrándose que para un valor p de 0,00, existen diferencias estadísticamente significativas entre la media de los niveles educativos respecto a los días estancia, la tabla 22, presenta la descripción de la variabilidad de los datos para los 28 grupos analizados; mientras en la tabla 23, se detallan los resultados del método de la diferencia mínima significativa (LSD) de Fisher, en donde se describen todos los pares de medias entre los niveles educativos, para lo cual, 86 pares de medias que presentan asterisco en la columna “Significativo” refieren diferencias estadísticamente significativas con un 95% de confianza.

FUENTE	Suma de cuadrados	Gl	Cuadrado medio	Razón-F	Valor-p
Entre grupos	10149,7	27	375,915	3,32	0,0000
Intra grupos	195258,	1725	113,193		
Total (Corr)	205407,	1752			

Tabla 22. Anova para días de estancia por nivel educativo para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

Contraste	*Significativo*	*Diferencia*	*+/- Límites*
ANALFABETA - Decimo bachillerato		1,62109	3,42427
ANALFABETA - EDUCACION_ESPECIAL		15,0309	21,0629
ANALFABETA - Estudiante universidad	*	3,64045	2,95111
ANALFABETA - LICENCIADO		1,91982	7,30191
ANALFABETA - Noveno bachillerato	*	4,34343	3,69801
ANALFABETA - Octavo bachillerato		4,48547	4,94842
ANALFABETA - Once bachillerato		1,17843	2,37166
ANALFABETA - PROFESIONAL		10,0309	21,0629
ANALFABETA - Primero primaria		2,52347	3,32881
ANALFABETA - Segundo primaria		-2,3095	3,72423
ANALFABETA - Septimo bachillerato		4,03093	14,9695
ANALFABETA - Tecnico		-9,96907	10,6914
ANALFABETA - Tecnologo	*	7,94002	4,94842
ANALFABETA - cuarto primaria		1,99584	3,49726
ANALFABETA - desconocido		-0,596979	3,83914
ANALFABETA - jardin		6,03093	14,9695
ANALFABETA - octavi bachilerato		10,0309	21,0629
ANALFABETA - octavo bachillerato	*	4,55593	3,93763
ANALFABETA - pensionado	*	-13,7691	9,60992
ANALFABETA - primero primaria		-0,302405	5,37796
ANALFABETA - profesional	*	5,55507	2,74871
ANALFABETA - quinto primaria		1,20054	2,46549
ANALFABETA - segundo primaria		1,86814	3,83914
ANALFABETA - septimo bachillerato		1,00529	3,97321
ANALFABETA - sexto bachillerato		1,2918	3,30015
ANALFABETA - tecnico	*	4,48093	3,93763
ANALFABETA - tercero primaria		0,969952	3,14358
Decimo bachillerato - EDUCACION_ESPECIAL		13,4098	21,1262
Decimo bachillerato - Estudiante universidad		2,01936	3,37353
Decimo bachillerato - LICENCIADO		0,298725	7,48261
Decimo bachillerato - Noveno bachillerato		2,72234	4,04313
Decimo bachillerato - Octavo bachillerato		2,86438	5,21139
Decimo bachillerato - Once bachillerato		-0,442664	2,88035
Decimo bachillerato - PROFESIONAL		8,40984	21,1262
Decimo bachillerato - Primero primaria		0,902373	3,70846
Decimo bachillerato - Segundo primaria		-3,93059	4,06713
Decimo bachillerato - Septimo bachillerato		2,40984	15,0585
Decimo bachillerato - Tecnico	*	-11,5902	10,8156
Decimo bachillerato - Tecnologo	*	6,31893	5,21139
Decimo bachillerato - cuarto primaria		0,374748	3,86037
Decimo bachillerato - desconocido		-2,21807	4,17261
Decimo bachillerato - jardin		4,40984	15,0585

Decimo bachillerato - octavi bachilerato		8,40984	21,1262
Decimo bachillerato - octavo bachillerato		2,93484	4,2634
Decimo bachillerato - pensionado	*	-15,3902	9,74793
Decimo bachillerato - primero primaria		-1,9235	5,62086
Decimo bachillerato - profesional	*	3,93397	3,19798
Decimo bachillerato - quinto primaria		-0,420553	2,95809
Decimo bachillerato - segundo primaria		0,247045	4,17261
Decimo bachillerato - septimo bachillerato		-0,615805	4,29629
Decimo bachillerato - sexto bachillerato		-0,329294	3,68275
Decimo bachillerato - tecnico		2,85984	4,2634
Decimo bachillerato - tercero primaria		-0,65114	3,54313
EDUCACION_ESPECIAL - Estudiante universidad		-11,3905	21,0547
EDUCACION_ESPECIAL - LICENCIADO		-13,1111	22,0887
EDUCACION_ESPECIAL - Noveno bachillerato		-10,6875	21,1723
EDUCACION_ESPECIAL - Octavo bachillerato		-10,5455	21,4261
EDUCACION_ESPECIAL - Once bachillerato		-13,8525	20,9813
EDUCACION_ESPECIAL - PROFESIONAL		-5,0	29,635
EDUCACION_ESPECIAL - Primero primaria		-12,5075	21,1109
EDUCACION_ESPECIAL - Segundo primaria		-17,3404	21,1769
EDUCACION_ESPECIAL - Septimo bachillerato		-11,0	25,6647
EDUCACION_ESPECIAL - Tecnico	*	-25,0	23,4286
EDUCACION_ESPECIAL - Tecnologo		-7,09091	21,4261
EDUCACION_ESPECIAL - cuarto primaria		-13,0351	21,1382
EDUCACION_ESPECIAL - desconocido		-15,6279	21,1974
EDUCACION_ESPECIAL - jardin		-9,0	25,6647
EDUCACION_ESPECIAL - octavi bachilerato		-5,0	29,635
EDUCACION_ESPECIAL - octavo bachillerato		-10,475	21,2155
EDUCACION_ESPECIAL - pensionado	*	-28,8	22,9552
EDUCACION_ESPECIAL - primero primaria		-15,3333	21,5294
EDUCACION_ESPECIAL - profesional		-9,47586	21,0273
EDUCACION_ESPECIAL - quinto primaria		-13,8304	20,9921
EDUCACION_ESPECIAL - segundo primaria		-13,1628	21,1974
EDUCACION_ESPECIAL - septimo bachillerato		-14,0256	21,2221
EDUCACION_ESPECIAL - sexto bachillerato		-13,7391	21,1064
EDUCACION_ESPECIAL - tecnico		-10,55	21,2155
EDUCACION_ESPECIAL - tercero primaria		-14,061	21,0825
Estudiante universidad - LICENCIADO		-1,72063	7,27825
Estudiante universidad - Noveno bachillerato		0,702976	3,65108
Estudiante universidad - Octavo bachillerato		0,845022	4,91345
Estudiante universidad - Once bachillerato	*	-2,46202	2,2978
Estudiante universidad - PROFESIONAL		6,39048	21,0547
Estudiante universidad - Primero primaria		-1,11699	3,27659
Estudiante universidad - Segundo primaria	*	-5,94995	3,67764
Estudiante universidad - Septimo bachillerato		0,390476	14,958
Estudiante universidad - Tecnico	*	-13,6095	10,6753
Estudiante universidad - Tecnologo		4,29957	4,91345
Estudiante universidad - cuarto primaria		-1,64461	3,44759
Estudiante universidad - desconocido	*	-4,23743	3,79396
Estudiante universidad - jardin		2,39048	14,958
Estudiante universidad - octavi bachilerato		6,39048	21,0547
Estudiante universidad - octavo bachillerato		0,915476	3,89359
Estudiante universidad - pensionado	*	-17,4095	9,59196
Estudiante universidad - primero primaria		-3,94286	5,34579
Estudiante universidad - profesional		1,91461	2,68523
Estudiante universidad - quinto primaria	*	-2,43991	2,39452
Estudiante universidad - segundo primaria		-1,77231	3,79396
Estudiante universidad - septimo bachillerato		-2,63516	3,92957

Estudiante universidad - sexto bachillerato		-2,34865	3,24748
Estudiante universidad - tecnico		0,840476	3,89359
Estudiante universidad - tercero primaria		-2,6705	3,08823
LICENCIADO - Noveno bachillerato		2,42361	7,61178
LICENCIADO - Octavo bachillerato		2,56566	8,29161
LICENCIADO - Once bachillerato		-0,741389	7,06319
LICENCIADO - PROFESIONAL		8,11111	22,0887
LICENCIADO - Primero primaria		0,603648	7,43941
LICENCIADO - Segundo primaria		-4,22931	7,62455
LICENCIADO - Septimo bachillerato		2,11111	16,3814
LICENCIADO - Tecnico		-11,8889	12,5925
LICENCIADO - Tecnologo		6,0202	8,29161
LICENCIADO - cuarto primaria		0,0760234	7,51629
LICENCIADO - desconocido		-2,5168	7,68133
LICENCIADO - jardin		4,11111	16,3814
LICENCIADO - octavi bachilerato		8,11111	22,0887
LICENCIADO - octavo bachillerato		2,63611	7,73103
LICENCIADO - pensionado	*	-15,6889	11,6882
LICENCIADO - primero primaria		-2,22222	8,5549
LICENCIADO - profesional		3,63525	7,19856
LICENCIADO - quinto primaria		-0,719278	7,09525
LICENCIADO - segundo primaria		-0,0516796	7,68133
LICENCIADO - septimo bachillerato		-0,91453	7,74921
LICENCIADO - sexto bachillerato		-0,628019	7,42663
LICENCIADO - tecnico		2,56111	7,73103
LICENCIADO - tercero primaria		-0,949864	7,35839
Noveno bachillerato - Octavo bachillerato		0,142045	5,3952
Noveno bachillerato - Once bachillerato		-3,165	3,20095
Noveno bachillerato - PROFESIONAL		5,6875	21,1723
Noveno bachillerato - Primero primaria		-1,81996	3,96261
Noveno bachillerato - Segundo primaria	*	-6,65293	4,30014
Noveno bachillerato - Septimo bachillerato		-0,3125	15,1231
Noveno bachillerato - Tecnico	*	-14,3125	10,9054
Noveno bachillerato - Tecnologo		3,59659	5,3952
Noveno bachillerato - cuarto primaria		-2,34759	4,10513
Noveno bachillerato - desconocido	*	-4,94041	4,40004
Noveno bachillerato - jardin		1,6875	15,1231
Noveno bachillerato - octavi bachilerato		5,6875	21,1723
Noveno bachillerato - octavo bachillerato		0,2125	4,48623
Noveno bachillerato - pensionado	*	-18,1125	9,84743
Noveno bachillerato - primero primaria		-4,64583	5,79169
Noveno bachillerato - profesional		1,21164	3,48951
Noveno bachillerato - quinto primaria		-3,14289	3,27108
Noveno bachillerato - segundo primaria		-2,47529	4,40004
Noveno bachillerato - septimo bachillerato		-3,33814	4,51749
Noveno bachillerato - sexto bachillerato		-3,05163	3,93857
Noveno bachillerato - tecnico		0,1375	4,48623
Noveno bachillerato - tercero primaria		-3,37348	3,80833
Octavo bachillerato - Once bachillerato		-3,30705	4,58887
Octavo bachillerato - PROFESIONAL		5,54545	21,4261
Octavo bachillerato - Primero primaria		-1,96201	5,14916
Octavo bachillerato - Segundo primaria	*	-6,79497	5,41321
Octavo bachillerato - Septimo bachillerato		-0,454545	15,4764
Octavo bachillerato - Tecnico	*	-14,4545	11,3903
Octavo bachillerato - Tecnologo		3,45455	6,31821
Octavo bachillerato - cuarto primaria		-2,48963	5,25963
Octavo bachillerato - desconocido		-5,08245	5,4929

Octavo bachillerato - jardin		1,54545	15,4764
Octavo bachillerato - octavi bachilerato		5,54545	21,4261
Octavo bachillerato - octavo bachillerato		0,0704545	5,56218
Octavo bachillerato - pensionado	*	-18,2545	10,3819
Octavo bachillerato - primero primaria		-4,78788	6,65998
Octavo bachillerato - profesional		1,06959	4,79461
Octavo bachillerato - quinto primaria		-3,28493	4,63805
Octavo bachillerato - segundo primaria		-2,61734	5,4929
Octavo bachillerato - septimo bachillerato		-3,48019	5,58743
Octavo bachillerato - sexto bachillerato		-3,19368	5,13068
Octavo bachillerato - tecnico		-0,00454545	5,56218
Octavo bachillerato - tercero primaria		-3,51552	5,0314
Once bachillerato - PROFESIONAL		8,8525	20,9813
Once bachillerato - Primero primaria		1,34504	2,76619
Once bachillerato - Segundo primaria	*	-3,48793	3,23121
Once bachillerato - Septimo bachillerato		2,8525	14,8545
Once bachillerato - Tecnico	*	-11,1475	10,5298
Once bachillerato - Tecnologo	*	6,76159	4,58887
Once bachillerato - cuarto primaria		0,817412	2,96675
Once bachillerato - desconocido		-1,77541	3,36301
Once bachillerato - jardin		4,8525	14,8545
Once bachillerato - octavi bachilerato		8,8525	20,9813
Once bachillerato - octavo bachillerato		3,3775	3,47502
Once bachillerato - pensionado	*	-14,9475	9,42981
Once bachillerato - primero primaria		-1,48083	5,04908
Once bachillerato - profesional	*	4,37664	2,0313
Once bachillerato - quinto primaria		0,0221113	1,62771
Once bachillerato - segundo primaria		0,689709	3,36301
Once bachillerato - septimo bachillerato		-0,173141	3,51528
Once bachillerato - sexto bachillerato		0,11337	2,73163
Once bachillerato - tecnico		3,3025	3,47502
Once bachillerato - tercero primaria		-0,208476	2,54025
PROFESIONAL - Primero primaria		-7,50746	21,1109
PROFESIONAL - Segundo primaria		-12,3404	21,1769
PROFESIONAL - Septimo bachillerato		-6,0	25,6647
PROFESIONAL - Tecnico		-20,0	23,4286
PROFESIONAL - Tecnologo		-2,09091	21,4261
PROFESIONAL - cuarto primaria		-8,03509	21,1382
PROFESIONAL - desconocido		-10,6279	21,1974
PROFESIONAL - jardin		-4,0	25,6647
PROFESIONAL - octavi bachilerato		0	29,635
PROFESIONAL - octavo bachillerato		-5,475	21,2155
PROFESIONAL - pensionado	*	-23,8	22,9552
PROFESIONAL - primero primaria		-10,3333	21,5294
PROFESIONAL - profesional		-4,47586	21,0273
PROFESIONAL - quinto primaria		-8,83039	20,9921
PROFESIONAL - segundo primaria		-8,16279	21,1974
PROFESIONAL - septimo bachillerato		-9,02564	21,2221
PROFESIONAL - sexto bachillerato		-8,73913	21,1064
PROFESIONAL - tecnico		-5,55	21,2155
PROFESIONAL - tercero primaria		-9,06098	21,0825
Primero primaria - Segundo primaria	*	-4,83296	3,98709
Primero primaria - Septimo bachillerato		1,50746	15,037
Primero primaria - Tecnico	*	-12,4925	10,7858
Primero primaria - Tecnologo	*	5,41655	5,14916
Primero primaria - cuarto primaria		-0,527625	3,77595
Primero primaria - desconocido		-3,12044	4,09463

Primero primaria - jardin		3,50746	15,037
Primero primaria - octavi bachilerato		7,50746	21,1109
Primero primaria - octavo bachillerato		2,03246	4,18712
Primero primaria - pensionado	*	-16,2925	9,71481
Primero primaria - primero primaria		-2,82587	5,56322
Primero primaria - profesional		3,0316	3,09554
Primero primaria - quinto primaria		-1,32293	2,84704
Primero primaria - segundo primaria		-0,655328	4,09463
Primero primaria - septimo bachillerato		-1,51818	4,2206
Primero primaria - sexto bachillerato		-1,23167	3,59416
Primero primaria - tecnico		1,95746	4,18712
Primero primaria - tercero primaria		-1,55351	3,45095
Segundo primaria - Septimo bachillerato		6,34043	15,1295
Segundo primaria - Tecnico		-7,65957	10,9143
Segundo primaria - Tecnologo	*	10,2495	5,41321
Segundo primaria - cuarto primaria	*	4,30534	4,12877
Segundo primaria - desconocido		1,71252	4,4221
Segundo primaria - jardin		8,34043	15,1295
Segundo primaria - octavi bachilerato		12,3404	21,1769
Segundo primaria - octavo bachillerato	*	6,86543	4,50787
Segundo primaria - pensionado	*	-11,4596	9,85731
Segundo primaria - primero primaria		2,00709	5,80847
Segundo primaria - profesional	*	7,86456	3,51729
Segundo primaria - quinto primaria	*	3,51004	3,30069
Segundo primaria - segundo primaria		4,17763	4,4221
Segundo primaria - septimo bachillerato		3,31478	4,53898
Segundo primaria - sexto bachillerato		3,6013	3,9632
Segundo primaria - tecnico	*	6,79043	4,50787
Segundo primaria - tercero primaria		3,27945	3,8338
Septimo bachillerato - Tecnico		-14,0	18,1477
Septimo bachillerato - Tecnologo		3,90909	15,4764
Septimo bachillerato - cuarto primaria		-2,03509	15,0752
Septimo bachillerato - desconocido		-4,62791	15,1582
Septimo bachillerato - jardin		2,0	20,9551
Septimo bachillerato - octavi bachilerato		6,0	25,6647
Septimo bachillerato - octavo bachillerato		0,525	15,1834
Septimo bachillerato - pensionado	*	-17,8	17,5323
Septimo bachillerato - primero primaria		-4,33333	15,619
Septimo bachillerato - profesional		1,52414	14,9194
Septimo bachillerato - quinto primaria		-2,83039	14,8698
Septimo bachillerato - segundo primaria		-2,16279	15,1582
Septimo bachillerato - septimo bachillerato		-3,02564	15,1927
Septimo bachillerato - sexto bachillerato		-2,73913	15,0307
Septimo bachillerato - tecnico		0,45	15,1834
Septimo bachillerato - tercero primaria		-3,06098	14,9971
Tecnico - Tecnologo	*	17,9091	11,3903
Tecnico - cuarto primaria	*	11,9649	10,839
Tecnico - desconocido		9,37209	10,9541
Tecnico - jardin		16,0	18,1477
Tecnico - octavi bachilerato		20,0	23,4286
Tecnico - octavo bachillerato	*	14,525	10,989
Tecnico - pensionado		-3,8	14,0571
Tecnico - primero primaria		9,66667	11,5834
Tecnico - profesional	*	15,5241	10,6211
Tecnico - quinto primaria	*	11,1696	10,5514
Tecnico - segundo primaria	*	11,8372	10,9541
Tecnico - septimo bachillerato		10,9744	11,0018

Tecnico - sexto bachillerato	*	11,2609	10,777
Tecnico - tecnico	*	14,45	10,989
Tecnico - tercero primaria	*	10,939	10,7301
Tecnologo - cuarto primaria	*	-5,94418	5,25963
Tecnologo - desconocido	*	-8,537	5,4929
Tecnologo - jardin		-1,90909	15,4764
Tecnologo - octavi bachilerato		2,09091	21,4261
Tecnologo - octavo bachillerato		-3,38409	5,56218
Tecnologo - pensionado	*	-21,7091	10,3819
Tecnologo - primero primaria	*	-8,24242	6,65998
Tecnologo - profesional		-2,38495	4,79461
Tecnologo - quinto primaria	*	-6,73948	4,63805
Tecnologo - segundo primaria	*	-6,07188	5,4929
Tecnologo - septimo bachillerato	*	-6,93473	5,58743
Tecnologo - sexto bachillerato	*	-6,64822	5,13068
Tecnologo - tecnico		-3,45909	5,56218
Tecnologo - tercero primaria	*	-6,97007	5,0314
cuarto primaria - desconocido		-2,59282	4,23271
cuarto primaria - jardin		4,03509	15,0752
cuarto primaria - octavi bachilerato		8,03509	21,1382
cuarto primaria - octavo bachillerato		2,56009	4,32224
cuarto primaria - pensionado	*	-15,7649	9,77381
cuarto primaria - primero primaria		-2,29825	5,66562
cuarto primaria - profesional	*	3,55923	3,27601
cuarto primaria - quinto primaria		-0,795301	3,04228
cuarto primaria - segundo primaria		-0,127703	4,23271
cuarto primaria - septimo bachillerato		-0,990553	4,35468
cuarto primaria - sexto bachillerato		-0,704043	3,75071
cuarto primaria - tecnico		2,48509	4,32224
cuarto primaria - tercero primaria		-1,02589	3,61371
desconocido - jardin		6,62791	15,1582
desconocido - octavi bachilerato		10,6279	21,1974
desconocido - octavo bachillerato	*	5,15291	4,60326
desconocido - pensionado	*	-13,1721	9,90129
desconocido - primero primaria		0,294574	5,88281
desconocido - profesional	*	6,15204	3,63874
desconocido - quinto primaria		1,79752	3,42982
desconocido - segundo primaria		2,46512	4,5193
desconocido - septimo bachillerato		1,60227	4,63373
desconocido - sexto bachillerato		1,88878	4,07137
desconocido - tecnico	*	5,07791	4,60326
desconocido - tercero primaria		1,56693	3,94552
jardin - octavi bachilerato		4,0	25,6647
jardin - octavo bachillerato		-1,475	15,1834
jardin - pensionado	*	-19,8	17,5323
jardin - primero primaria		-6,33333	15,619
jardin - profesional		-0,475862	14,9194
jardin - quinto primaria		-4,83039	14,8698
jardin - segundo primaria		-4,16279	15,1582
jardin - septimo bachillerato		-5,02564	15,1927
jardin - sexto bachillerato		-4,73913	15,0307
jardin - tecnico		-1,55	15,1834
jardin - tercero primaria		-5,06098	14,9971
octavi bachilerato - octavo bachillerato		-5,475	21,2155
octavi bachilerato - pensionado	*	-23,8	22,9552
octavi bachilerato - primero primaria		-10,3333	21,5294
octavi bachilerato - profesional		-4,47586	21,0273

octavi bachilerato - quinto primaria		-8,83039	20,9921
octavi bachilerato - segundo primaria		-8,16279	21,1974
octavi bachilerato - septimo bachillerato		-9,02564	21,2221
octavi bachilerato - sexto bachillerato		-8,73913	21,1064
octavi bachilerato - tecnico		-5,55	21,2155
octavi bachilerato - tercero primaria		-9,06098	21,0825
octavo bachillerato - pensionado	*	-18,325	9,93989
octavo bachillerato - primero primaria		-4,85833	5,94755
octavo bachillerato - profesional		0,999138	3,7425
octavo bachillerato - quinto primaria		-3,35539	3,53972
octavo bachillerato - segundo primaria		-2,68779	4,60326
octavo bachillerato - septimo bachillerato		-3,55064	4,71565
octavo bachillerato - sexto bachillerato		-3,26413	4,16437
octavo bachillerato - tecnico		-0,075	4,68571
octavo bachillerato - tercero primaria		-3,58598	4,04142
pensionado - primero primaria	*	13,4667	10,5933
pensionado - profesional	*	19,3241	9,53163
pensionado - quinto primaria	*	14,9696	9,45385
pensionado - segundo primaria	*	15,6372	9,90129
pensionado - septimo bachillerato	*	14,7744	9,95404
pensionado - sexto bachillerato	*	15,0609	9,70503
pensionado - tecnico	*	18,25	9,93989
pensionado - tercero primaria	*	14,739	9,65291
primero primaria - profesional	*	5,85747	5,23678
primero primaria - quinto primaria		1,50294	5,09383
primero primaria - segundo primaria		2,17054	5,88281
primero primaria - septimo bachillerato		1,30769	5,97117
primero primaria - sexto bachillerato		1,5942	5,54612
primero primaria - tecnico		4,78333	5,94755
primero primaria - tercero primaria		1,27236	5,4544
profesional - quinto primaria	*	-4,35453	2,14011
profesional - segundo primaria	*	-3,68693	3,63874
profesional - septimo bachillerato	*	-4,54978	3,77992
profesional - sexto bachillerato	*	-4,26327	3,06471
profesional - tecnico		-1,07414	3,7425
profesional - tercero primaria	*	-4,58511	2,89543
quinto primaria - segundo primaria		0,667598	3,42982
quinto primaria - septimo bachillerato		-0,195252	3,57926
quinto primaria - sexto bachillerato		0,0912583	2,81348
quinto primaria - tecnico		3,28039	3,53972
quinto primaria - tercero primaria		-0,230587	2,62807
segundo primaria - septimo bachillerato		-0,86285	4,63373
segundo primaria - sexto bachillerato		-0,57634	4,07137
segundo primaria - tecnico		2,61279	4,60326
segundo primaria - tercero primaria		-0,898185	3,94552
septimo bachillerato - sexto bachillerato		0,286511	4,19803
septimo bachillerato - tecnico		3,47564	4,71565
septimo bachillerato - tercero primaria		-0,0353346	4,07609
sexto bachillerato - tecnico		3,18913	4,16437
sexto bachillerato - tercero primaria		-0,321845	3,42332
tecnico - tercero primaria		-3,51098	4,04142

Tabla 23. Detalle de resultados del método de diferencia máxima significativa por nivel educativo para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

Los días de permanencia en hospitalización fueron analizados con respecto al estado civil de los pacientes (Ver Tabla 24), dado que el mayor número de casos pertenece a soltero(a) (39%), casado(a) (24%), viudo (4%) y separado(a) (13%). Se utiliza un análisis de variancia para contrarrestar la hipótesis alternativa de existencia de diferencias estadísticamente significativas frente a la hipótesis nula del caso contrario a un nivel de confianza del 95%; posteriormente una prueba de múltiples rangos (LSD) para identificar los estados civiles que presentaron diferencias significativas de acuerdo a la variable respuesta, días de permanencia en hospitalización. Los resultados de la tabla 25, indican que para un valor-p de 0,0014 existe una diferencia significativa entre la media del estado civil y los días de estancia; posteriormente en la tabla de resultados para la prueba de múltiples rangos (Ver Tabla 26) se identifica que existen diferencias estadísticamente significativas entre los pares de medias que presentan un asterisco situado en la columna “significativo”; en este caso se hicieron evidentes las diferencias entre el estado civil “menor” y los estados civiles “soltero(a)”, “viudo(a)”, “separado (a)” y “casado (a).

Estado civil	Casos	Media
Menor	212.	11,2642
Casado(a)	420	13,6167
Viudo(a)	70	14,0714
Otro	8	14,125
Unión libre	220	14,8455
Soltero(a)	675	14,8622
Separado(a)	148	15,2905

Tabla 24. Detalle de resultados del método de diferencia mínima significativa por nivel educativo para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

FUENTE	Suma de cuadrados	GI	Cuadrado Medio	Razón-F	Valor-p
Entre grupos	2525,93	6	420,988	3,62	0,0014
Intra grupos	202881,	1746	116,198		
Total (Corr)	205407,	1752			

Tabla 25. Anova para días de estancia por estado civil para pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

Contraste	Sig	Diferencia	+ / - Límites
Casado(a) - Menor	*	2,35252	1,77998
Casado(a) - Otro		-0,508333	7,5405
Casado(a) - Separado(a)		-1,67387	2,01961
Casado(a) - Soltero(a)		-1,24556	1,31304
Casado(a) - Unión libre		-1,22879	1,75834
Casado(a) - Viudo(a)		-0,454762	2,72755
Menor - Otro		-2,86085	7,60933
Menor - Separado(a)	*	-4,02639	2,26308
Menor - Soltero(a)	*	-3,59807	1,66337
Menor - Unión libre	*	-3,5813	2,03334
Menor - Viudo(a)		-2,80728	2,91243
Otro - Separado(a)		-1,16554	7,66892
Otro - Soltero(a)		-0,737222	7,51383
Otro - Union libre		-0,720455	7,6043
Otro - Viudo(a)		0,0535714	7,88499
Separado(a) - Soltero(a)		0,428318	1,91763
Separado(a) - Union libre		0,445086	2,2461
Separado(a) - Viudo(a)		1,21911	3,06476
Soltero(a) - Union libre		0,0167677	1,6402
Soltero(a) - Viudo(a)		0,790794	2,65293
Union libre - Viudo(a)		0,774026	2,89926

Tabla 26. Múltiples rangos por estdo civil pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

Entre los diagnósticos relacionados al TAB por los cuales fueron hospitalizados estos pacientes, se destaca que el 27.3% de los casos, obedecieron a trastornos bipolares, episodio actual maníaco con síntomas psicóticos (F31.2 según CIE10); un 13.9% perteneciente a trastornos bipolares, episodio actual depresivo grave sin síntomas psicóticos (F31.4 según CIE10) y un 13.2% a otros tipo de trastornos bipolares (F31.9 según CIE10) (Ver Figura 32).

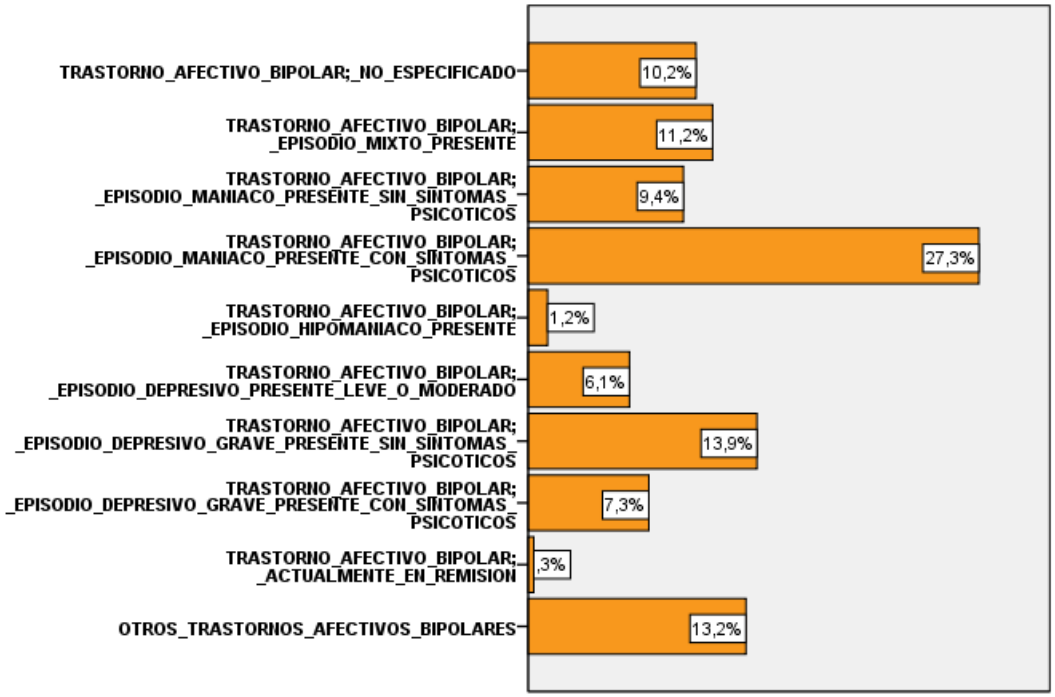


Figura 32. Distribución de los pacientes hospitalizados entre 2013 y 2014 por TAB según el tipo de diagnóstico en la CSJDM. Fuente: propia.

Al desagregar esta información según el género se encontró que en los hombres predominó en Trastorno bipolar, episodio maníaco presente con síntomas psicóticos (F31.2 según CIE10) con un 29.6%; el trastorno bipolar, episodio maníaco presente sin síntomas psicóticos (F31.1 según CIE10) y Otros trastornos bipolares (F31.8 según CIE10). En vista de lo anterior, el género femenino predominó para los demás diagnósticos asociados al TAB (Ver Figura 33).

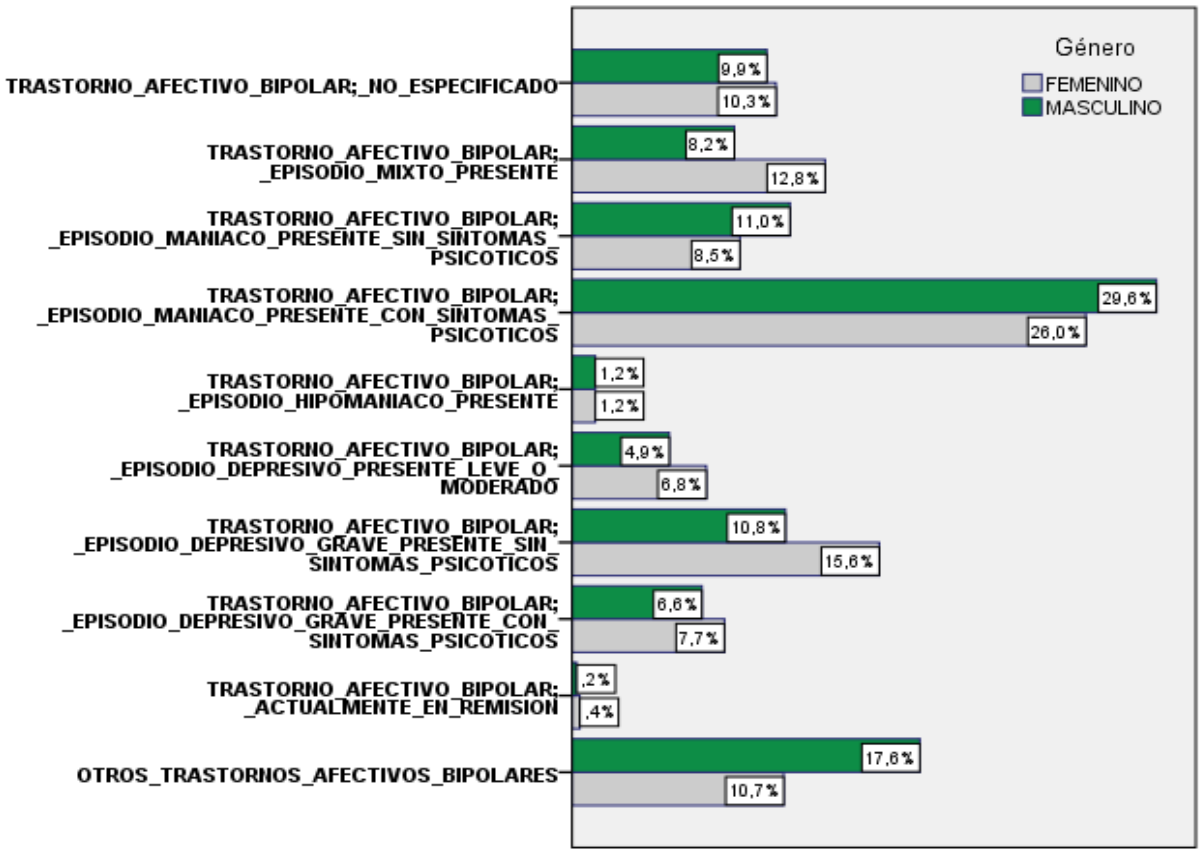


Figura 33. Distribución de los pacientes hospitalizados entre 2013 y 2014 por TAB según el tipo de diagnóstico y género en la CSJDM. Fuente: propia.

Una vez revisado dicho fenómeno desde el punto geográfico, se encontró que en el municipio de Filadelfia predominaron los pacientes con Trastorno bipolar, episodio mixto presente (F31.6 según CIE10); en los municipios de Belalcazar y Aguadas predominaron los casos de Trastorno bipolar, no especificado (F31.9 según CIE10); en los municipios de Pácora y Victoria predominaron los pacientes con Trastorno bipolar, episodio depresivo grave presente con síntomas psicóticos (F31.5 según CIE10) y en el resto del municipios del departamento de Caldas predominaron los casos de Trastorno bipolar, episodio maníaco presente con síntomas psicóticos (F31.2 según CIE10) (Ver Figura 34).

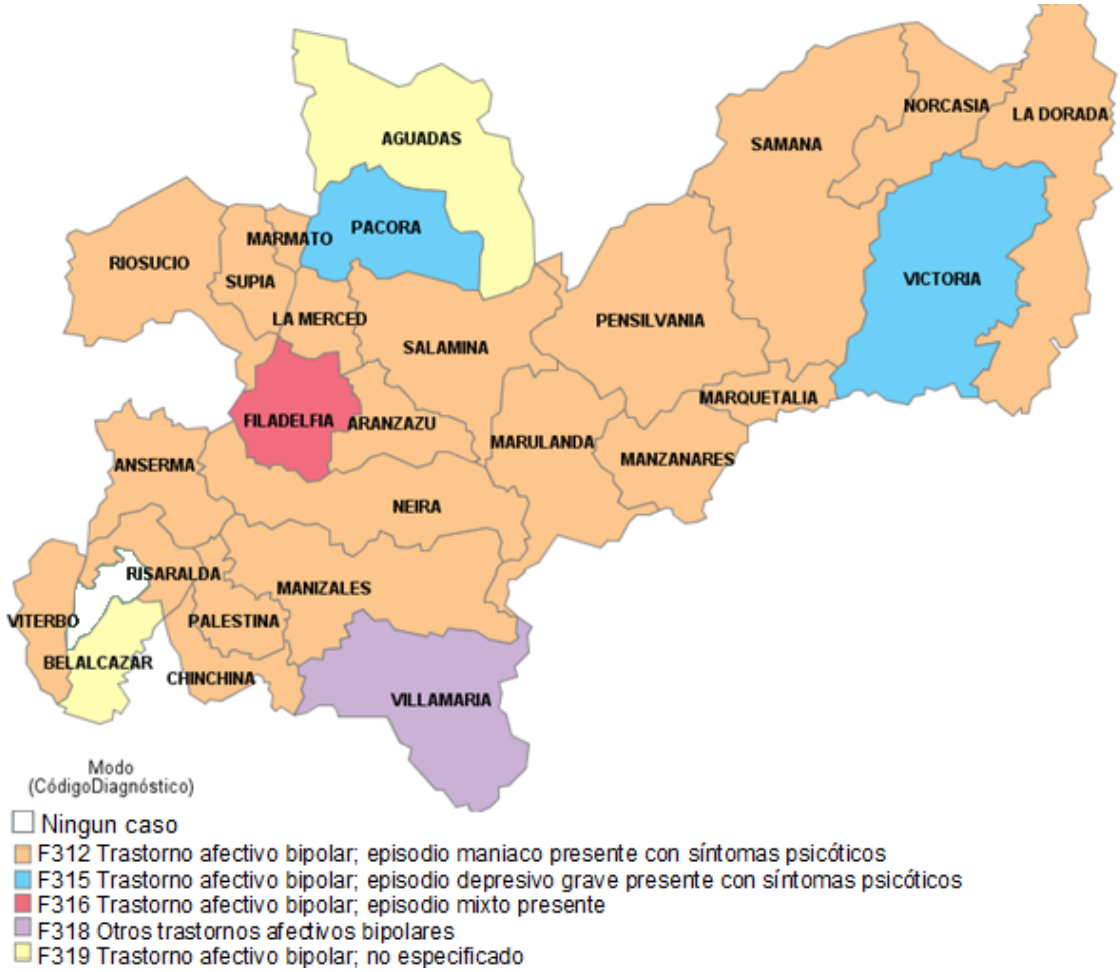


Figura 34. Distribución espacial en el departamento de Caldas del tipo del diagnóstico de los pacientes hospitalizados por TAB entre 2013 y 2014 en la CSJDM. Fuente: propia.

Estas personas hospitalizadas registraron una media de 13.52 días estancia, con una desviación estándar de 9.6 días, lo cual sugiere, que al evidenciar datos extremos de estancia, entre 1 y 122, la dispersión de los mismos es elevada al arrojar un coeficiente de variación de Pearson del 71,4%, presentado un comportamiento heterogéneo en los resultados. Al igual que el análisis de las hospitalizaciones realizado anteriormente, los días de estancia presentaron una asimetría positiva, valor estadístico de 2,550, lo que sugiere que hay mayor cantidad de pacientes hospitalizados con de menos de 13.52 días de estancia, así mismo, la media recortada al 5%, se encuentra más cerca a la mediana que a la media aritmética, demostrando la presencia de valores extremos. En este contexto, en la figura 35, podemos observar como los valores extremos se encuentran a partir de los 34 días de estancia. (Ver Tabla 27).

DÍAS DE ESTANCIA			
	ITEM	ESTADÍSTICO DESCRIPTIVO	ERROR ESTÁNDAR
Días Estancia	Media	13,52	,280
	Media recortada al 5%	12,75	
	Mediana	13,00	
	Varianza	93,420	
	Desviación estándar	9,665	
	Mínimo	1	
	Máximo	122	
	Rango	121	
	Rango intercuartil	10	
	Asimetría	2,550	,071
	Curtosis	17,867	,142

Tabla 27. Estadísticos descriptivos de los días estancia en pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

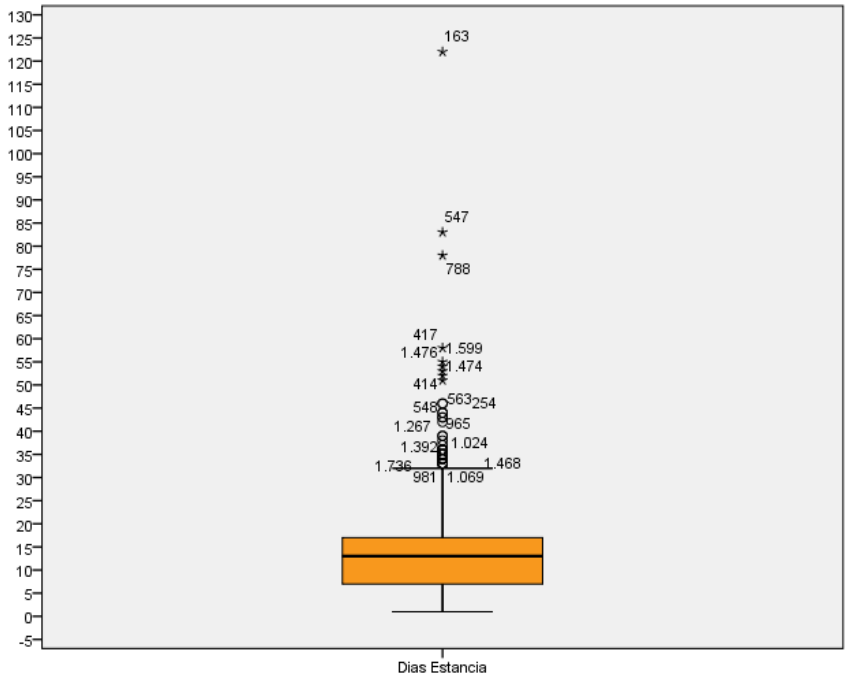


Figura 35. Diagrama de caja y valores atípicos extremos de los días de estancia de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

Al observar los días estancia de estos pacientes hospitalizados, según el municipio de procedencia, se evidencia que en el municipio de Salamina predominaron los pacientes con mayor número de días de hospitalización con una estancia que se encontraba entre 22 y 28 días. En los municipios de Belalcazar, Aguadas y Marquetalia predominaron los pacientes con una hospitalización de 15 a 21 días; por su parte los municipios de Villamaria, Pensilvania, Manzanares, Samaná y Norcasia se destacaron por presentar pacientes hospitalizados con estancias entre 1 y 7 días; los demás municipios registraron un predominio de pacientes que fueron hospitalizados entre 8 y 14 días (Ver Figura 36).

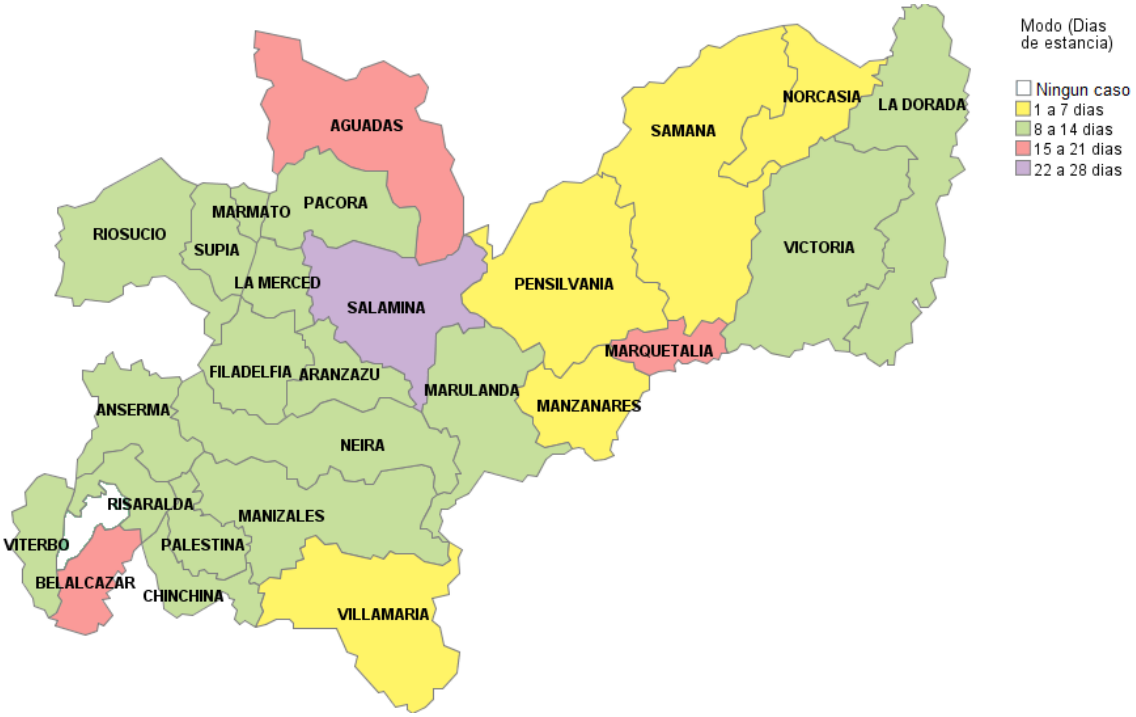


Figura 36. Distribución espacial en el departamento de Caldas de los días estancia de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

Al desagregar los días estancia por el género de los pacientes hospitalizados por TAB, se encontró que la media de los días estancia del género femenino fue de 12.93 días con una desviación estándar de 8.8 días; por su parte el género masculino registró una media de 14.59 días con una desviación estándar de 10.97 días. Estos resultados sugieren que para ambos géneros la variación de los días estancia fue elevada de acuerdo con los valores del coeficiente de Pearson, los cuales fueron de 62,5% y 75,2% para el género femenino y masculino, respectivamente; además, en ambos casos se presentó una asimetría a la derecha, lo que evidencia que los datos se concentran a la izquierda del promedio, permitiendo que se presentaran menor cantidad de días estancia que el promedio calculado. Con respecto a la curtosis, en ambos géneros también se revelaron valores elevados indicando un comportamiento leptocúrtico (Ver Tablas 28 y 29).

De acuerdo a la variabilidad de los datos, se realizó una prueba de hipótesis bajo el supuesto de que la diferencia entre medias es igual a cero (hipótesis nula) y no igual (hipótesis alternativa), encontrando que para un valor del estadístico t de -3,22 y un valor-p de 0,0012, se rechaza la hipótesis nula con un 95% de confianza; por lo tanto las medias de los géneros femenino y masculino de acuerdo a los días de estancia difieren.

Ahora bien, para determinar si los días estancia del género masculino eran mayores que los del género femenino, se realizó la prueba no paramétrica de Mann Whitney para muestras independientes y los resultados presentados (Ver Tabla 30) sugieren que los días estancia del género masculino fueron mayores que los del género femenino, siendo esta diferencia estadísticamente significativa a un alfa de 0.05 (valor p de 0.016).

DÍAS ESTANCIA EN GÉNERO FEMENINO		
ÍTEM	**ESTADÍSTICO DESCRIPTIVO**	**ERROR ESTÁNDAR**
Media	12,93	,318
Media recortada al 5%	12,27	
Mediana	12,00	
Varianza	77,573	
Desviación estándar	8,808	
Mínimo	1	
Máximo	83	
Rango	82	
Rango intercuartil	10	
Asimetría	1,628	,088
Curtosis	6,903	,177

Tabla 28. Estadísticos descriptivos de los días estancia de pacientes de género femenino hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

DÍAS ESTANCIA EN GÉNERO MASCULINO		
ÍTEM	ESTADÍSTICO DESCRIPTIVO	ERROR ESTÁNDAR
Media	14,59	,532
Media recortada al 5%	13,64	
Mediana	13,00	
Varianza	120,351	
Desviación estándar	10,970	
Mínimo	1	
Máximo	122	
Rango	121	
Rango intercuartil	11	
Asimetría	3,274	,118
Curtosis	24,034	,236

Tabla 29. Estadísticos descriptivos de los días estancia de pacientes de género Masculino hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

Estadísticos de prueba[a]	
	Días Estancia
U de Mann-Whitney	149222,500
W de Wilcoxon	442217,500
Z	-2,414
Sig. asintótica (bilateral)	,016

a. Variable de agrupación: Género

Tabla 30. Prueba de Mann Whitney para muestras independientes de los días de estancia y género de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

Al tomar los días estancia según el año de hospitalización de estos pacientes, se encontró que la media de los días estancia del año 2014 fue de 13.12 días con una desviación estándar de 10.87 días. Por su parte el año 2013 registró una media de 14.02 días con una desviación estándar de 7.84 días, (Ver Tablas 31 y 31) por lo cual se evidenció que la dispersión en los datos fue mayor para el año 2014, sin embargo al tener un coeficiente de variación de Pearson de 82,85% y 55,92% para el año 2013, ambos resultados indicaron baja concentración de los resultados con respecto a la media. Para determinar si los días estancia del año 2014 era mayores que los del año 2013, se realizó la prueba no paramétrica de Mann Whitney para muestras independiente con el fin de determinar la existencia de diferencias estadísticamente significativas de los días de estancia en función del año (hipótesis alternativa) frente al supuesto del caso contrario (hipótesis nula)

y los resultados presentados en la tabla 32 sugieren que los días estancia del año 2014 fueron menores que los del año 2013, siendo esta diferencia estadísticamente significativa a un alfa de 0.05 (Valor p de 0.000).

DÍAS DE ESTANCIA EN PACIENTES (2014)		
ITEM	**ESTADÍSTICO DESCRIPTIVO**	**ERROR ESTÁNDAR**
Media	13,12	,421
Media recortada al 5%	12,03	
Mediana	12,00	
Varianza	118,283	
Desviación estándar	10,876	
Mínimo	1	
Máximo	122	
Rango	121	
Rango intercuartil	10	
Asimetría	3,044	,095
Curtosis	20,002	,189

Tabla 31. Estadísticos descriptivos de los días estancia de pacientes hospitalizados en el año 2014 por TAB en la CSJDM. Fuente: propia.

DÍAS DE ESTANCIA EN PACIENTES (2013)		
ITEM	**ESTADÍSTICO DESCRIPTIVO**	**ERROR ESTÁNDAR**
Media	14,02	,343
Media recortada al 5%	13,63	
Mediana	13,00	
Varianza	61,481	
Desviación estándar	7,841	
Mínimo	1	
Máximo	44	
Rango	43	
Rango intercuartil	10	
Asimetría	0,716	,107
Curtosis	0,462	,213

Tabla 32. Estadísticos descriptivos de los días estancia de pacientes hospitalizados en el año 2013 por TAB en la CSJDM. Fuente: propia.

Estadísticos de prueba[a]	
	Días Estancia
U de Mann-Whitney	153710,500
W de Wilcoxon	376488,500
Z	-3,575
Sig. asintótica (bilateral)	,000

a. Variable de agrupación: Año

Tabla 33. Prueba de Mann Whitney para muestras independientes de los días de estancia y el año de hospitalización de pacientes con TAB en la CSJDM. Fuente: propia.

Análisis para los diagnósticos

Dando continuidad al tema de los días estancia de los pacientes hospitalizados, pero esta vez desde el punto de vista del diagnóstico asociado, se encontró que el trastorno bipolar, episodio maníaco presente con síntomas psicóticos (F31.2 según CIE-10), registró una media de 17.5 días de estancia y desviación estándar de 10.7 días. El trastorno bipolar, episodio depresivo grave presente sin o con síntomas psicóticos (F31.4 y F31.5 según CIE-10), presentaron 14.8 y 14.3 días de estancia con una desviación estándar de 8.1 y 8.6 días respectivamente (Ver Tabla 34). De forma general para todos los tipos de diagnóstico, la dispersión en los días de estancia fue elevada, es decir que no existe un comportamiento similar en todos los pacientes de cada diagnóstico frente al número de días en hospitalización.

Tipo diagnóstico	Días Estancia		
	Media	Desviación estándar	CV Pearson %
F31.2 trastorno bipolar; episodio maníaco presente con síntomas psicóticos	17,5	10,7	61,1
F31.1 trastorno bipolar; episodio maniaco presente sin síntomas psicóticos	14,8	8,1	54,7
F31.5 trastorno bipolar; episodio depresivo grave presente con síntomas psicóticos	14,3	8,6	60,1
F31.6 trastorno bipolar; episodio mixto presente	13,6	9,6	70,6
F31.7 trastorno bipolar; actualmente en remisión	12,0	8,8	73,3
F31.4 trastorno bipolar; episodio depresivo grave presente sin síntomas psicóticos	11,3	9,1	80,5
F31.8 otros trastornos afectivos bipolares	11,1	7,7	69,4
F31.9 trastorno bipolar; no especificado	10,3	8,6	83,5
F31.3 trastorno bipolar; episodio depresivo presente leve o moderado	9,5	9,3	97,9
F31.0 trastorno bipolar; episodio hipomaniaco presente	7,3	5,2	71,2

Tabla 34. Media de los días estancia según el diagnóstico de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Durante los días de hospitalización, los pacientes fueron tratados con medicamentos, principalmente con Ácido Valproico 250 mg, Clozapina 25 mg Tableta, Omeprazol 20 mg Cápsula y Lorazepam 2 mg Tableta, cabe aclarar que estos medicamentos se contaron por pacientes al que se suministró y no excluye el hecho de que un paciente se le haya suministrados una combinación de diferentes medicamentos (Ver Tabla 35).

Adicionalmente, después de permanecer hospitalizados en la CSJDM, el 85% de los pacientes fueron dados de alta por mejoría, un 11% por petición voluntaria y un caso por muerte (Ver Figura 37).

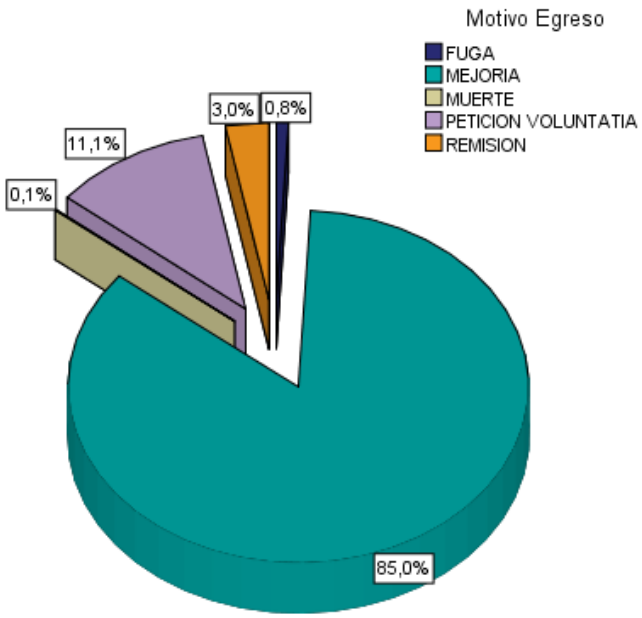


Figura 37. Distribución del egreso hospitalario de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

MEDICAMENTOS SUMINISTRADOS	CANTIDAD PACIENTES	% DE PACIENTES
Ácido Valproico 250 mg Cápsula	411	35
Clozapina 25 mg Tableta	293	25
Omeprazol 20 mg Capsula	262	22
Lorazepam 2 mg Tableta	195	16
Clozapina 100 mg Tableta	179	15
Carbonato de Litio 300 mg Tableta	170	14
Lorazepam 1 mg Tableta	169	14
Difenhidramina 50 mg Capsula	144	12
Midazolam 5 mg5 mL Solucion inyectable	139	12
Haloperidol 5 mg Tableta	128	11
Trazodona 50 mg Tableta	114	10
Clonazepam 05 mg Tableta	110	9
Haloperidol 5 mgml Solución inyectable	103	9
Sodio cloruro 09 500 ml liquido	90	8
Bisacodil 5 mg Tableta	73	6
Levotiroxina sodica 50 ug Tableta	72	6
Fluoxetina 20 mg Capsula	67	6
SERTRALINA 50 mg TABLETA	53	4
Acetil salicilico acido 100 mg Tableta	47	4
Metformina HCL 850 mg Tableta	34	3
Hidroclorotiazida 25 mg Tableta	32	3
Captopril 25 mg Tableta	30	3
Clonazepam 2 mg Tableta	30	3
Lovastatina 20 mg Tableta	28	2
Losartan 50 mg tab	27	2
Enalapril 5 mg Tableta	24	2
Cefalexina 500 mg Capsula	16	1
Risperidona 2 mg tab	15	1
Dicloxacilina 500 mg Capsula	14	1
Enalapril 20 mg Tableta	13	1
Furosemida 40 mg Tableta	13	1
Glibenclamida 5 mg Tableta	13	1
Metoprolol 50 mg Tableta	13	1
QUETIAPINA XR 300 mg TABLETA	12	1
Gemfibrozil 600 mg Tableta	10	1
Risperidona 1 mg tableta	10	1
Carbamazepina 200 mg Tableta	7	1
Quetiapina 100 mg Tableta	7	1
Tiamina 300 mg Tableta	6	1
Verapamilo Clorhidrato 80 mg Tableta	6	1

Tabla 35. Listado de medicamentos suministrados a los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Consideraciones para la construcción del modelo multivariado

Para la construcción de un modelo multivariado es necesario identificar la relación existente entre la variable dependiente y aquellas independientes, para la presente investigación la variable dependiente días estancia se tiene que someter a un análisis de correlación de Spearman, con las variables edad y estrato para identificar lo dicho acerca de la relación que tiene y además la fuerza y dirección de dicha relación tal cual lo establece la ciencia estadística.

Para determinar la relación existente entre los días de estancia con la edad de los pacientes hospitalizados y el estrato socioeconómico de los mismos, se utiliza la prueba no paramétrica de coeficientes de correlación de Spearman, los resultados presentados (Ver Tabla 36) indican que los días de estancia y la edad se encuentran relacionados a un nivel de significancia de 0.000; dicha relación fue levemente positiva, lo cual sugiere que en la medida que la edad de los pacientes sea mayor, los días estancia aumentan (Ver Figura 38) y que la correlación es mínima al presentar un valor cercano al cero (0,148).

CORRELACIONES					
			Edad	**Estrato**	**Días Estancia**
Rho de Spearman	**Edad**	**Coeficiente de correlación**	1,000	,017	,148**
		Sig. (unilateral)	.	,284	,000
		N	1191	1191	1191
	Estrato	**Coeficiente de correlación**	,017	1,000	-,080**
		Sig. (unilateral)	,284	.	,003
		N	1191	1191	1191
	Días Estancia	**Coeficiente de correlación**	,148**	-,080**	1,000
		Sig. (unilateral)	,000	,003	.
		N	1191	1191	1191

**. La correlación es significativa en el nivel 0,01 (1 cola).

Tabla 36. Correlaciones de spearman para los días estancia, edad y el estrato socioeconómico de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Al analizar la relación existente entre el estrato socioeconómico y la variable dependiente días de estancia (ver tabla 36) se evidencia una correlación mínima (-0,08) a un nivel de significancia de 0.003, siendo esta correlación inversa (negativa), lo que sugiere que en la medida que el estrato socioeconómico de los pacientes sea menor, el tiempo de hospitalización es mayor (Ver Figura 39).

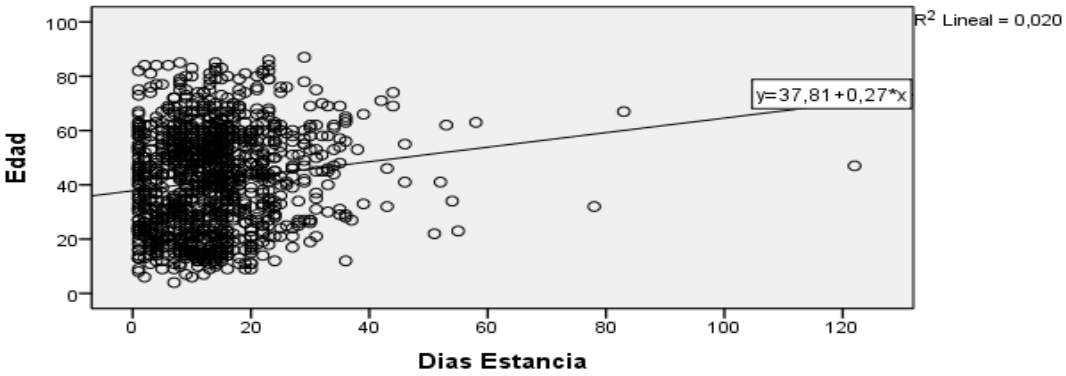


Figura 38. Nube de puntos de la correlación de los días estancia y la edad de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia

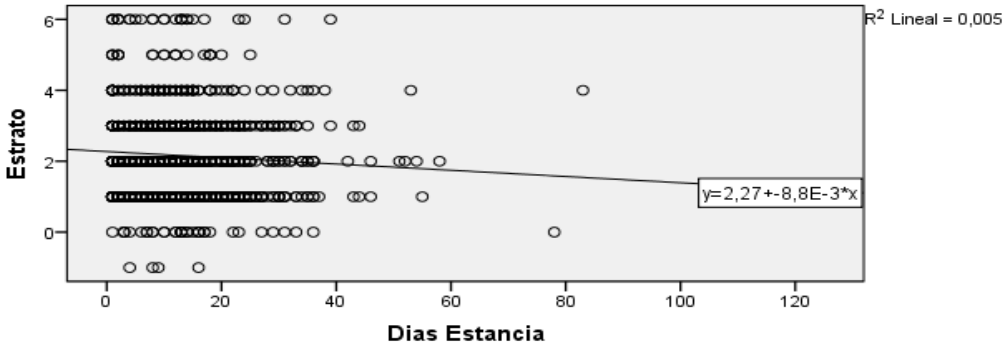


Figura 39. Nube de puntos de la correlación de los días estancia y el estrato socioeconómico de los pacientes hospitalizados entre 2013 y 2014 por TAB en la CSJDM. Fuente: propia.

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

El análisis de componentes principales descubre la estructura latente (dimensiones) de un conjunto de variables; es utilizado para dos fines, primero reafirmar la correlación existente entre los factores que componen la categoría y segundo permitir la reducción de factores que la componen, generando nuevas variables que expresen la información contenida en un conjunto de datos para la construcción de modelos multivariados.

Los modelos multivariados se construyen para identificar la relación existente entre una variable dependiente y unos factores o variables independientes y a partir de las variables o factores independientes predecir el comportamiento de la variable dependiente, tanto la precisión como otros aspectos a considerar, dependen de la capacidad predictiva de los factores o variables independientes y no de la cantidad de los mismos.

El análisis factorial en este contexto permite, por un lado, identificar aquellas variables o factores más relacionados en el conjunto de variables, y por el otro eliminar la multicolinealidad entre variables, por lo tanto, es una técnica inicial para identificar la capacidad predictiva de algunas variables o factores que se deben incluir para la construcción de un modelo multivariado.

Es de aclarar, que para este análisis no se incluyeron las variables relacionadas con los medicamentos, dado que la frecuencia que presentaron la gran mayoría no fueron suficientes para incluirlas en esta técnica. Las variables incluidas fueron: la edad, el estado civil, el género, el motivo de egreso, estrato socioeconómico, régimen de seguridad social y diagnóstico asociado.

Dentro de los supuestos para la aplicación del ACP los datos deben ser del orden cuantitativo. Sin embargo, Kim y Mueller (1978) mencionan que los datos ordinales se pueden utilizar si se piensa que la asignación de categorías ordinales a los datos no distorsiona gravemente la escala métrica subyacente. Asimismo, los autores permiten el uso de los datos dicotómicos si las correlaciones subyacentes métricas entre las variables se cree que son moderadas. Tal cual como se realizó para la presente investigación, de tal forma que las variables cualitativas, como el género, al ser de carácter binario, fue transformada en 1 y 0, de tal forma que la media se calculó a través de la proporción de unos.

El análisis de componentes principales comienza con La prueba de Bartlett, la cual está referida a la matriz de correlaciones. Se contrasta la siguiente hipótesis nula (Ho): La matriz de correlaciones es una matriz de identidad; versus la hipótesis alternante: la matriz de correlaciones no es una matriz de identidad. En caso de rechazarla (Ho), se concluye que las variables están correlacionadas entre sí, lo que da sentido al análisis componentes principales a realizar. La prueba de Kaiser-Meyer-Olkin (KMO) estima un valor que de acuerdo a su ubicación en una escala permitirá concluir si el análisis realizado es conveniente. En la medida que los primeros sean más altos, el valor estimado estará más cerca de uno, y por lo tanto el modelo factorial empleado será más efectivo.

PRUEBA DE KMO Y BARTLETT		
Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,594
Prueba de esfericidad de Bartlett	**Aprox. Chi-cuadrado**	727,067
	Gl	28
	Sig.	,000

Tabla 37. Prueba de Kaiser-Meyer-Olkin (KMO) y Bartlett para la reducción de variables en el estudio sobe los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia

Debido a que es necesario evaluar la correlación entre las variables involucradas para determinar la necesidad de aplicar el análisis de componentes principales, se realizó la prueba de esfericidad de Bartlett, encontrándose que, para un valor de p de 0,000 la hipótesis nula se rechaza, por lo anterior existe evidencia de que la matriz de coeficientes de correlación es significativamente diferente a la matriz identidad; indicando que se puede realizar el ACP dado el indicio de alta correlación entre las variables analizadas.

En el cuadro de comunalidades (Ver Tabla 38), se mide el porcentaje de la varianza de cada uno de los factores que se explica por el resto de los factores que componen la categoría, en el caso del factor estrato socioeconómico, es explicado en un 70.9% por los demás factores. En general, las comunalidades en esta tabla son todas de valores medios, lo que indica que el componente extraído representa bien los demás factores, por lo tanto los 8 factores incluidos en este análisis se encuentran ampliamente relaciones entre sí.

COMUNALIDADES

	Inicial	Extracción
Edad	1,000	,633
Estado Civil	1,000	,582
Género	1,000	,342
Nivel Educativo	1,000	,594
Motivo Egreso	1,000	,502
Estrato	1,000	,709
Régimen	1,000	,650
Diagnóstico	1,000	,281

Método de extracción: análisis de componentes principales.

Tabla 38. Tabla de comunalidades para la extracción de componentes principales en el estudio sobre los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014en la CSJDM. Fuente: propia.

El cuadro varianza total explicada muestra un modelo que es determinado por 3 tres componentes. En la columna "Sumas de extracción de cargas al cuadrado" se indica que el primer componente explica el 23.763% de la variación total de los datos, el segundo el 16.594% y el tercero un 13.295%. En general los tres componentes explican 53.652% de la variación total de los datos, puesto que, durante el análisis solo se toman los valores de la columna “Autovalores iniciales” cuyo valor total para cada componente sea mayor a 1, así cada componente muestra los valores propios, que son la proporción de la varianza total en todas las variables que se explica por ese factor en el porcentaje de varianza.

VARIANZA TOTAL EXPLICADA									
	Autovalores iniciales			Sumas de extracción de cargas al cuadrado			Sumas de rotación de cargas al cuadrado		
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	1,901	23,763	23,763	1,901	23,763	23,763	1,896	23,706	23,706
2	1,328	16,594	40,357	1,328	16,594	40,357	1,305	16,306	40,012
3	1,064	13,295	53,652	1,064	13,295	53,652	1,091	13,639	53,652
4	,988	12,351	66,003						
5	,938	11,730	77,733						
6	,775	9,691	87,424						
7	,566	7,071	94,496						
8	,440	5,504	100,000						

Método de extracción: análisis de componentes principales.

Tabla 39. Varianza total explicada para la extracción de componentes principales en el estudio sobe los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

A continuación, en la matriz de componentes (Ver Tabla 39), se proporciona el factor de cargas, y esta es la salida central para el análisis de componentes principales. El factor de cargas, también llamado saturación en componentes en análisis de componentes principales, son los coeficientes de correlación entre las variables, en otras palabras, las variables con mayor puntuación al interior de cada componente son las variables que se encuentran más relacionadas con las demás del mismo componente, por lo tanto, es la variable que representa al componente en el que se encuentra. Debido a que los tres componentes principales explican el 53,6% de los resultados, se consideraron las mismas para explicar la mayor variabilidad de los datos, a su vez, el primer componente explica el tamaño de las variables, mientras el segundo y tercero la forma de las mismas. En este sentido los factores de mayor carga con respecto a los días de estancia en la CSJDM, para el componente 1 fue el estrato (0.825) y el nivel educativo (0,74); para el componente 2 fue la edad (0.787) y en un menor valor el diagnóstico (0,449); y finalmente para el componente 3 fue el motivo de egreso (0.696) y el estado civil (0,501) (Ver ecuaciones de componentes principales). De esta manera, estos factores son de mayor relevancia, por lo tanto, serían los candidatos para la construcción de un modelo multivariado. Sin embargo, dado que un modelo predictivo lo que busca es anticiparse a un evento, en este caso en particular la variable tipo de egreso no sería incluida en la construcción del modelo dado que es una característica o factor que va posterior a los días estancia y lo que se busca son factores que predigan cuantos días estancia estará una persona dadas unas características o factores previos como la edad y el estrato socioeconómico.

MATRIZ DE COMPONENTE [a]			
	Componente		
	1	2	3
Edad	-,117	,787	-,001
Estado Civil	-,048	-,573	,501
Género	,116	-,273	-,504
Nivel Educativo	,744	-,201	-,028
Motivo Egreso	,128	,025	,696
Estrato	,825	,162	,049
Régimen	-,783	-,192	,003
Diagnóstico	-,093	,449	,266

Método de extracción: análisis de componentes principales.

a. 3 componentes extraídos.

Tabla 40. Matriz de componentes para la extracción de componentes principales en el estudio sobe los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

$$Componente\ principal\ 1 = -0{,}117 * edad - 0{,}048 * estado\ civil + 0{,}116 * gènero + 0{,}744 * nivel\ educativo + 0{,}128 * motivo\ egreso + 0{,}825 * estrato - 0{,}783 * regimen - 0{,}093 * diagnòstico$$

$$Componente\ principal\ 2 = 0{,}787 * edad - 0{,}573 * estado\ civil - 0{,}273 * gènero - 0{,}201 * nivel\ educativo + 0{,}025 * motivo\ egreso + 0{,}162 * estrato - 0{,}192 * regimen + 0{,}449 * diagnòstico$$

$$Componente\ principal\ 3 = -0{,}001 * edad + 0{,}501 * estado\ civil - 0{,}504 * gènero - 0{,}028 * nivel\ educativo + 0{,}696 * motivo\ egreso + 0{,}049 * estrato + 0{,}003 * regimen + 0{,}266 * diagnòstico$$

MODELOS MULTIVARIADOS

Dada la naturaleza cuantitativa de la variable dependiente (días estancia), se procedió a recodificarla en dos categorías: en estancias de 15 días o menos, y más de 15 días de estancia, con el objetivo de presentar un análisis comparativo entre dos modelos multivariados que presentaran similitudes técnicas, los más apropiados fueron regresión logística y análisis discriminante en ambos se incluyeron las mismas variables en las mismas condiciones.

La única recodificación que se realizó en la presente investigación fue en la variable días estancia, la cual se hizo estrictamente necesaria debido a que los modelos de regresión logística bivariada y análisis discriminante así los requieren; los criterios utilizados para recodificar dicha variable en dos fueron la medida de tendencia central media poblacional en su intervalo superior de 14.64 días; además del criterio experto del grupo de médicos especialistas en psiquiatría que participaron en la investigación.

Los resultados del modelo se presentan a continuación:

MODELO DE REGRESIÓN LOGÍSTICA

Para la construcción de los modelos de regresión logística se utilizó el 91% de los datos como conjunto de entrenamiento, lo que representó trabajar con 1084 registros, los demás registros serán utilizados como conjunto de pruebas para el modelo. Los factores seleccionados para la construcción del modelo fueron: edad y el estrato socioeconómico, al ser las variables con mayor tamaño dentro de los resultados arrojados en el análisis de componentes principales. La técnica utilizada para construir el modelo fue de paso a paso hacia delante, mediante el procedimiento de máxima verosimilitud (hacia adelante: wald).

Los resultados de esta prueba arrojaron dos modelos de inclusión de variables. En el primer modelo se incluyó únicamente la variable edad; en el segundo modelo se incluyó la variable, estrato. Es de recordar que cada vez que se incluye una variable nueva en cada modelo, las demás permanecen en él, es decir, en el modelo 2 quedaron las variables edad y estrato.

En la tabla 41 se pueden ver los dos modelos descritos anteriormente, y de ella podemos señalar que los dos modelos fueron estadísticamente significativos al obtener valores p (sig.) menores del alfa del 0.05; es decir, las variables independientes (edad y estrato) tomadas juntas tienen un efecto discriminante altamente significativo sobre la variable dependiente (días estancia).

Pruebas ómnibus de coeficientes de modelo				
		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	20,949	1	,000
	Bloque	20,949	1	,000
	Modelo	20,949	1	,000
Paso 2	Escalón	5,685	1	,017
	Bloque	26,634	2	,000
	Modelo	26,634	2	,000

Tabla 41. Pruebas omnibus sobre los coeficientes del modelo de regresión logística en predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

De igual forma la prueba de bondad de ajuste (Ver Tabla 42), sugiere que en cada modelo no existen diferencias significativas entre los valores observados y los pronosticados; lo que implica que en cada modelo estimado los datos observados y los datos esperados son similares en buena medida. Es de recordar, en este aspecto, que en la medida en que el valor de la significancia se encuentra mucho más alto del alfa de 0.05, el ajuste de los datos es mucho mejor; y en este caso el segundo modelo fue el que más se ajustó a los datos.

Prueba de Hosmer y Lemeshow			
Escalón	Chi-cuadrado	gl	Sig.
1	12,588	8	,127
2	6,223	8	,622

Tabla 42. Prueba de Hosmer y Lemeshow modelo de regresión logística en predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Ahora se presentan las funciones del análisis de regresión logística (Ver Tabla 43), se puede observar en ella que las variables incluidas en cada modelo son significativas; es decir, cada coeficiente (B) es diferente a cero, lo cual sugiere que dicha variable tiene un valor en cada función discriminante de regresión logística. Otro aspecto que corrobora lo anterior es el hecho de que el intervalo de confianza del 95% para Exp (B), no contiene el 1. Así mismo el signo de cada coeficiente (B) nos está sugiriendo la dirección en la cual dicha variable discrimina a los días estancia; por ejemplo en el modelo 2, el coeficiente (B) de 0.016 para la variable edad nos indica que un aumento en una (1) unidad en esta variable, aumenta en un 0.016 la probabilidad de que un paciente permanezca hospitalizado más de 15 días; y lo contrario sucede para el coeficiente (B) de -0.132 para la variable estrato, la cual sugiere que un aumento en una (1) unidad en esta variable, disminuye en un 0.132 la probabilidad de que un paciente permanezca

hospitalizado más de 15 días. Dicho en otras palabras, un paciente presentará mayor probabilidad de ser hospitalizado más de 15 días si presenta de manera simultánea mayor edad y menor estrato socioeconómico. Por otra parte, un paciente que presente de manera simultánea menor edad y mayor estrato socioeconómico tendrá mayor probabilidad de permanecer hospitalizado 15 o menos

Variables en la ecuación									
								95% C.I. para EXP(B)	
		B	Error estándar	Wald	gl	Sig.	Exp(B)	Inferior	Superior
Paso 1[a]	Edad	,016	,004	20,533	1	,000	1,016	1,009	1,023
	Constante	-1,434	,168	72,887	1	,000	,238		
Paso 2[b]	Edad	,016	,004	21,276	1	,000	1,017	1,010	1,024
	Estrato	-,132	,056	5,553	1	,018	,876	,785	,978
	Constante	-1,170	,201	34,018	1	,000	,310		

a. Variables especificadas en el paso 1: Edad.

b. Variables especificadas en el paso 2: Estrato.

Tabla 43. Variables en la ecuación modelo de regresión logística en predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Finalmente, se puede determinar, que los modelos de esta tabla sirven para discriminar o diferenciar la permanencia hospitalaria, sin embargo solo debe seleccionarse uno de los modelos, y para ello la tabla de clasificación (Ver Tabla 44), presenta las clasificaciones que se realizaron para cada modelo. En ella se observa que todos los modelos tienen un porcentaje global de diferenciación del 67,9%, pero varían al momento de clasificar correctamente a cada paciente según su grupo. En el primer modelo, el porcentaje de diferenciación en cada grupo varia del 100% al 0% (Es de recordar que en este modelo la única variable incluida fue la edad). En cuanto al segundo modelo presentó un porcentaje global de discriminación del 67,9%, pero en dicho modelo, se incluyeron las variables edad y estrato; este modelo puede diferenciar no solo a los pacientes que permanecerán 15 días o menos, sino también, aquellos que permanecerán más de15 días. Por lo tanto se sugiere que el modelo 2, es el que mejor diferenciación realiza entre los grupos, además de ser el modelo que mejor se ajusta a los datos según lo expuesto anteriormente con la prueba de Hosmer y Lemeshow**.**

Tabla de clasificación [a]

Observado			Pronosticado		
			Días estancia codificado		Corrección de porcentaje
			15 días o menos	Más de 15 días	
Paso 1	Días estancia codificado	15 días o menos	736	0	100,0
		Más de 15 días	348	0	,0
	Porcentaje global				67,9
Paso 2	Días estancia codificado	15 días o menos	732	4	99,5
		Más de 15 días	344	4	1,1
	Porcentaje global				67,9

a. El valor de corte es ,500

Tabla 44. Tabla de clasificación de variables modelo de regresión logística en predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

MODELO DE ANÁLISIS DISCRIMINANTE

Para la construcción de este modelo se utilizaron las mismas condiciones técnicas que se utilizaron para construir el modelo de regresión logística, la tabla 45 nos ofrece un resumen de la cantidad de registros y su peso porcentual, para cada grupo al interior de la variable dependiente (días estancia) recodificada, el 67,9% de los pacientes hospitalizados presentaron 15 o menos días estancia, mientras el 32.1% presento más de 15 días estancia.

DÍAS ESTANCIA CODIFICADO					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	15 días o menos	736	67,9	67,9	67,9
	Más de 15 días	348	32,1	32,1	100,0
	Total	1084	100,0	100,0	

Tabla 45. Estadísticos descriptivos del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Los resultados encontrados en la aplicación de esta técnica, sugieren que solo hay una función que explica el 100% de las diferencias existentes entre los pacientes que pertenecen a cada grupo; además, con un autovalor obtenido bastante próximo a 0 y una correlación canónica baja debemos suponer que las variables discriminantes utilizadas permiten distinguir moderadamente a cada grupo de pacientes según sus días estancia recodificados (Ver Tabla 46). Aunque datos suministrados en este apartado son importantes para describir el proceso de construcción del modelo, es importante también ver conjuntamente la tabla 47, la cual se encuentra estrechamente ligada a los autovalores, en ella el estadístico de Lambda Wilks expresa la proporción de variabilidad total no debida a diferencia entre los 2 grupos, en este caso, el valor de Lambda Wilks de 0,976 indica que existe bastante solapamiento entre los grupos, sin embargo, este valor transformado en Chi-cuadrada de 26,559 y una valor p de 0,000 permite suponer que los dos grupos comparados tienen en cada variable diferentes medias multivariadas (centroides) (Ver Tabla 47).

Autovalores				
Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	,025[a]	100,0	100,0	,156

a. Se utilizaron las primeras 1 funciones discriminantes canónicas en el análisis.

Tabla 46. Autovalores del modelo de análisis discriminante en la predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Lambda de Wilks				
Prueba de funciones	Lambda de Wilks	Chi-cuadrado	Gl	Sig.
1	,976	26,559	2	,000

Tabla 47. Prueba de Lambda de Wilks del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Uno de los supuesto del análisis discriminante, es que los grupos proceden de la misma población, más concretamente, que las matrices de varianzas-covarianzas poblacionales correspondientes a cada grupo son iguales entre si, en este caso el estadístico sugiere que este supuesto se cumple, valor p de 0,307 con un alfa de 0.05, por lo tanto ambos grupos proceden de una misma población (Ver Tabla 48). En la tabla 49 se pueden apreciar las diferencias mencionadas entre los grupos, se observa como el grupo de pacientes que pertenecía a aquellos con 15 días o menos de hospitalización presentaron determinantes logarítmicos mayores que los pacientes del grupo perteneciente a pacientes con más de 15 días de estancia hospitalaria.

Resultados de pruebas

M de Box		3,618
F	**Aprox.**	1,203
	df1	3
	df2	10910507,026
	Sig.	,307

Prueba la hipótesis nula de las matrices de covarianzas de población iguales.

Tabla 48. Prueba de M Box del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM en la CSJDM. Fuente: propia

Log determinante

Días estancia codificado	Rango	Determinante de logaritmo
15 días o menos	2	6,268
Más de 15 días	2	6,044
Dentro de grupos combinados	2	6,200

Los logaritmos naturales y los rangos de determinantes impresos son los de las matrices de covarianzas de grupo.

Tabla 49. Log determinante del modelo de análisis discriminante en la predicción de los días de estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Al observar la tabla 50, podemos afirmar que el grupo de pacientes hospitalizados con 15 días o menos de estancia tienden a tener puntuaciones negativas en la función discrimínate, mientras el grupo de pacientes hospitalizados con más de 15 días de estancia tiende a obtener puntuaciones positivas, ahora bien, atendiendo al valor de los coeficientes estandarizados, podemos sugerir que la variable edad tiene mayor importancia que la variable estrato a la hora de predecir el grupo de días estancia al cual pertenecerá un paciente durante el tratamiento de TAB. En combinación estos dos aspectos, nos indica que un incremento en la edad de los pacientes hará más probable que estos sean hospitalizados con más de 15 días estancia. Por el contrario, si la edad disminuye es más probable que los pacientes sean hospitalizados durante 15 días o menos.

La variable estrato muestra un escenario diferente, mientras más alto el estrato más probable que la estancia de un paciente sea de 15 días o menos, y de igual forma si el estrato disminuye la probabilidad de ser hospitalizado durante 15 días o más aumenta.

.

Días estancia codificado	Función 1
15 días o menos	-,108
Más de 15 días	,229

Las funciones discriminantes canónicas sin estandarizar se han evaluado en medias de grupos

Tabla 50. Funciones en centroides de grupo en el modelo de análisis discriminante para la predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM Fuente: propia.

	Función 1
Edad	,906
Estrato	-,460

Tabla 51. Coeficientes de la función discriminante canónica estandarizadas en el modelo de análisis discriminante para la predicción de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

COMPARACIÓN Y EVALUACIÓN DE LOS MODELOS MULTIVARIADOS.

La Curva ROC permite describir que tan separadas están las distribuciones de la sensibilidad y la especificidad de los modelos multivariados construidos. Entre más alejada esté la curva de la línea diagonal más eficiente será el modelo para la discriminación de los días estancia de pacientes hospitalizados por TAB, en este en particular se comparó el modelo regresión logística (en color azul) con el modelo de análisis discrimínate (en color naranja), ambos presentaron la misma capacidad de discriminar a pacientes con 15 días o menos de estancia de aquellos con más de 15 días estancia (Ver Figura 40).

Ahora bien, el área bajo la curva para ambos modelos fue de 59.4%, (Ver Tabla 52) en perspectiva una prueba con un 50% de área bajo la curva equivale a decir que se tiene la misma probabilidad de calificar a un paciente que permanecerá 15 o menos días de estancia, de aquel que permanecerá 15 días o más, es decir, no se cuenta con la capacidad para discriminar por estar ambos modelos muy cercanos a la línea diagonal. Se interpreta entonces que ambos modelos son pruebas no concluyentes ya que no reducen el grado de incertidumbre previo acerca de los días estancia de los pacientes que van hacer hospitalizados.

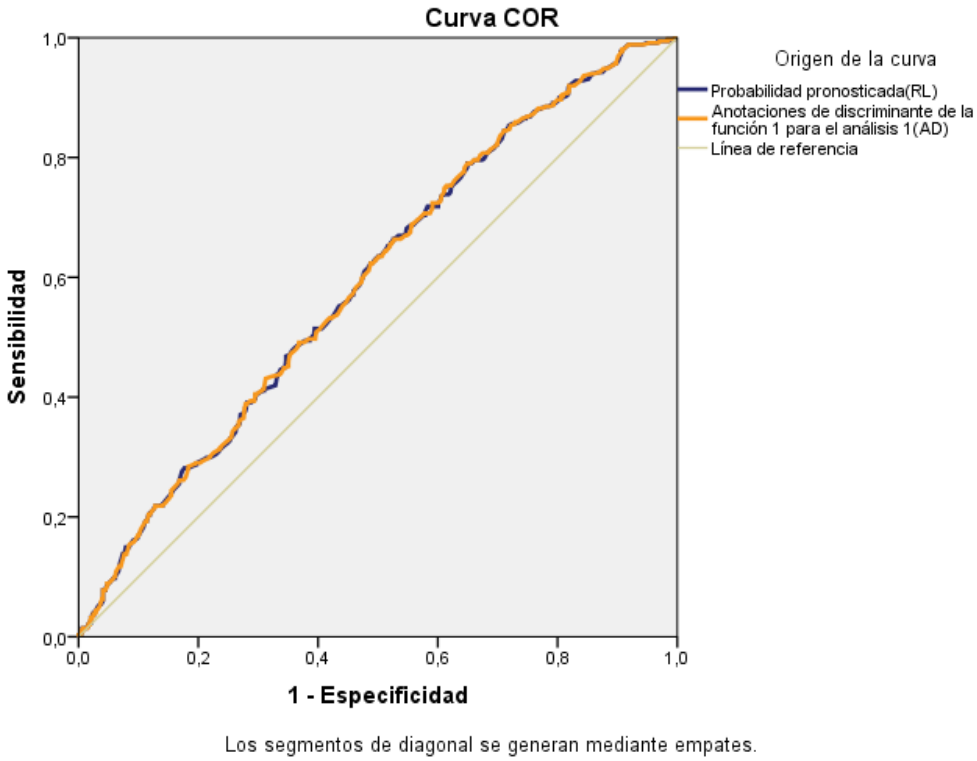


Figura 40. Curvas ROC para la evaluación de los modelos predictivos del tiempo de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

Área bajo la curva

Variable(s) de resultado de prueba	Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
				Límite inferior	Límite superior
Probabilidad pronosticada	,593	,018	,000	,558	,629
Anotaciones de discriminante de la función 1 para el análisis 1	,594	,018	,000	,558	,629

La(s) variable(s) de resultado de prueba: Probabilidad pronosticada, Anotaciones de discriminante de la función 1 para el análisis 1 tiene, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Tabla 52. Área bajo la curva curvas ROC para la evaluación de los modelos predictivos del tiempo de los días estancia de los pacientes con TAB hospitalizados entre 2013 y 2014 en la CSJDM. Fuente: propia.

DISCUSIÓN

Las metodologías de minería de datos requieren de una interpretación de los resultados de parte de un profesional en la materia de estudio, para este caso en particular, era necesaria la vinculación de profesionales en la rama de la medicina y especialmente en la salud mental. En este sentido, para explicar adecuadamente los resultados de investigación, se contó con la participación de médicos especialistas en psiquiatría, quienes tuvieron un papel fundamental a la hora de validar los resultados con su conocimiento y experiencia. Los temas tratados en esta sección fueron sometidos a revisión por parte de profesionales asistenciales, quienes identificaron elementos importantes dentro de la investigación que pueden generar valor e impacto.

En el total de la población estudiada se presentan dos tipos de análisis, el primero, hace referencia al total de eventos de hospitalización que se presentaron durante 2013 y 2014 de pacientes con TAB en la CSJDM, y el segundo, se concentra en analizar la población como individuos únicos dentro del estudio, con la intención de caracterizar los pacientes atendidos en la CSJDM en la geografía del departamento de Caldas.

Desde la perspectiva de los eventos de hospitalización, se puede evidenciar que el aumento de casos de pacientes hospitalizados por TAB entre 2013 y 2014 es del 74.3% Además, se estima con una confianza del 95% que la duración del total de las hospitalizaciones se encuentra entre 13.6 y 14.6 días. Durante este periodo, la mayor cantidad de eventos de hospitalización se debe a Trastornos Bipolares, episodios presentes maníacos con síntomas psicóticos (Clasificado en el CIE-10 como F31.2), catalogado como TAB 1 con el 30.1% de los casos.

En cuanto a la caracterización de la población atendida en la CSJDM con respecto al departamento de Caldas se encontró lo siguiente:

Según estudios relacionados a los trastornos bipolares en general, a nivel mundial la prevalencia es de 0.5% a 1.5% [Negash et al, 2005], de los cuales el Trastorno Bipolar 1 cuenta con una prevalencia del 0.8%, mientras que para el Trastorno Bipolar 2 es del 1.1%. En el análisis de los datos, se presentó la distribución espacial del TAB en general en la geografía del departamento de Caldas, en donde se evidencia una concentración elevada de los individuos que presentan esta enfermedad en el centro del departamento. El municipio de Aranzazu fue el que presentó mayor tasa de pacientes afectados por TAB con 29 pacientes por cada 10000 habitantes, seguido por Manizales y Filadelfia con 18 y 17; y municipios como La Merced, Salamina, Marulanda, Villamaría y Neira que presentaron entre 11 y 15 pacientes por cada 10000 habitantes respectivamente.

El caso de la población de Aranzazu, confirma algunos hallazgos realizados en investigaciones previas realizadas por la Universidad de Antioquia [Bedoya, et al., 2006], en donde se postula que la incidencia de estas enfermedades es de orden genético, pues son resultado de un efecto fundador de la colonización antioqueña, la cual se estableció en esta zona a finales del siglo XIX, que al tratarse de un grupo pequeño y a la práctica de matrimonios consanguíneos, se presentaron casos de endogamia, los cuales acarrearon alteraciones genéticas. Adicionalmente, se resalta el hecho de que en este municipio no había presencia de personal asistencial especializado, solo hasta el año 2011 una psicóloga comenzó a prestar sus servicios después de la manifestación de diversos eventos adversos en la población.

En el presente estudio, los pacientes del departamento de Caldas hospitalizados por TAB en la CSJDM, presentaron mayores casos debido a trastornos bipolares, episodio actual maníaco con síntomas psicóticos (F31.2 según CIE-10) con el 27.3% de los casos, seguido por un 13.9% perteneciente a trastornos bipolares, episodio actual depresivo grave sin síntomas psicóticos (F31.4 según CIE10), y un 13.2% a otros tipo de trastornos bipolares (F31.9 según CIE10). Es importante aclarar que el diagnóstico F31.2 asociado a TB 1, predominó en 20 de los 26 municipios afectados por TAB.

Según [Negash et al, 2005], el TAB se presenta igualmente en hombres y mujeres, sin embargo, los resultados no se comportaron de dicha manera, pues se evidencia un 64.23% de los casos en mujeres y 53.77% en hombres. Geográficamente, se destaca que en 25 de los 26 municipios afectados por TAB en el departamento de Caldas, predomina el género femenino.

La distribución de la población con respecto al sistema de salud indica que el 98.4% de los pacientes presentaban afiliación, en donde el 45.4% pertenecían al régimen subsidiado, y el 46.7% al contributivo. La población que más predominó en el estudio fue la de provenientes de estratos 1, 2 y 3 con el 85.8% de los casos, en donde los estratos 1 y 2 presentaron 30.5% y 30.3% respectivamente.

Los egresos por mejoría alcanzaron un 85% del total de los pacientes reportados, es oportuno mencionar, que este tipo de egreso se presenta por una recuperación sindrómica (ausencia del episodio afectivo con los criterios diagnósticos), sintomática (ausencia de síntomas desde una perspectiva dimensional) o funcional (condiciones para regreso a actividades laborales y psicosociales) [Grupo de Trabajo de la Guía de Práctica Clínica sobre Trastorno Bipolar, 2012]. Entre tanto, el 11.1% de los egresos se presentó por petición voluntaria, lo cual sugiere que los pacientes y sus familias carecen de información sobre la enfermedad y la

importancia de su tratamiento adecuado. Por tal motivo es fundamental involucrar a la familia en el tratamiento del paciente desde el principio.

Las intervenciones psicosociales tienen como objetivo proveer al paciente y su familia las técnicas y recursos necesarios para convivir con la enfermedad y prevenir las recaídas. Por este motivo, se han convertido en un elemento fundamental para el abordaje del TAB por la baja adherencia terapéutica, factores ambientales que tienen repercusiones en el curso de la enfermedad, y el hecho de que los tratamientos farmacológicos no aseguran la ausencia de recaídas.

Para los resultados de la investigación, existía un factor que afectaba significativamente el tiempo de estancia, y del cual no había un mecanismo de medición que permitiera contemplarlo para el estudio, se trata de la ubicación y asistencia de la red de apoyo, quien generalmente está compuesta por la familia, un acudiente o una institución responsable del paciente. Esta red de apoyo, según los médicos especialistas, en algunas ocasiones ya sea por dificultades locativas u otras circunstancias, no podían hacer presencia en las instalaciones de la clínica para acompañar el egreso cuando el profesional médico determinaba el alta médica del paciente, ya que el paciente había alcanzado los objetivos del tratamiento terapéutico y farmacológico, y al no poder hacer efectivamente el egreso del paciente, se prolongaba su estancia innecesariamente.

En las variables cuantitativas se incluyen la gran mayoría de estadísticos descriptivos como por ejemplo el intervalo de confianza del 95% de para la media, la media recortada al 5%, la mediana, etc. con el propósito de tener una idea general de la dichas variables, cada uno los estadísticos allí calculados tiene una función específica, en el caso del intervalo de confianza del 95% para la media su función es identificar la media de los días estancia de la población con TAB en la CSJDM no solo para los años 2013 y 2014 sino para todo el tiempo de funcionamiento. Aunque en el transcurso de la investigación no se realizó ningún tipo de inferencia estadística, la información presentada es útil no solo para la investigación actual, sino también para futuras investigaciones, por lo tanto incluir la mayor cantidad de información posible amplia no solo la presente investigación, sino también genera un punto de partida para aquellas que se generen de la misma.

La única recodificación que se realizó en la presente investigación fue en la variable días estancia, la cual se hizo estrictamente necesaria por la razón que los modelos de regresión logística bivariada y análisis discriminante así los requieren, los criterios utilizados para recodificar dicha variable en dos secciones fue la medida de tendencia central media poblacional en su intervalo superior de 14.64 días y, adicionalmente se tuvo en cuenta el criterio experto de grupo de médicos

especialistas en psiquiatría que participaron en la investigación. Con referencia al análisis de regresión logística y análisis discriminante, las variables independientes edad y estrato socioeconómico no fueron recodificadas, dado que para ambos análisis el nivel de medición que presentaron eran apropiados para los análisis, como se mencionó anteriormente, la única variable que se recodificó fue días estancia y se hizo con el criterio ya expuesto y estrictamente necesario para estos análisis multivariados donde la variable dependiente tiene que ser cualitativa bivariada.

Aunque el teorema del límite central establece que la distribución de la media de una muestra aleatoria de una población con varianza finita, tiene una distribución aproximadamente normal cuando el tamaño de la muestra es grande, independientemente de la forma de la distribución de la población. Dicho argumento genera confusión debido a la gran proximidad que tiene la distribución binomial a la distribución binomial, debido a esta confusión y gracias a los avances tecnológicos actuales podemos calcular con bastante precisión la distribución real que presenta una variable cuantitativa, por esta razón se realizó la identificación de la distribución de la variable días estancia y no en una suposición de la misma. A partir de los hallazgos encontrados, en el sentido de que la variable días estancia no procede de una distribución normal se utilizaron pruebas no paramétricas la para obtención de los estadísticos.

CONCLUSIONES

En el planteamiento inicial del problema, se formuló la siguiente pregunta de investigación, ¿Cuáles son los factores demográficos que influyen en la duración de la estancia hospitalaria en pacientes diagnosticados y hospitalizados con Trastorno Afectivo Bipolar en la Clínica San Juan de Dios de Manizales?. La investigación se centró entonces en la construcción de dos modelos predictivo de la estancia hospitalaria de los pacientes con TAB atendidos en la CSJDM, y de acuerdo a la estrategia metodológica utilizada durante la investigación se lograron los siguientes resultados:

La construcción de los modelos multivariados que permitieran predecir el tiempo de estancia, se realizó con base a las variables resultantes de la utilización de la técnica de análisis de componentes principales. En esta técnica, se resaltan 3 variables de las cuales solo 2 pueden ser utilizadas debido a su naturaleza (previas al ingreso del paciente a hospitalización). La variable tipo de egreso, resaltó desde el análisis exploratorio, debido a la cantidad de pacientes que egresan de la hospitalización por petición voluntaria, mas no por la mejoría de los síntomas que causaron la hospitalización.

Ambos modelos multivariados encuentran que los pacientes con mayor edad y menor estrato socio económico, tienden a permanecer durante el tratamiento establecido por el médico psiquiatra. Caso contrario sucede con los pacientes con menor edad y mayor estrato socio económico, que tienden a egresar antes de cumplir con las metas terapéuticas determinadas por el médico especialista.

Este fenómeno se puede presentar debido a algunas situaciones; en cuanto a la edad, entre menor sea esta, y la familia no cuente con la suficiente educación psicosocial, facilitará el egreso temprano del paciente y no permitirán una adherencia óptima al tratamiento, propiciando un egreso por petición voluntaria. Caso opuesto a una persona de mayor edad, quien posiblemente tendrá más conciencia de su estado de salud y tratará en lo posible por terminar su tratamiento adecuadamente. El estrato socioeconómico también responde a un fenómeno que se presenta en la institución; cuando los pacientes provienen de un menor estrato socio económico, cuentan con una serie de necesidades básicas suplidas satisfactoriamente como una alimentación balanceada, un patrón de sueño adecuado y actividades sociales asistidas por profesionales asistenciales; condiciones que en algunos casos son mejores que las tienen en su cotidiano vivir, motivando a los paciente y sus familias a continuar con el tratamiento hasta terminarlo adecuadamente. Por otra parte, los pacientes provenientes de estratos socio económicos altos, al contar con todas las necesidades suplidas en sus hogares y, al encontrarse compartiendo indistintamente durante la hospitalización

con los diferentes pacientes sin importar su condición social, tanto pacientes como familiares tienden a propiciar un egreso temprano sin cumplir satisfactoriamente las metas del tratamiento.

Los resultados de los modelos predictivos luego del análisis de las curvas ROC, no demuestra ser concluyente con una precisión cercana al 60%. Un enfoque para alcanzar mejores resultados en modelos de predicción de estancias hospitalarias en psiquiatría, debe contar con variables objetivo según el tipo de egreso, pues para esta clase de estudios se debe tener en cuenta que la estancia no depende únicamente de la evolución médica según criterios profesionales, sino que la voluntad del paciente y su familia juega un papel determinante para tal fin.

Los resultados de esta investigación proveerán información importante para la psiquiatría especialmente en el departamento de Caldas, pues se identificaron focos de atención que pueden representar una detección oportuna y una mejor calidad en la atención. La presente investigación representa también un beneficio intangible para la Clínica San Juan de Dios de Manizales, pues al ser el centro de referencia para la salud mental en el departamento; y al ser este departamento uno de los que cuenta con las más altas concentraciones de trastornos mentales de esa enfermedad, se genera la necesidad de impulsar nuevas investigaciones que permitan mejorar las condiciones de la población afectada por esta enfermedad.

RECOMENDACIONES

Con la intención de contar con la mayor cantidad de datos relacionados al caso de estudio para realizar un adecuado análisis de los datos, se recomienda partir de fuentes de información existentes que permitan una extracción de información lo suficientemente contextualizada como para darle explicación a determinados fenómenos y permitir un mayor grado de precisión en la construcción de los modelos predictivos.

En la exploración de los datos se resaltó un atributo que vale la pena estudiar detenidamente, los egresos por petición voluntaria determinan en gran medida el tiempo de estancia de un paciente, se recomienda analizar este comportamiento en este tipo de pacientes para identificar las causas que originan este egreso temprano que detiene impetuosamente el tratamiento del paciente.

Es necesario brindar un tratamiento adecuado a aquellas personas que no pueden acceder oportunamente a consultas médicas con personal especializado en las zonas de mayor concentración de esta enfermedad en el departamento de Caldas.

Algunas veces las herramientas utilizadas para realizar la construcción de los modelos multivariados ofrecen diferentes configuraciones que pueden cambiar en cierta medida los resultados de la investigación, en tal medida, se recomienda que para futuras investigaciones se utilice más de un programa estadístico o más de una plataforma de software para el aprendizaje automático, que permita confrontar los resultados para identificar las herramientas que mejor se ajustan a los diferentes entornos.

Se recomienda como elemento importante en el tratamiento farmacológico de los pacientes, hacer seguimiento más riguroso en los mecanismos de recolección de información electrónica, que permitan hacer el uso de técnicas de aprendizaje automático para resolver temas de aprovisionamiento farmacológico según la población que se esté atendiendo en los Institutos prestadores de salud según las proyecciones.

Fortalecer los mecanismos de recolección de información en las organizaciones a través de las tecnologías de la información, pues es esta información la que provee con mayor rigor las tendencias y comportamientos futuros de las diferentes patologías, entregando ventajas competitivas en el sector.

Se recomienda realizar informes desagregados de los resultados del análisis univariado y bivariado, con el fin de encontrar elementos relevantes dentro de las investigaciones que promuevan nuevas hipótesis para futuras investigaciones.

REFERENTE BIBLIOGRÁFICO

Alderete, A. (2006). Fundamentos del Análisis de Regresión Logística en la Investigación Psicológica. In *Evaluar, 6*, 52-67. Córdoba: Universidad Nacional de Córdoba.

Arango, C., Rojas, J., Moreno, M. (2008). Análisis de los aspectos asociados a la enfermedad mental en Colombia y la formación en psiquiatría. Revista Colombiana de Psiquiatría. vol 37, No 4

Ayuso, J. (2000). Global Burder of bipolar disorder in the year 2000. Geneva: World Health Organization: 2000.

Barnett, J., Smoller, J. (2009). The genetics of bipolar disorder. Neuroscience. 2009;164(1):331-43

Bedoya, G., Garcia, J., Montoya, P., Rojas, W., Amézquita, M., et al., (2006). Análisis de isonimia entre poblaciones del noeste de Colombia. Biomédica, Revista del Instituto Nacional de Salud. Vol24, Num 4

Britos, P. (2008). Procesos de explotación de información basados en sistemas inteligentes. Tesis doctoral en ciencias informáticas. Universidad Nacional de La Plata, Argentina

Burch, S. (2005). Palabras en Juego: Enfoques Multiculturales sobre las Sociedades de la Información. Recuperado el 25 de marzo de 2015 [http://www.analfatecnicos.net/archivos/76.SociedadDeLaInformacionYConocimiento-SallyBurch.pdf]

Cabena, P., Hadjinain, P., Stadler, R., Verhees, J., Zanasi, A., International Business Machines Corporation (San Jose, California)., International Technical Support Organization (San Jose, California). (1998). Discovering data mining: from concept to implementation. Prentice Hall.

Cerda, J., Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. Revista chilena de infectología, 29(2), 138-141.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). CRISP-DM 1.0: Step by Step Data mining Guide. The CRISP-DM consortium

CIMM (2015). Clinical Information Model Manager. Recuperado el 25 de marzo de 2015 [http://www.en13606.org/]

DANE (2015). Estimación y proyección de población nacional, departamental y municipal total por área 1985-2020. Recuperado el 1 de Agosto de 2015 [http://www.dane.gov.co/index.php/poblacion-y-demografia/proyecciones-de-poblacion]

De la Fuente, S. (2011). Análisis discriminante. Facultad de ciencias económicas y empresariales. Universidad Autónoma de Madrid. Recuperado el 08 de agosto de 2015

[http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/DISCRIMINANTE/analisis-discriminante.pdf]

Dreiseitl, S., Ohno, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics, 35(5), 352-359.

Edelstein, H. (1999). Introduction to data mining and knowledge discovery, third edition. Paperback, Two Crows Corporation (October 8, 1999).

Fayyad, U., Simoudis, E. (1997). Data Mining and Knowledge Discovery in Databases. Advances in knowledge and data mining

Fernández, E. (2015). La importancia de la analítica de datos en Colombia. Mayo de 2015. Recuperado el 26 de Agosto de 2015. [http://increnta.com/es/blog/analitica-de-datos-en-colombia/]

Freitas, A., Silva, T., Lopes, F., Garcia, I., Teixeira, A., Brazdil, P., & Costa, A. (2012). Factors influencing hospital high length of stay outliers. BMC health services research, 12(1), 265.

Gomez, V., Abasolo, J. (2009). Usind data mining to describe long hospital stays. Paradigma 2009;3(1):1 - 10

Grupo de Trabajo de la Guía de Práctica Clínica sobre Trastorno Bipolar (2012). Guía de Práctica Clínica sobre Trastorno Bipolar. Madrid: Plan de Calidad para el Sistema Nacional de Salud del Ministerio de Sanidad, Servicios Sociales e Igualdad. Universidad de Alcalá. Asociación Española de Neuropsiquiatría. 2012. UAH / AEN Núm.

Hernandez, J., Ramirez, M., Ferri, C. (2004). Introducción a la minería de datos. Pearson Education, Valencia.

Hidalgo, M. Gómez, J., Padilla, J. (2005). Regresión logística: Alternativas de análisis en la detección del funcionamiento diferencial del ítem. Psicothema. 17(3), 509-515.

Jimenez, I., Martinez, S., Rosero, C., Bonilla, M. (2015). Guia de práctica clínica trastorno afectivo bipolar ICSN 2015. Clínica Monserrat Bogotá Colombia

Judd, L., Akiskal, H., Schettler, P., Endicott, J., Maser, J., et al. (2002). The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry 2002; 59(6):530-7.

Kim, J., Mueller, C. (1978). Factor Analysis: Statistical Methods and Practical Issue. Thousand Oaks, CA, Sage Publications

Kimball, R., Caserta, J. (2004). The Data WarehouseETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis: Wiley Publishing, Inc.

Kimball, R., Reeves, L., Ross, M., Thornthwaite, W. (1998). The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses John Wiley & Sons, Ney York.

Li, J. S., Tian, Y., Liu, Y. F., Shu, T., & Liang, M. H. (2013). Applying a BP neural network model to predict the length of hospital stay. In Health Information Science (pp. 18-29). Springer Berlin Heidelberg.

Lizcano, L., Gallardo, J., Moncada, L., Nieto, K., Ortiz, Y., Carrillo, A., Durán, E. (2011). Evaluación del nivel del conocimiento que tiene el personal de enfermería sobre la guía de manejo de Trastorno Afectivo Bipolar (TAB). Ciencia y cuidado. Vol 8 No 1.

Martín, P., Díaz, A., Torres, E., Garnica, E. (1994). Una aplicación del análisis de componentes principales en el área educativa.Economía, 19(9), 55-72.

MedCalc. (2015). ROC curve analysis in MedCalc. Recuperado el 20 de agosto de 2015 [https://www.medcalc.org/manual/roc-curves.php]

Mestre, M., Samper, P., & Frías, M. (2002). Procesos cognitivos y emocionales predictores de la conducta prosocial y agresiva: La empatía como factor modulador. Psicothema, 14(2), 227-232.

Michalski, R. (1983). A Theory and Methodology of Inductive Learning. Artificial Intelligence, 20: 111-161.

Ministerio de la Protección Social. (2005). Estudio Nacional de Salud Mental Colombiana 2003. Bogotá: MPS, FES; 2005.

Ministerio de la Protección Social. (2005). Estudio Nacional de Salud Mental Colombiana 2003. Bogotá: MPS, FES; 2005.

Moine, J., Haedo, A. & Gordillo, S. (2012). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. XIII Workshop de Investigadores en Ciencias de la Computación. p. 278-281. [ISBN: 978-950-673-892-1] [Red de Universidades con Carreras en Informática (RedUNCI)]

Negash, A., Alem, A., Kebede, D., Deyessa, N., Shibre, T., & Kullgren, G. (2005). Prevalence and clinical characteristics of bipolar I disorder in Butajira, Ethiopia: a community-based study. Journal of affective disorders, 87(2), 193-201.

Nigro,H.,Xodo,D., Corti, G. (2004) KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario. Recuperado el día 15 de abril de 2015 [http://sedici.unlp.edu.ar/bitstream/handle/10915/21220/Documento_completo.pdf?sequence=1]

Noguera, T. (2010). Metodología ROC en la evaluación de medidas antropométricas como marcadores de la hipertensión arterial. Tesis de Maestría en Técnicas Estadísticas. Universidad de Santiago de Compostela.

Organización Mundial de la Salud. (1992). Clasificación Estadística Internacional de Enfermedades y Problemas Relacionado con la Salud (CIE - 10)

Organización Mundial de la Salud (OMS). (2001). The world health report 2001, mental health: New Understanding New Hope: Geneva; 2001

Ottenbacher, K., Linn, R., Smith, P., Illig, S., Mancuso, M., & Granger, C. (2004). Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Annals of epidemiology, 14(8), 551-559.

Pardo, E., Fierro, M. & Ibañez, M. (2011). Prevalencia y factores asociados a la no adherencia en el tratamiento farmacológico de mantenimiento en adultos con trastorno afectivo bipolar. Revista Colombiana de Psiquiatría. vol 40, No 1

Posada, J., Aguilar, S., Magaña, C. & Gómez, L. (2004). Prevalencia de trastornos mentales y uso de servicios:
resultados preliminares del Estudio nacional de salud mental Colombia, 2003. Revista Colombiana de Psiquiatría. vol 33, No 3

Rengifo, L., Gaviria, D., Salazar, L., Velez, J. & Lozano, S. (2012). Polimorfismos en el gen del transportador de serotonina (SLC6A4) y el trastorno afectivo bipolar en dos centros regionales de salud mental del eje cafetero. Revista Colombiana de Psiquiatría. vol 41, No 1

PSICOMED. (2015). Trastornos mentales y del comportamiento. Recuperado el 14 de abril de 2015 [http://www.psicomed.net/cie_10/cie10_F31.html]

Pyle, D. (2003). Business Modeling and Business intelligence. Morgan Kaufmann Publishers

República de Colombia, Ministerio de Salud. (1993).Resolución N 008430 de 1993. Recuperado el 8 de Mayo de 2015
[http://www.unisabana.edu.co/fileadmin/Documentos/Investigacion/comite_de_etica/Res__8430_1993_-_Salud.pdf].

República de Colombia, Ministerio de Salud. (1999).Resolución 1995 de 1999. Recuperado el 10 de Mayo de 2015
[http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=16737].

Romeu, P. & Pardo, J. (2010). Minería de datos aplicada al análisis del tratamiento informativo de la drogadicción. DSpace Universidad San Pablo CEU 2010

Rumelhart, D. & McClelland, J. (1986). Parallel Distributed Processing Vol 1: Foundations. MIT Press

Sánchez, R. & Paredes, R. (1998). Modelo de regresión logística para la predicción de tratamiento intrahospitalario prolongado en pacientes de la Unidad de Salud Mental del

Hospital San Juan de Dios de Santafé de Bogotá. Revista de la facultad de medicina de la Universidad Nacional de Colombia. Vol 46 No 1 P(8-15)

Santana, A. (2009). Efecto de la razón de tamaños sobre la detección del funcionamiento diferencial del ítem mediante regresión logística. Universidad Nacional de Colombia Facultad de Ciencias Humanas. (Tesis de maestría)

SAS Institute. Data Mining and the Case for Sampling. Recuperado el 16 de abril de 2015 [www.sasenterpriseminer.com/documents/SAS-SEMMA.pdf].

SNOMED Clinical Terms. (2015). Unified Medical Language System. Recuperado el 25 de marzo de 2015 [http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html].

Stefanovic, N., Majstorovic. V. & Stefanovic, D. (2006). Supply Chain Business Intelligence Model. Proceedings 13th International Conference on Life Cycle Engineering. P. 613-618.

Toro, E., Pérez, L. & bernal, M. (2007). Reducción de la dimensionalidad con componentes principales y técnicas de búsqueda de la proyección aplicada a la clasificación de nuevos datos. Tecnura vol11 No 21 pp, 29-40

Umapathy, K. (2007). Towards Co-Design of Business Processes and Information Systems Using Web Services. Proceedings 40th Annual Hawaii International Conference on System Sciences. Pág. 172-181.

Usuga, O. & Patiño, C. (2008). Análisis discriminante no métrico y regresión logística en el problema de clasificación.

Vila, A. (2012). Estudio de viabilidad para la mejora del sistema de información de salud de los establecimientos rurales de Perú (caso de estudio región de Loreto) utilizando la herramienta DHIS2. Máster en redes de telecomunicaciones para países en desarrollo. Universidad Rey Juan Carlos.

WebMining. (2011). KDD: Proceso de extracción de conocimiento. Recuperado el 15 de abril de 2015 [http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/]

ANEXOS

ANEXO 1: definición operacional y valores permitidos o categorías de las variables

Nombre	Variable	Tipo variable	Definición para el estudio	Valores límite
Edad	Cuantitativa	Discreta	Edad del paciente en años	[0,…, 87]
Estado_Civil	Cualitativa	Nominal	Estado civil del paciente	[Casado, Soltero, Viudo, Divorciado, Unión libre]
Género	Cualitativa	Nominal	Género del paciente	[Masculino, Femenino]
Educación	Cuantitativa	Ordinal	Nivel educativo del paciente	[Universitario incompeto, Universitario completo, tecnólogo, técnico, bachiller incompleto, bachiller completo, primaria incompleta, primaria completa, sin estudios]
Estrato	Cualitativa	Ordinal	Estrato socioeconómico del paciente	[1, 2, 3, 4, 5, 6]
IdHospitalizacion	Cuantitativa	Discreta	Número correspondiente a la hospitalización de cada paciente (autoincremento con cada nueva hospitalización)	[1,…, 35]
Motivo_Egreso	Cualitativa	Ordinal	Motivo del egreso de la estancia hospitalaria del paciente	[Fuga, mejoría, muerte, petición voluntaria, remisión]
Fecha_ingreso	Cuantitativa	Discreta	Fecha de inicio del periodo de hospitalización del paciente	[2013-01-01,…, 2014-12-31]
Fecha_Egreso	Cuantitativa	Discreta	Fecha final del periodo de hospitalización del paciente	[2013-01-01,…, 2014-12-31]
Codigo_Diagnostico	Cualittiva	Ordinal	Código correspondiente por el cual el paciente fue hospitalizado según Clasificación Internacional de enfermedades CIE-10	[F31.0 Trastorno bipolar, episodio actual hipomaníaco, F31.1 Trastorno bipolar, episodio actual maníaco sin síntomas psicóticos, F31.2 Trastorno bipolar, episodio actual maníaco con

				síntomas psicóticos, F31.3 Trastorno bipolar, episodio actual depresivo leve o moderado, F31.4Trastorno bipolar, episodio actual depresivo grave sin síntomas psicóticos, F31.5Trastorno bipolar, episodio actual depresivo grave con síntomas psicóticos, F31.6 Trastorno bipolar, episodio actual mixto, F31.7 Trastorno bipolar, actualmente en remisión, F31.8 Otros trastornos bipolares, F31.9 Trastorno bipolar sin especificación.]
Regimen	Cualitativa	Ordinal	Afiliación al Sistema de Salud	[AOferta, Contributivo, Especial, Particular, Subsidiado, Otros]
Municipio	Cualitativa	Ordinal	Municipio de procedencia del paciente	[Aguadas, Anserma, Aranzazu, Arauza, Armenia, Belalcazar, Cartago, Chinchiná, Filadelfia, La Dorada, La Merced, La Vega, Manizales, Manzanares, Marmato, Marquetalia, marulanda, Neira, Norcasia, Pácora, Palestina, Pensilvania, Pereira, Quinchia, Riosucio, Risaralda, Salamina, Sabaná,

				Sevilla, Supia, Victoria, Villamaría, Viterbo]
Dias_estancia	Cuantitativa	Discreta	Cantidad de días transcurridos desde el inicio de la hospitalización hasta el egreso	[1,…, 122]

ANEXO 2: Resumen Ejecutivo presentado al comité de Bioética de la Clínica San Juan de Dios de Manizales

INVESTIGACIÓN DE DATOS CLÍNICOS EN LA CLÍNICA SAN JUAN DE DIOS DE MANIZALES

a. La Idea

Con el ánimo de promover los procesos de investigación tecnológica enfocados a la psiquiatría, se ha encontrado la oportunidad de realizar estudios de descubrimiento de conocimiento en bases de datos (KDD), esto gracias a los grandes volúmenes de información que a la fecha no han sido explorados en la Clínica San Juan de Dios de Manizales. Inicialmente se lograron identificar inicialmente dos posibles focos de investigación para promover la generación de conocimiento a partir de los datos existentes en la bases de datos institucional. El primero de ellos se centra en la construcción de un modelo que permita identificar cual es la combinación y concentración de medicamentos que han generado mejores indicadores de evolución en los pacientes para reducir los tiempos de estancia hospitalaria y optimizar el uso de medicamentos para ciertas patologías, este estudio se puede realizar con la información de la clínica usando adecuadamente las técnicas y la tecnología disponible, adicionalmente este estudio puede promover futuros semilleros de investigación. El segundo foco de investigación se encontró en la posible identificación de modelos para el diagnóstico de patologías según los síntomas presentados por el paciente, todo esto por medio de técnicas de redes neuronales computacionales, las cuales se encargan de identificar patrones y comportamientos específicos para indicar un posible diagnóstico, este estudio se realiza por medio de técnicas de aprendizaje automático en donde el modelo tiene conjuntos de entrenamiento (aprendizaje) y conjuntos de prueba (identificación de precisión del modelo).

b. ¿Por qué?

La motivación para realizar estos estudios está sustentada en la oportunidad de entregar a la clínica resultados de estudios innovadores que permitan identificar el comportamiento a medida de la población objetivo para fortalecer los modelos de atención ya existentes, y brindar herramientas a la comunidad científica asistencial para obtener conocimiento aparentemente oculto de la información médica que se encuentra en las bases de datos, ya que es evidente que en Colombia existe un déficit en investigación que promueva la generación de nuevo conocimiento y por esta razón se ha visto un gran nicho de trabajo en el sector salud para ayudar a promover a tan importante labor.

c. Desarrollo tecnológico

El crecimiento desmedido del volumen de datos generado por los sistemas de gestión empresariales ha hecho necesario desde hace algunos años la utilización de tecnologías que permitan su organización y adecuado procesamiento. Esta necesidad ha motivado el empleo de técnicas y herramientas informáticas que permitan procesar adecuadamente la información, que posibiliten extraer conocimiento útil de la información almacenada. La Clínica San Juan de Dios de Manizales cuenta con un sistema de información que registra el contenido de las historias clínicas de los pacientes y su tratamiento durante la atención hospitalaria o ambulatoria. Sin embargo, no se aprovecha el conocimiento oculto en estos datos, este conocimiento oculto puede sustentar determinadas acciones estratégicas dentro de la institución.

El volumen y variedad de información que se encuentra informatizada en bases de datos digitales ha crecido exponencialmente en las últimas décadas, de tal forma que las empresas e instituciones en el mundo se han visto en ocasiones abarrotadas de datos históricos que no aprovechan al máximo. Esta información, bien tratada y analizada, puede reportar grandes beneficios a las organizaciones al explicar problemáticas y abrir nuevos horizontes y frentes de trabajo. Para dar respuesta a este tipo de problemas es empleado el proceso de KDD (Knowledge Discovery in Databases), se denomina descubrimiento de conocimiento en bases de datos, que posibilita la extracción de conocimiento oculto en los datos.

En el ámbito médico la aplicación de procesos de KDD tiene interés en varios campos: 1. En el ámbito clínico resulta de ayuda para la identificación y diagnóstico de patologías. Asimismo tiene importancia para el descubrimiento de posibles interrelaciones entre diversas enfermedades. 2. Al nivel de medicina preventiva, resulta de interés para la detección de pacientes con factores de riesgo para sufrir una patología. 3. Al nivel de gestión hospitalaria, se puede usar para obtener predicciones temporales que permitiesen optimizar los recursos disponibles y priorizar el uso de los diversos tratamientos para una misma patología.

d. Los Promotores

Los ingenieros en sistemas y computación Juan Sebastián Gonzales y Cristian Daniel Zuluaga de la Universidad de Caldas que actualmente adelantan los estudios de maestría en gestión y desarrollo de proyectos de software en la Universidad Autónoma de Manizales en compañía de la Dra en ciencias computacionales María Helena Mejía de la Universidad de Arizona, han identificado una viabilidad técnica de dichos estudios dada la importancia de los temas a tratar.

El ingeniero Cristian Daniel Zuluaga labora como gestor de sistemas de la Clínica San Juan de Dios.

La Dra María Helena Mejía es experta en minería de datos y se desempeña como docente de la Universidad de Caldas.

e. Objetivos principales.

Identificar un modelo para la posible reducción de la estancia hospitalaria por medio del suministro óptimo de medicamentos para determinadas patologías según datos históricos de las bases de datos de la Clínica san Juan de Dios.

Identificar un modelo para la predicción del diagnóstico psiquiátrico según las características patológicas de determinadas enfermedades.

f. Puntos fuertes y ventajas.

La clínica San Juan de Dios de Manizales cuenta la con información necesaria para realizar los estudios propuestos, esta información se encuentra en las bases de datos de la institución.

Las técnicas y herramientas tecnológicas están a libre disposición para realizar las investigaciones necesarias para alcanzar los resultados propuestos.

Dada la experiencia en el sector salud y la formación académica de los promotores, se cuenta con el conocimiento necesario para obtener los modelos propuestos.

El desarrollo de la investigación se llevará a cabo bajo la asesoría de una persona experta en la materia quien podrá orientar los procesos investigativos de la manera más idónea.

Ya que en la región no se han encontrado estudios que hagan referencia a los temas de investigación propuestos, se ha considerado que ésta es innovadora y generadora de conocimiento. Adicionalmente puede promover futuros semilleros de investigación que aporten valor a los modelos ya existentes.

g. Política de confidencialidad

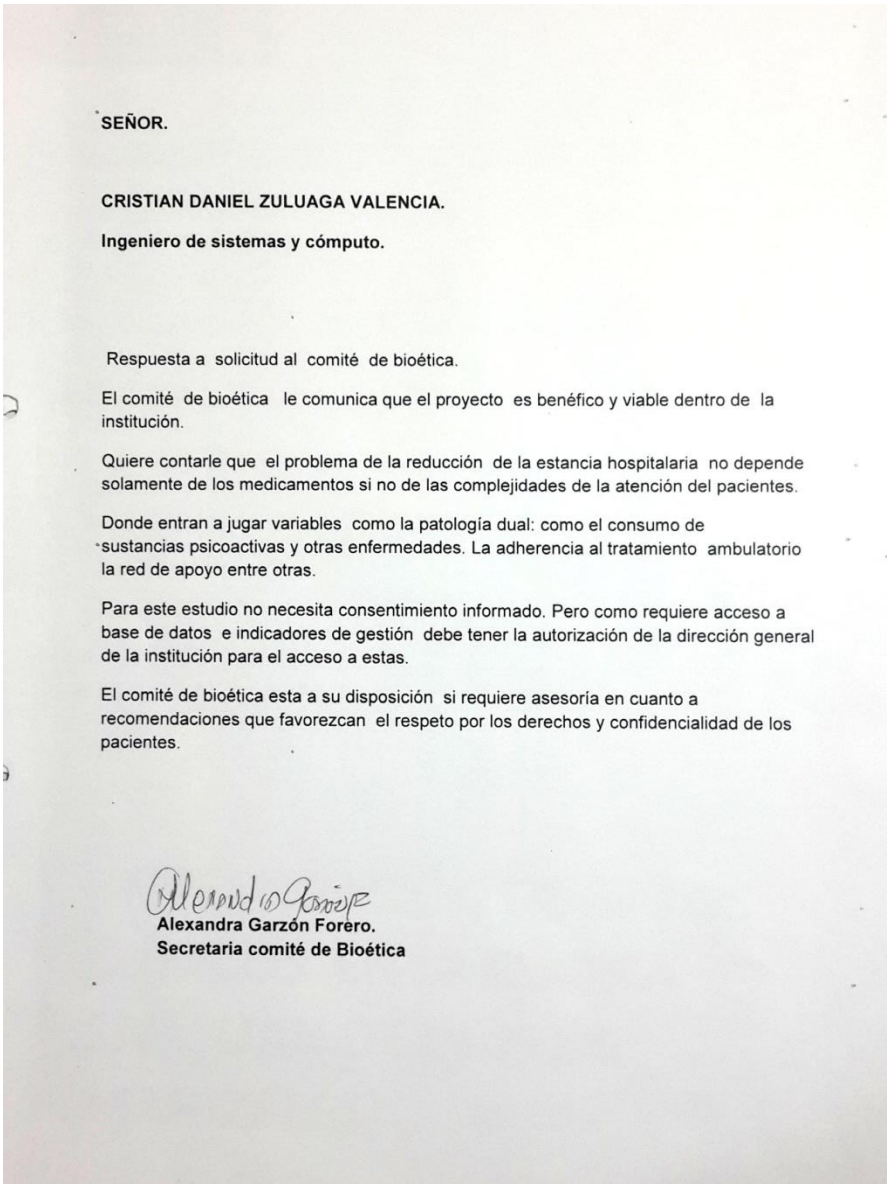
Los estudios propuestos anteriormente requieren contar con un insumo de información el cual se puede encontrar en las bases de datos de la clínica, es de conocimiento que la información de las historias clínicas contempla datos personales y confidenciales que solo le competen a la institución y al paciente. Teniendo en cuenta lo anterior, para dicho estudio se usará únicamente información epidemiológica, estadística de las estancias y su respectivo tratamiento

Cabe resaltar que no se utilizará ni divulgará ningún tipo de información que pueda ir en contra de las políticas de confidencialidad establecidas por la institución, tales como nombres de pacientes, cédulas, direcciones, teléfonos y ningún tipo de relación existente entre un paciente y su diagnóstico asociado.

h. Comunicación de resultados

Los resultados de las investigaciones propuestas se comunicarán a la Clínica San Juan de Dios y al comité de bioética una vez se finalice el estudio para su respectivo conocimiento y análisis.

ANEXO 3: Autorización de comité de bioética para el uso de los datos clínicos



SEÑOR.

CRISTIAN DANIEL ZULUAGA VALENCIA.

Ingeniero de sistemas y cómputo.

Respuesta a solicitud al comité de bioética.

El comité de bioética le comunica que el proyecto es benéfico y viable dentro de la institución.

Quiere contarle que el problema de la reducción de la estancia hospitalaria no depende solamente de los medicamentos si no de las complejidades de la atención del pacientes.

Donde entran a jugar variables como la patología dual: como el consumo de sustancias psicoactivas y otras enfermedades. La adherencia al tratamiento ambulatorio la red de apoyo entre otras.

Para este estudio no necesita consentimiento informado. Pero como requiere acceso a base de datos e indicadores de gestión debe tener la autorización de la dirección general de la institución para el acceso a estas.

El comité de bioética esta a su disposición si requiere asesoría en cuanto a recomendaciones que favorezcan el respeto por los derechos y confidencialidad de los pacientes.

Alexandra Garzón Forero.
Secretaria comité de Bioética