Tesis de Maestría en Ingeniería con Énfasis en Ingeniería de Sistemas
"Analíticas Visuales de la Información del Registro Poblacional de Cáncer de Cali"

Por:
Paola Andrea Collazos Rodríguez, Ing.
paola.collazos@gmail.com
Código: 1001425

Directora de Tesis
Beatriz Eugenia Florián Gaviria, MSc. Ph.D.
Profesora Asistente
Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle

Co-Director
Luis Eduardo Bravo Ocaña, M.D. Msc.
Profesor Emérito
Escuela de Salud, Universidad del Valle

Universidad del Valle
Facultad de Ingeniería
Escuela de Ingeniería de Sistemas y Computación
Maestría en Ingeniería con Énfasis en Ingeniería de Sistemas

Santiago de Cali, Junio de 2016.

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

Agradecimientos

Quiero comenzar agradeciendo a mi madre y padre por su apoyo incondicional siendo ejemplo de vida, y que me han brindado todo su cariño el cual es correspondido de igual forma.

También quiero agradecer de forma muy especial al Registro Poblacional de Cáncer de Cali (RPCC) por la oportunidad de desempeñarme y de desarrollarme como profesional; y principalmente al Dr. Luis Eduardo Bravo por todas las enseñanzas constante que me ha brindado durante todo este tiempo y por colaborarme en el desarrollo de este trabajo de grado. A mis compañeros de trabajo que me han ayudado a mejorar cada día como profesional y persona.

Un fuerte agradecimiento a mi tutora, Beatriz Eugenia Florián, por todos los conocimientos que compartió conmigo constantemente y por la dedicación y el gran apoyo brindado en la elaboración del presente proyecto de grado.

También quiero agradecer a mis amigos que jamás dejaron de alentar, apoyar y acompañar; y a la persona que se encuentra brindándome su amor y comprensión para seguir adelante y culminar este proyecto.

Y a la Universidad el Valle, por permitirme adquirir muchos más conocimientos, y por la oportunidad de hacer parte de la academia, que me han permitido madurar y crecer a nivel personal y profesional.

Solo me resta dar gracias a Dios por los altos y bajos, las pruebas y bendiciones.

1 Introducción

En Cali, existe un sistema información basado en un repositorio de datos de pacientes con cáncer desde 1962. Este registro (Registro Poblacional de Cáncer de Cali -RPCC) es considerado como la fuente de epidemiología descriptiva de cáncer más importante de Latinoamérica[1] y ha permitido determinar el impacto de cáncer en Cali mediante la disponibilidad de información sobre incidencia (1962-2008), mortalidad (1984-2012); siendo uno de los más importantes del mundo[2], y único de base poblacional en Colombia que ha existido por más de 50 años.

En el área de la salud, las estadísticas se centran en cálculos sobre los diferentes aspectos como son: nacimiento, muertes, enfermedades crónicas, tratamientos, estadios de la enfermedad, datos demográficos, etc. Usando esta información ha permitido identificar, idear planes de prevención y medir factores de riesgo. En general esta información es útil para la toma de decisiones y debe ser confiable, oportuna de fácil acceso para ser analizada y brindar óptimos resultados.

El RPCC presenta en su página Web[1] la información estadística de pacientes con cáncer en períodos quinquenales por sexo y grupos de edad de las diferentes localizaciones de los tumores.

Figura 1. Portal Web Registro Poblacional de Cáncer de Cali

[1] http://rpcc.univalle.edu.co/es/index.php
[2] http://gicr.iarc.fr/files/resources/20120329-GICRCancerRegistriesInfoLatAmericaS.pdf

Este servicio Web presenta una serie de limitantes como:

Los datos presentados en las diferentes tablas y gráficas no son dinámicos, porque la información carga los resultados procesados de un Software estadístico de cáncer (Seer*Stat), y posteriormente estos resultados deben ser almacenados en otra base de datos que contiene las estadísticas realizadas de los otros períodos publicados

- Las estadísticas son limitadas, porque existen procesos internos en el RPCC en donde es necesario tener frecuencias relativas de los datos que se encuentran actualmente en la base de datos, de esta forma conocer la cantidad de información existente, para evaluar el funcionamiento de la recolección de información y determinar la cantidad de datos que podrían faltar por recolectar y así definir que un periodo es completo.

- No hay personalización de la información presentada, ni navegación para diferentes tipos de usuarios.

- No están organizadas las analíticas por control de mandos, sino que se presentan tablas y gráficas en general.

Este repositorio se ha convertido en un problema de grandes volúmenes de datos que se va alimentando diariamente, y los datos son analizados al transcurrir demasiado tiempo; sin embargo, el RPCC es el principal repositorio para almacenar la información de cáncer en la ciudad de Cali, teniendo reconocimiento por la calidad de sus datos[3,4,5].

Con el presente trabajo fue posible proponer una arquitectura de software que permitió el análisis de datos del RPCC, identificando un conjunto de variables útiles en la toma de decisiones mediante los procesos de minería de datos y descubrimiento de conocimiento en bases de datos– Knowledge Discovery in Databases, KDD(Fayyad & Shapiro, 1996), de esta forma se incluyeron las analíticas visuales adaptados a los diferentes usuarios (público en general, Personal de Registros de Cáncer y Médico especialista) utilizando tableros o controles de mando (Ballvé, 2007).

El documento está estructurado en 6 secciones. La primera contextualiza proyecto presentando una introducción, glosario de términos, planteamiento del problema, justificación y objetivos. La segunda sección inicia en el capítulo 6, el cual presenta el marco teórico y contextual permitiendo identificar los temas relacionados y de interés para este proyecto como es: Introducción y conocimiento básico sobre Cáncer y Registros Poblacionales de Cáncer, minería de datos y descubrimiento de conocimiento, arquitectura de software, Big data, Inteligencia de Negocios y analíticas visuales; además el estado del arte. En la sección 3, el capítulo 8 explica un resumen del

[3] http://www.pohema.org/noticia/ganadores-del-premio-nacional/
[4] http://rpcc.univalle.edu.co/es/Publicaciones/Libros/libros.php
[5] http://comunicaciones.univalle.edu.co/InformesPrensa/2012/julio/OC-156-2012.html

desarrollo del proyecto y el capítulo 9 describe el proceso de minería de datos identificando las relaciones existentes entre los datos a presentar. En la sección 4 en el capítulo 10 y 11 se describe la arquitectura de software implementada y el resultado obtenido según el objetivo del proyecto. Finalmente en la sección 5 se realiza la evaluación y validación del aplicativo web. Finalmente, se incluyen las conclusiones, referencias bibliográficas y anexos al documento.

2 Glosario de Términos

A continuación se detalla un conjunto de términos que son utilizados en las estadísticas del área de salud que es necesario dejar en claro para garantizar el entendimiento de la presente propuesta de tesis de grado.

API. Application Programming Interface. Interfaz de programación de aplicaciones, es el conjunto de funciones y procedimientos que ofrece una biblioteca para ser utilizado por otro software como una capa de abstracción.

Arquitectura de software. Las estructuras de un sistema, compuestas de elementos con propiedades visibles de forma externa y las relaciones que existen entre ellos

Base de datos: Es un repositorio en donde guardamos información integrada que podemos almacenar y recuperar.

Datos Dinámicos: Conjunto de datos en donde su tamaño y forma es variable (o puede serlo) a lo largo de un programa, por lo que se crean y destruyen en tiempo de ejecución.

Framework. Estructura conceptual y tecnológica de soporte definido que puede servir de base para la organización y desarrollo de software. Típicamente, puede incluir soporte de programas, bibliotecas, y un lenguaje interpretado, entre otras herramientas, para así ayudar a desarrollar y unir los diferentes componentes de un proyecto.

Frecuencias: es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente. La frecuencia es el número de veces que aparece un determinado valor en un estudio estadístico[6].

Incidencia: el número de nuevos "casos" en un periodo de tiempo. Es un índice dinámico que requiere seguimiento en el tiempo de la población de interés. Los estudios de incidencia se inician con poblaciones susceptibles libres de la enfermedad, en los cuales se observa la presentación de

[6] http://www.vitutor.com/estadistica/descriptiva/a_3.html

casos nuevos a lo largo de un período de seguimiento. De esta manera, los resultados no sólo indican el volumen final de casos nuevos aparecidos durante el seguimiento, sino que permiten establecer relaciones de causa efecto entre determinadas características de la población y enfermedades específicas[7].

JSON. Javascript Object Notation. Formato ligero de intercambio de datos, completamente independiente del lenguaje de programación.

Neoplasia: Formación anormal en alguna parte del cuerpo de un tejido nuevo de carácter tumoral, benigno o maligno.

Repositorio de datos. Es un sitio centralizado donde se almacena y mantiene información digital, habitualmente bases de datos o archivos informáticos.

Sistema información. Es un conjunto de elementos orientados al tratamiento y administración de datos e información, organizados y listos para su uso posterior, generados para cubrir una necesidad o un objetivo

STATA (Stata Statistical Software): Paquete de software estadístico creado en 1985 por StataCorp. Stata: StataCorp. 2007. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Tendencias: Una de tres medidas (media, mediana y modo) en estadística, diseñada para indicar el lugar en el que se concentra el mayor número de elementos en una curva de distribución.

WEKA(Waikato Environment for Knowledge Analysis*)*: es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL.

[7] http://sameens.dia.uned.es/Trabajos7/Trabajos_Publicos/Trab_3/Caracuel_Pedraza_3/tasa/index_tasa.htm

3 Planteamiento del Problema

"El cáncer es una de las principales causas de muerte en todo el mundo; en 2008 causó 7,6 millones de defunciones (aproximadamente un 13% del total)"[8].

En 2008, la Organización Mundial de la Salud (OMS) puso en marcha el plan de acción sobre enfermedades no transmisibles, que abarca intervenciones específicas contra el cáncer.

"El cometido fundamental del Programa de la OMS de Lucha contra el Cáncer es promover políticas, planes y programas nacionales de control del cáncer que estén integrados en las iniciativas de lucha contra las enfermedades no transmisibles y los otros problemas conexos. Nuestras funciones básicas consisten en establecer normas y criterios, promover la vigilancia y fomentar la prevención basada en datos científicos, la detección precoz y el tratamiento y los cuidados paliativos adaptados a los diferentes contextos socioeconómicos"[9].

El RPCC ha permitido conocer y monitorizar la tendencia en la incidencia de cáncer en la ciudad de Cali en los últimos 50 años. Este conocimiento ha sido empleado por las autoridades locales y nacionales para la planeación de actividades de prevención y control de cáncer en la región y en el país[10] brindando conocimiento de las tendencias temporales en la incidencia que podrían usarse para proyectar las tasas de incidencia en el futuro, el número de casos y las necesidades de instalaciones sanitarias para brindar una asistencia oncológica integral (Bravo, Collazos, Collazos, García, & Correa, 2012). En el ámbito académico ha facilitado la investigación Epidemiológica con universidades nacionales y extranjeras [4,11] y ha asesorado y capacitado al talento humano de todos los Registros Poblacionales de Cáncer que existen actualmente en Colombia (Pasto, Bucaramanga, Manizales, Valledupar, Medellín, Barranquilla)[11] y algunos Latinoamericanos (Costa Rica, Perú, Venezuela y Ecuador).

El RPCC cuenta con cinco procesos (Carrascal, Guerrero, & Llanos, 2002) que han sido mejorados y optimizando, a través de las implementaciones tecnológicas. Estos procesos se muestran en la Figura 2, y se describen a continuación:

1. Búsqueda, Recolección, Procesamiento e Ingreso al sistema de información.
La información que se recolecta en el RPCC es buscado por el personal que labora en éste, el cual visita las diferentes instituciones de salud, tales como, clínicas públicas y privadas, hospitales, centros médicos y laboratorios de oncología o patología, también se obtiene información de la Secretaría de Salud Pública Municipal (SSPM). Esta información es

[4] http://rpcc.univalle.edu.co/es/Publicaciones/Libros/libros.php
[8] http://www.who.int/mediacentre/factsheets/fs297/es/
[9] http://www.etceter.com/c-conocimiento/p-dia-mundial-contra-el-cancer-2013-stopcancer/
[10] http://www.cali.gov.co/publicaciones.php?id=48468
[11] http://www.redalyc.org/articulo.oa?id=28324856002

recolectada mediante un formulario físico e ingresada al sistema de información SISCAN[12]; en algunos casos cuando existe conexión a la Web en la institución de recolección, el caso es ingresado directamente omitiendo el formulario.

2. Análisis de la información.
La información ingresada en SISCAN, debe ser verificada y analizada dependiendo de los diferentes proyectos de investigación o las publicación[4] que se estén llevando a cabo en el RPCC. Para la verificación de los datos es necesario el uso de unas herramientas[13,14,15], que dependen del conocimiento adquirido por el personal a través de los años, logrando la calidad de los datos a analizar.

3. Difusión de información
El RPCC pública la información analizada en las herramientas estadística que se muestran en la página Web[1], este proceso hace que los datos no sean dinámicos, porque se necesita exportar un archivo de texto con los resultados obtenidos de las estadísticas analizadas en la herramienta SEER*STAT[31], y posteriormente se importa estos datos a la base de datos que contiene los resultados obtenidos de los periodos anteriores.

4. Asesoría técnica
El propósito del RPCC es extender sus metodologías y experiencia en el manejo de información a otros registros de cáncer del país y América Latina[16] .

5. Investigación
EL RPCC brinda apoyo en el desarrollo de las diferentes investigaciones[17] referentes al cáncer en Cali.

Estos procesos se manejan en el RPCC y se han adaptado a las tecnologías de información brindando un óptimo y mejor funcionamiento, pero a pesar de estos cambios tecnológicos no son procesos automáticos porque dependen de procesos manuales en donde se deben exportar los datos de la base de datos Figura 2, quedando en primer lugar desactualizados porque la recolección e ingreso de la información de pacientes con cáncer a la base de datos es activo que es realizado diariamente por el personal que labora en este. Lo más importante de estos procesos es divulgar la información de pacientes con cáncer de Cali de forma oportuna y siendo útil para los especialistas en Oncología y Salud Pública, quienes pueden implementar una respuesta terapéutica y preventiva a esta temida enfermedad[18].

[12] http://rpcc.univalle.edu.co/siscan
[13] http://www.iacr.com.fr/iarccrgtools.htm
[14] http://www.seer.cancer.gov/seerprep/
[15] http://www.seer.cancer.gov/seerstat/
[16] http://gicr.iarc.fr/files/resources/20121217-CaliMeeting2012_es.pdf
[17] http://rpcc.univalle.edu.co/es/proyectos/proyectos.php
[18] http://www.mineducacion.gov.co/cvn/1665/w3-article-311021.html

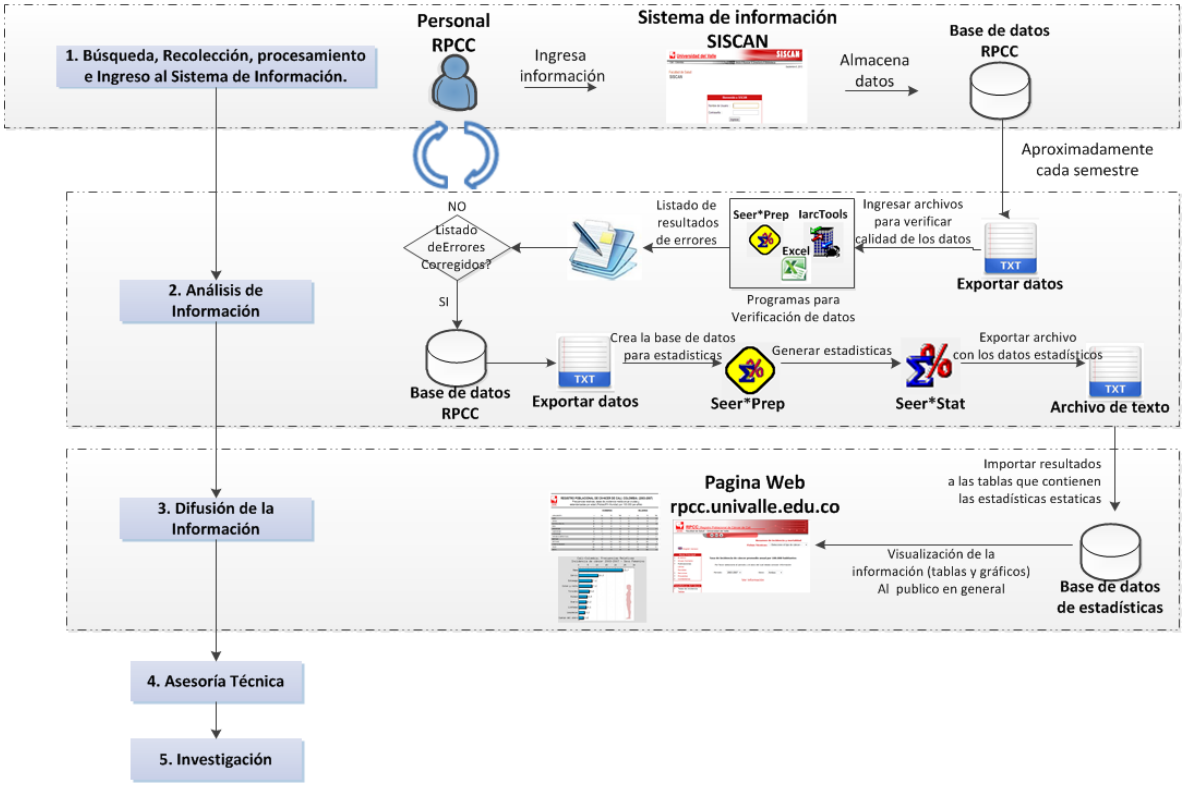


Figura 2. Mapa de procesos del Registro de Cáncer de Cali.

El RPCC presenta sus datos estadísticos en la página Web; teniendo una diferencia aproximada de 5 años a la fecha actual según la información recolectada de casos nuevos con cáncer en la base de datos, ya que no existe un mecanismo automatizado que realice el análisis de forma inmediata, según los procesos definidos y explicados anteriormente; además estos datos presentados no son dinámicos y no son adaptadas a los diferentes usuarios, como por ejemplo un usuario tipo RPCC necesitaría visualizar la información de los datos en general del ingreso de casos siendo frecuencias básicas, el cual no están reflejadas en la página actual del RPCC; y un médico especialista necesitaría ver el comportamiento del cáncer de determinada localización del tumor, y de esta forma realizar estudios específicos con estos datos.

El Programa Surveillance, Epidemiology and End Results (SEER)[19], es la única fuente de información basada en la población en Estados Unidos, que presenta la información desde la etapa del cáncer

[19] http://seer.cancer.gov/

en el momento del diagnóstico. La información presentada genera un informe personalizado según los criterios de búsqueda, y proporciona un documento en formato PDF (para impresión), tablas en HTML o archivos de texto delimitados por comas[20] para análisis propios del usuario. Conociendo esta fuente de información se quiere manejar de forma similar la información del RPCC, solo que los datos estadísticos (tablas y gráficas) serían frecuencias relativas que se analizan directamente de la base de datos y se visualicen según el tipo de usuario.

El impacto social de brindar la disponibilidad de esta información es colaborar para el cumplimiento de la Resolución 4496 de 28 diciembre 2012, por la cual se organiza el Sistema Nacional de Información en Cáncer y se crea el Observatorio Nacional de Cáncer, en donde el Artículo No. 2 incluye a los Registros de Cáncer como fuentes de información que se integran en el Sistema Nacional de Información en Cáncer[21]. La importancia de los sistemas de información en cáncer es establecer la obligatoriedad del reporte de los casos de cáncer para planificar y evaluar servicios asistenciales, monitorear el comportamiento del cáncer en el país, y retroalimentar las acciones del Plan Nacional para el Control del Cáncer[22]; de esta forma la Liga colombiana Contra el Cáncer[23] realiza campañas Nacionales con el objetivo de fomentar actividades de prevención y la detección oportuna del cáncer dirigidas a la población Colombiana.

El Plan Decenal pretende contribuir con la prevención del cáncer con la política 4x4 (alimentación saludable, actividad física, eliminación del consumo de tabaco y del alcohol) para así prevenir el 30% de las muertes. Además, se consolidan estrategias para crear el Sistema de Información en Cáncer, lo que permitirá mantener un análisis actualizado de la situación[24].

La información del RPCC, ha permitido colaborar como referencia epidemiológica de las cifras de estimaciones de Cáncer para la ciudad de Cali[25], y mediante lo publicado en la página Web ha podido mostrar un panorama de la situación del Cáncer y ha mantenido su permanencia durante todo este tiempo.

4 Justificación del Proyecto de Tesis

El cáncer es una enfermedad de interés en salud pública y prioridad nacional en salud para la República de Colombia. Para contribuir de manera significativa a un programa de control del cáncer es fundamental organizar el Sistema nacional de Información en Cáncer para Colombia y en este sistema, los Registros Poblacionales de Cáncer son una fuente primordial de información.

[20] http://seer.cancer.gov/cgi-bin/csr/1975_2010/search.pl
[21] Ministerio de Salud y Protección Social. INS. Plan nacional para el control del cáncer en Colombia 2012-20
[22] http://www.redalyc.org/pdf/120/12026437014.pdf
[23] http://www.ligacancercolombia.org/
[24] http://www.usergioarboleda.edu.co/altus/articulo-panorama-general-del-cancer-en-Colombia.htm
[25] http://www.scielo.org.co/pdf/rcc/v8n1/v8n1a02.pdf

Las Direcciones de salud de las entidades territoriales son responsables del monitoreo, seguimiento y control de la garantía de la atención de pacientes con cáncer, de conformidad con lo establecido en los Decretos 1011 y 3518, del 2006 y tienen que presentar ante los consejos departamentales para la atención del Cáncer, un plan que contenga como mínimo: acciones de vigilancia de la detección temprana de cáncer, estrategias de entrega oportuna de la información al SIVIGILA y organización de unidades de análisis para la toma de decisiones que garanticen la atención de los pacientes con cáncer o investigar las razones asociadas a los fallecimientos.

El desarrollo de la propuesta trae beneficios, en particular, para contribuir de manera significativa al programa de control del cáncer; siendo fundamental implementar el Observatorio Epidemiológico del Cáncer en Colombia del Instituto Nacional de Cancerología (INS), impulsando, fortaleciendo e integrando los Registros de Cáncer Poblacionales e Institucionales. Ante la imposibilidad de contar con un Registro Poblacional de Cáncer (RPC) en cada una de las ciudades capitales de Colombia, es prioritario consolidar los RPC existentes.

El propósito del RPCC es aportar con su experiencia a fortalecer los Registros Poblacionales de Cáncer ya existentes en algunas ciudades, brindando apoyo tanto a nivel médico como informativo y teniendo en cuenta aplicaciones informáticas que sean económicamente asequibles para ellos. Al ofrecer analíticas visuales de los datos almacenados se avanzaría a divulgar la información de forma actualizada y podrán ser de soporte para la toma de decisiones.

Este nuevo avance posiciona al RPCC a nivel mundial, demostrando una vez más las capacidades que se pueden desarrollar con una buena orientación y colaborando con la investigación.

Según la justificación mencionada se presenta esta propuesta como tesis de grado de maestría a desarrollar y de esta forma obtener el título de Maestría en Ingeniería con Énfasis en Ingeniería de Sistemas.

5 Objetivos

En esta sección se presentará el objetivo general y los objetivos específicos propuestos para el desarrollo del siguiente trabajo de grado.

5.1 Objetivo General

Proponer una arquitectura de software en el contexto médico para el almacenamiento, recuperación, análisis y visualización de la información del Registro de Cáncer de Cali, con el fin de generar datos estadísticos y analíticas visuales adaptativas y dinámicas para los diferentes tipos de usuarios (Médicos especialistas “Hemato-oncólogo pediatra”, Personal de Registros de Cáncer y público en general) mediante controles de mando especializados.

5.2 Objetivo Específicos

1. Caracterizar la información demográfica y seleccionar los principales tumores de interés según los datos publicados del RPCC.

2. Realizar un estudio exploratorio de los datos almacenados en las bases de datos del RPCC identificando las variables de interés de los diferentes usuarios, implementando la minería de datos utilizando el proceso de descubrimiento de conocimiento en bases de datos (KDD).

3. Analizar y diseñar una arquitectura de software para el almacenamiento, recuperación, procesamiento y visualización de la información almacenada en la base de datos del RPCC aplicando minería de datos a las variables de interés.

4. Generar controles de mando personalizados Web con analíticas visuales de la información del RPCC, dependiendo de las variables que influyan en su caracterización para tres diferentes usuarios (Médicos especialistas – Hemato-oncólogo pediatra, Personal de Registros de Cáncer y público en general).

5. Realizar pruebas funcionales del aplicativo con los tres diferentes usuarios (Médicos especialistas – Hemato-oncólogo pediatra, Personal de Registros de Cáncer y público en general).

6 Marco Teórico y Contextual

En esta sección se presentan los conceptos relacionados a la solución planteada al problema definido anteriormente, que colaborarán a optimizar los procesos realizados en el RPCC y dotan al lector del conocimiento necesario de dicha implementación.

Esta sección se estructura de 4 sub-secciones. En la sección 6.1, se introducen los conceptos asociados al Cáncer. En la sección 6.2, Se presenta la definición de Registro de Cáncer poblacional y en la 6.3 se presenta el Registro Poblacional de Cáncer de Cali. En la sección 6.4 se presenta la minería de datos y técnicas del descubrimiento de conocimiento, mostrando algunas de las herramientas que permiten la visualización de la información (sistemas de soporte a la decisión y controles de mando). En la sección 6.5 se presentan los componentes de una arquitectura de software y el modelo a implementar en el proyecto.

6.1 Que es el Cáncer?

El cáncer es un proceso de crecimiento y diseminación incontrolados de células. Puede aparecer prácticamente en cualquier lugar del cuerpo[26]. El tumor suele invadir el tejido circundante y puede provocar metástasis en puntos distantes del organismo.

El cáncer es un problema mundial que genera gran carga de enfermedad, en especial para los países en desarrollo (Nelson & Arias, 2013). Según globocan[27] se estimaron 12,7 millones de casos nuevos de cáncer (Figura 3) y 7,6 millones de muertes por esta causa para el año 2008; el 56% de los casos incidentes y el 63% de las muertes registradas ocurrieron en las regiones del mundo menos desarrolladas.

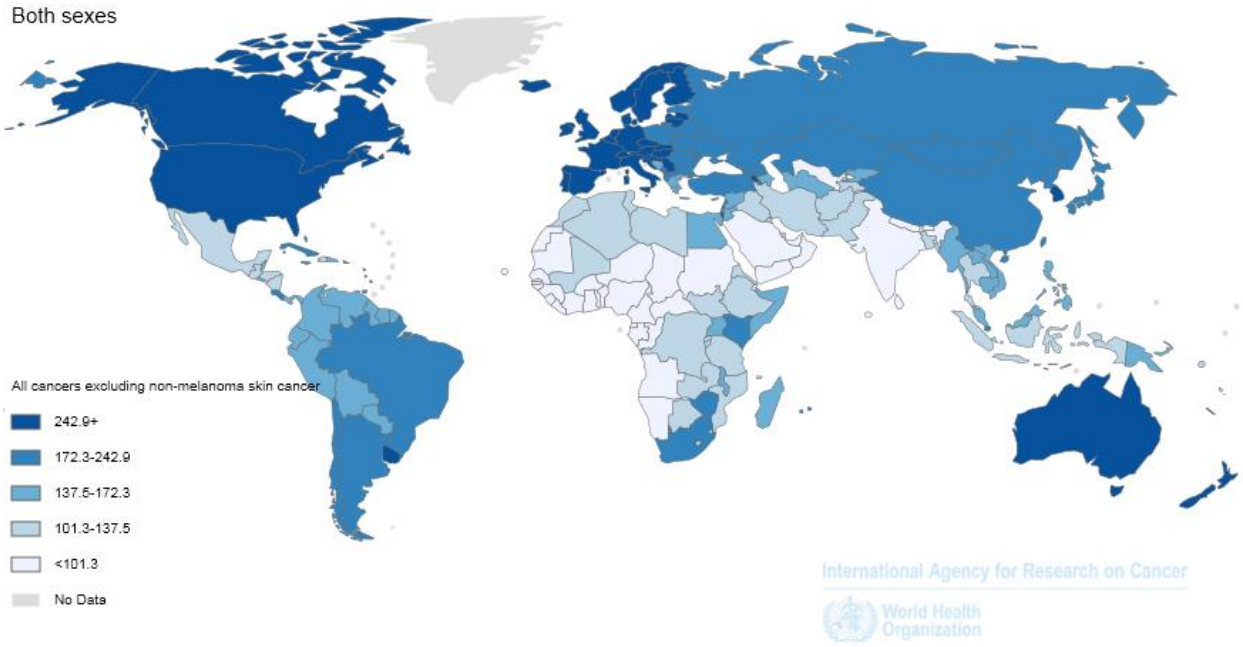


Figura 3. Incidencia de cáncer en el mundo 2008. Tasas estandarizadas por 100.000 habitantes. Globocan

6.2 Registro Poblacional de Cáncer

Los registros de cáncer de población[28] (RCP) juegan un papel esencial en el control del cáncer, ya que realizan una labor continua y sistemática de recopilación, análisis e interpretación de datos,

[26] http://www.who.int/topics/cancer/es/
[27] http://globocan.iarc.fr/Default.aspx
[28] http://cancergranada.org/es/registros_de_cancer_poblacion_historia.cfm

sobre las características personales de los pacientes con cáncer del área en la que está ubicado el Registro, incorporando también datos clínicos y anatomopatológicos del tumor y datos sobre el seguimiento de los pacientes, para conocer su supervivencia.

Alrededor del mundo existen más de 300 registros de cáncer de base poblacional, de los cuales 225 fueron incluidos en el volumen IV de Cancer Incidence in Five Continents (ci5c). De éstos, sólo 11 se encuentran en Latinoamérica y el Caribe y representan el 4,3% de la población de la región (Nelson & Arias, 2013).

En el año 2010, el Legislativo colombiano declaró el cáncer como problema de salud pública de prioridad nacional con la promulgación de las Leyes 1384 y 1388. En el articulado de estas normas se reconoce la importancia de los sistemas de información en cáncer y se establece la obligatoriedad del reporte de los casos de cáncer por parte de todos los prestadores de servicios de salud involucrados en la atención de pacientes con cáncer.

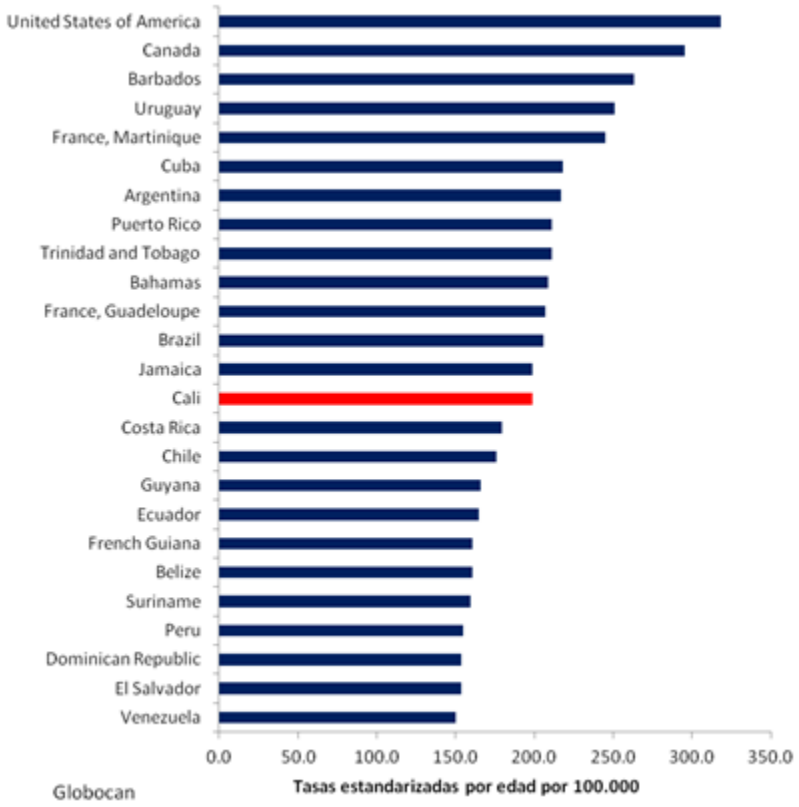
En la actualidad, se encuentran activos los registros poblacionales municipales de Barranquilla, Bucaramanga (Área Metropolitana), Cali, Manizales y Pasto, y el registro departamental de Antioquia.

6.3 Registro Poblacional de Cáncer de Cali

El Registro de Cáncer de Cali (RPCC) es el más antiguo y vigente en Latinoamérica que se encuentra en el departamento de patología de la Universidad del Valle el que ha mantenido un registro de cáncer de base poblacional desde 1962. De hecho, el Registro Poblacional de Cáncer de Cali es uno de los más importantes del mundo y el único que tiene una base poblacional de tan larga trascendencia en el contexto regional(Correa, 2012).

En el año 2012, la Asociación Internacional de Registros de Cáncer concedió un reconocimiento a 17 registros poblacionales que proporcionaron datos de incidencia a los nueve volúmenes de Cancer Incidence in Five Continents durante 50 años consecutivos. Dentro de estos registros, el de Cali fue el único incluido que provenía de países de medio y bajo ingreso; los otros 16 registros fueron de América del Norte, Europa, Japón y Australia.

A nivel mundial en América según los datos de Globocan 2008, esta es la ubicación del Cáncer del RPCC según se muestra a continuación:



EL RPCC ha permitido conocer y monitorizar la tendencia en la incidencia de cáncer en la ciudad en los últimos 50 años. Este conocimiento ha sido empleado para la planeación de actividades de prevención y control de cáncer en la región y en el país, en la asesoría para la formación de nuevos registros de cáncer en otras ciudades, en la docencia en epidemiología de cáncer a nivel de pre y postgrado y en proyectos de investigación terminados, en curso y en fase de planeación.

6.4 Minería de Datos

La cantidad de información que se encuentra almacenada en la base de datos del RPCC durante 50 años y que sigue creciendo diariamente excede nuestra habilidad para analizar los datos de una forma óptima, haciendo que se tenga en cuenta el uso de técnicas para el análisis de esta información; por este motivo se realiza la minería de datos que será explicada a continuación utilizando la metodología CRIPS-DM y la técnica de clustering es la utilizada para el desarrollo de este proyecto.

En la literatura se cuenta con varias definiciones para la minería de datos, algunas de ellas son:

- La minería de datos es el proceso de descubrir patrones de los datos. Los datos se presentan en grandes cantidades. Los patrones descubiertos deben ser significativos de manera que se permitan ventajas, por lo general, de tipo económicas. (Witten & Frank, 2005).

- Es la exploración y análisis de grandes cantidades de datos para descubrir reglas y patrones significativos. (Berry & Linoff, 2000)

- Es el conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos. (Frawley, Piatetski-Shapiro, & Matheus, 1991).

En estas tres definiciones tienen en común la extracción de información potencialmente útil (patrones, asociaciones o relaciones entre los datos) para los usuarios finales.

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- **Pronóstico:** cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.

- **Riesgo y probabilidad**: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.

- **Recomendaciones:** determinación de los productos que se pueden vender juntos y generación de recomendaciones.

- **Búsqueda de secuencias**: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.

- **Agrupación:** distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

La minería de datos es un paso esencial de un proceso más amplio cuyo objetivo es el descubrimiento de conocimiento en bases de datos (Knowlegde Discovery in Databases o KDD). KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en

última instancia comprensibles a partir de los datos (Fayyad & Shapiro, 1996), siendo un proceso iterativo e interactivo dividido en una secuencia de pasos según se muestra en la Figura 4.

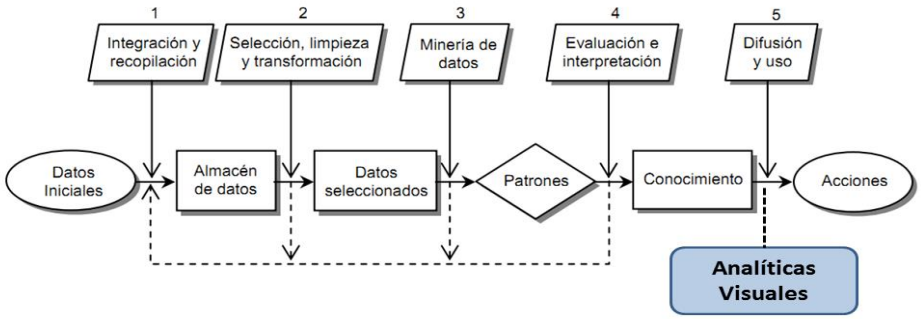


Figura 4. . Proceso de descubrimiento de conocimiento en bases de datos (KDD) aplicando analíticas visuales.

1. **Integración y recopilación**. En esta fase inicial se determina las diferentes fuentes de información, que pueden ser obtenidos de bases de datos y/o repositorios de datos; y se decidir cuáles van a ser las variables objetivas, las independientes y los datos a utilizar.

2. **Selección, limpieza y transformación**. En este paso se seleccionan el número de variables, se eliminan las redundancias en los datos y si es necesario se filtran aquellos que son relevantes para el proceso. Los métodos para la selección de características son básicamente dos:

- Aquellos basados en la elección de los mejores atributos del problema,
- Y aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

3. **Minería de datos**. En este paso se define la técnica a utilizar. Luego de aplicar esta técnica sobre los datos, debe ser validado de acuerdo al conocimiento y dominio existente.

4. **Evaluación e interpretación**. Los patrones extraídos en el paso anterior son interpretados y evaluados, garantizando la identificación de patrones verdaderamente significativos.

5. **Difusión y uso del conocimiento**. El conocimiento extraído es incorporado a algún sistema para su difusión y uso de los usuarios finales. Una de las formas para visualizar esta información es utilizar analíticas visuales adaptativas.

Por su parte, para (Han & Kamber, 2006), la minería de datos surge de la idea bajo la cual los datos, de manera organizada, ocultan información que puede ser relevante y que puede conformar nuevas formas de conocimiento. Se reconoce, de otro lado, su importancia creciente

como disciplina en ciencias de la computación impactando niveles tanto académicos como industriales (Kriegel, Borgwardt, Kröger, Pryakhin, & Zimek, 2007).

El KDD es un proceso general de descubrir conocimiento desde bases de datos, mientras que la minería de datos viene a ser la aplicación de los métodos de aprendizaje y estadísticos (Hernández, Ramírez, & Ferri, 2004), de los cuales la minería de datos es considerada uno de los pasos más importantes en todo el proceso (Wong & Leung, 2002).

6.4.1 Técnicas de Minería de Datos y Herramientas Automatizadas

Como ya se ha comentado, las técnicas de Minería de Datos (una etapa dentro del proceso completo de KDD (Fayyad & Shapiro, 1996)) intentan tener patrones o modelos a partir de los datos recopilados. Decidir si los modelos obtenidos son útiles o no suele requerir una valoración subjetiva por parte del usuario. Las técnicas de Minería de Datos se clasifican en dos grandes categorías (Figura 5): supervisadas o predictivas y no supervisadas o descriptivas (Weiss, 1998).

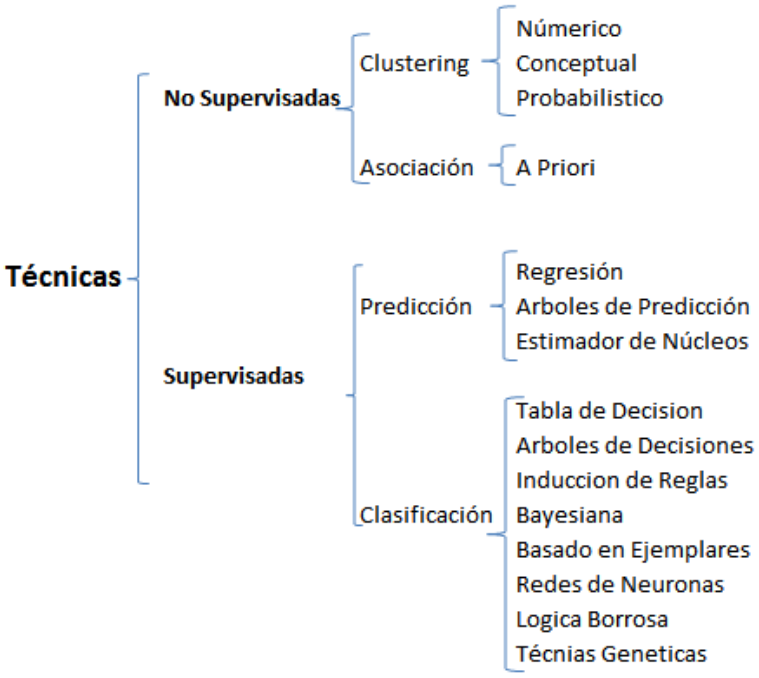


Figura 5. Técnicas de Minería de datos

Una técnica constituye el enfoque conceptual para extraer la información de los datos, y, en general es implementada por varios algoritmos. Estos algoritmos representan la manera de

desarrollar determinada técnica y es necesario tener un entendimiento de estos para saber cuál es la técnica más apropiada para cada problema.

Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión. De esta forma, hay algoritmos o técnicas que pueden servir para distintos propósitos.
El aprendizaje inductivo no supervisado estudia el aprendizaje sin la ayuda del maestro; es decir, se aborda el aprendizaje sin supervisión, que trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la guía del atributo especial clase. Éste es el proceder de los sistemas que realizan clustering conceptual y de los que se dice también que adquieren nuevos conceptos.

En el aprendizaje inductivo supervisado existe un atributo especial, normalmente denominado clase, presente en todos los ejemplos que especifica si el ejemplo pertenece o no a un cierto concepto, que será el objetivo del aprendizaje.

6.4.1.1 Clustering (Agrupaciones)

También llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Las herramientas de segmentación se basan en técnicas de carácter estadístico, de empleo de algoritmos matemáticos, de generación de reglas y de redes neuronales para el tratamiento de registros. Esta técnica suele servir de punto de partida para después hacer un análisis de clasificación sobre los clusters.

La principal característica de esta técnica es la utilización de una medida de similaridad que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar data mining sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización.

6.4.1.1.1 Clustering Numérico (k-medias)

Uno de los algoritmos más utilizados para hacer clustering es el k-medias (k-means) (MacQueen, 1967) que se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado

cuantos clusters se van a crear, éste es el parámetro k, para lo cual se seleccionan k elementos aleatoriamente, que representaran el centro o media de cada clúster. A continuación cada una de las instancias, ejemplos, es asignada al centro del clúster más cercano de acuerdo con la distancia Euclidea que le separa de él. Para cada uno de los clusters así construidos se calcula el centroide de todas sus instancias. Estos centroides son tomados como los nuevos centros de sus respectivos clusters. Finalmente se repite el proceso completo con los nuevos centros de los clusters. La iteración continúa hasta que se repite la asignación de los mismos ejemplos a los mismos clusters, ya que los puntos centrales de los clusters se han estabilizado y permanecerán invariables después de cada iteración. Para obtener los centroides, se calcula la media [mean] o la moda [mode] según se trate de atributos numéricos o simbólicos.

El algoritmo EM (Expectation Maximization) empieza adivinando los parámetros de las distribuciones (dicho de otro modo, se empieza adivinando las probabilidades de que un objeto pertenezca a una clase) y, a continuación, los utiliza para calcular las probabilidades de que cada objeto pertenezca a un clúster y usa esas probabilidades para re-estimar los parámetros de las probabilidades, hasta converger.

Este algoritmo recibe su nombre de los dos pasos en los que se basa cada iteración: el cálculo de las probabilidades de los grupos o los valores esperados de los grupos, mediante la ecuación denominado expectation; y el cálculo de los valores de los parámetros de las distribuciones, denominado maximization, en el que se maximiza la verosimilitud de las distribuciones dados los datos.

Aunque EM garantiza la convergencia, ésta puede ser a un máximo local, por lo que se recomienda repetir el proceso varias veces, con diferentes parámetros iniciales para las distribuciones. Tras estas repeticiones, se pueden comparar las medidas de verosimilitud obtenidas y escoger la mayor de todas ellas.

6.4.2 Metodologías para Minería de Datos

La metodología es un conjunto de procedimientos o actividades que sirven para alcanzar un determinado objetivo, estas principalmente ayudan a saber cómo planear lo que se debe hacer y las tareas a realizarse y cómo llevarlas a cabo.

A partir del año 2000, surgen tres metodologías que plantean un enfoque sistemático para llevar a cabo un proceso (Brito, 2008): SEMMA (SAS Entreprise Miner, 2012), Catalyst (conocida como P3TQ)(Pyle, 2003)y CRISP-DM(Pete et al., 2000). Como se puede observar en la Figura 6, CRISP-DM se ha convertido en la metodología más utilizada, según un estudio publicado en el año 2007 por la comunidad KDnuggets (Data Mining Community's Top Resource)

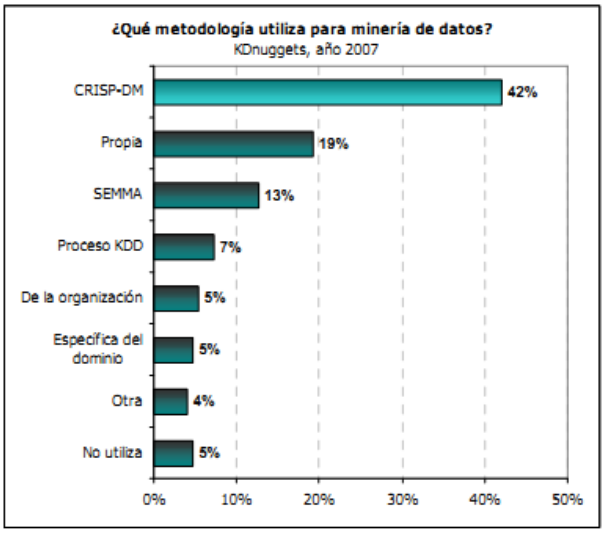


Figura 6. Encuesta realizada por la KDnuggets en el año 2007

6.4.2.1 SEMMA

La metodología SEMMA (Sample, Explore, Modify, Model and Asses) o (Muestreo, Exploración, Modificación, Modelado y Valoración en español) es una metodología desarrollada por el instituto SAS Inc., actualmente, uno de los productores más grandes de software para inteligencia de negocios, orientada al descubrimiento de conocimiento a través de patrones en los datos.

SEMMA incluye un análisis exploratorio estadístico de los datos, pasando por la utilización de técnicas de visualización, selección y transformación de variables, hasta la producción de modelos que predicen de algún modo comportamiento y cuya precisión es objeto de confirmación. Esta metodología está representado por 5 fases (Figura 7) que significa

sample=muestreo, explore=explora, modify=modifica, model=modeliza y assess=evalua, a sus siglas SEMMA (En inglés).

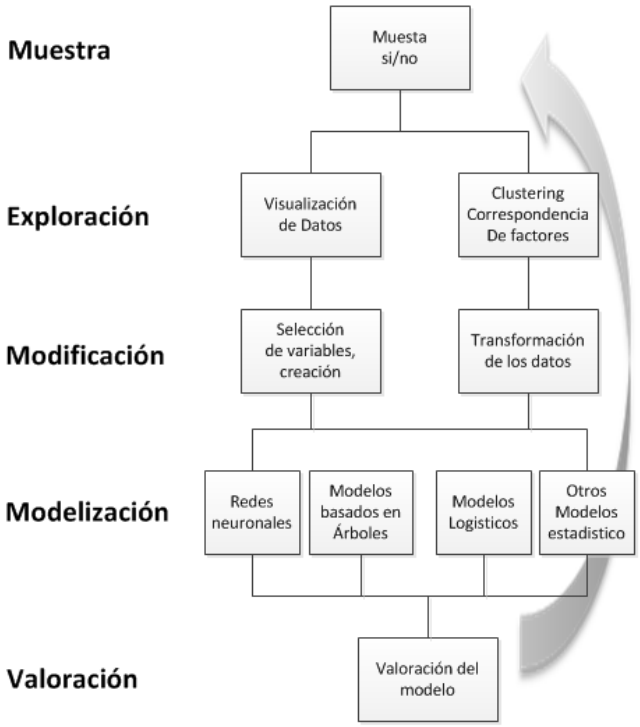


Figura 7. Metodología SEMMA

1. **Muestra:** de un gran volumen de información, extraemos una muestra lo suficientemente significativa y con el tamaño apropiado para poder manipularla con agilidad. Esto nos permite realizar los análisis de una forma rápida e inmediata

2. **Explora:** en esta fase de exploración el usuario busca tendencias imprevistas o anomalías para obtener una mejor comprensión del conjunto de datos. En esta fase se explora visualmente y numéricamente buscando tendencias o agrupaciones. Esta exploración ayuda a refinar y a redirigir el proceso.

3. **Modifica:** aquí es donde el usuario, crea, selecciona y transforma las variables con el objetivo puesto en la construcción del modelo. Basándonos en los descubrimientos de la fase de exploración, modificaremos los datos para incluir información de las agrupaciones o para introducir nuevas variables que pueden ser relevantes, o eliminar aquellas que realmente no lo son.

4. **Modeliza:** cuando encontramos una combinación de variables que predice de forma fiable un resultado deseado. En este momento estamos preparados para construir un modelo que explique los patrones en los datos. Las técnicas de modelado incluyen las redes neuronales, árboles de decisión, modelos logísticos o modelos estadísticos como series de tiempo, razonamientos basados en memoria, etc.

5. **Evalúa:** en esta fase el usuario evalúa la utilidad y fiabilidad de los descubrimientos realizados en el proceso de minería de datos. Verificaremos aquí lo bien que funciona un modelo. Para ello, podremos aplicarlo sobre muestreos de datos diferentes (de test) o sobre otros datos conocidos, y así confirmar su validez.

6.4.2.2 Catalyst (P3TQ)

La metodología CatalystoP3TQ: Producto (Product), Lugar (Place), Precio (Price), Tiempo (Time) y Cantidad (Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología básicamente propone dos modelos (Figura 8): el "Modelo de Negocio (MII)" y el "Modelo de Explotación de Información (MIII)".

El modelo de Negocio (MII), cuenta diferentes circunstancias para el proyecto de explotación de datos, proponiendo acciones concretas según el contexto desde el cual se parte. En el caso de aquellos proyectos donde no existe una definición real del problema u oportunidad de negocio, se recomienda iniciar analizando las relaciones P3TQ que existen en la cadena de valor organizacional y que son significativas para la empresa.

Según lo expresado por (Brito, 2008) el modelado en MII depende del contexto en el cual está inmerso el negocio, lo que promueve el planteamiento de distintos escenarios. Ellos son: dato, oportunidad, prospectiva, definido y estratégico:

1. **Dato.** El proyecto se inicia con un conjunto de datos y la premisa es explorar este conjunto para encontrar relaciones interesantes.
2. **Oportunidad.** El proyecto se inicia con una situación de negocio (problema u oportunidad) que debe ser explorada. En este caso, se debe:
3. **Prospectiva.** El proyecto es diseñado con el fin de descubrir dónde la explotación de Información puede brindar un valor en el entorno de la organización.
4. **Definido.** El proyecto comienza con la premisa de crear la especificación del modelo de explotación de datos con un propósito específico.

5. **Estratégico.** El proyecto comienza con una estrategia de análisis para brindar soporte a un escenario planeado por la organización.

El Modelo de Explotación de Información (MIII), Según (Brito, 2008) brinda una guía de pasos para la realización y ejecución de modelos de explotación de Información a partir del modelo de negocio desarrollado (MII). Los pasos a seguir en MIII son:

1. Preparación de los datos.
2. Selección de herramientas y modelado inicial.
3. Ejecución
4. Evaluación de resultados
5. Comunicación de resultados

La metodología P3TQ, en sus dos modelos, está constituida por un conjunto de pasos denominados "cajas" (boxes). Conceptualmente, dicha metodología determina que luego de ejecutar una acción se deben evaluar los resultados obtenidos y determinar cuál es el paso que se debe ejecutar posteriormente.

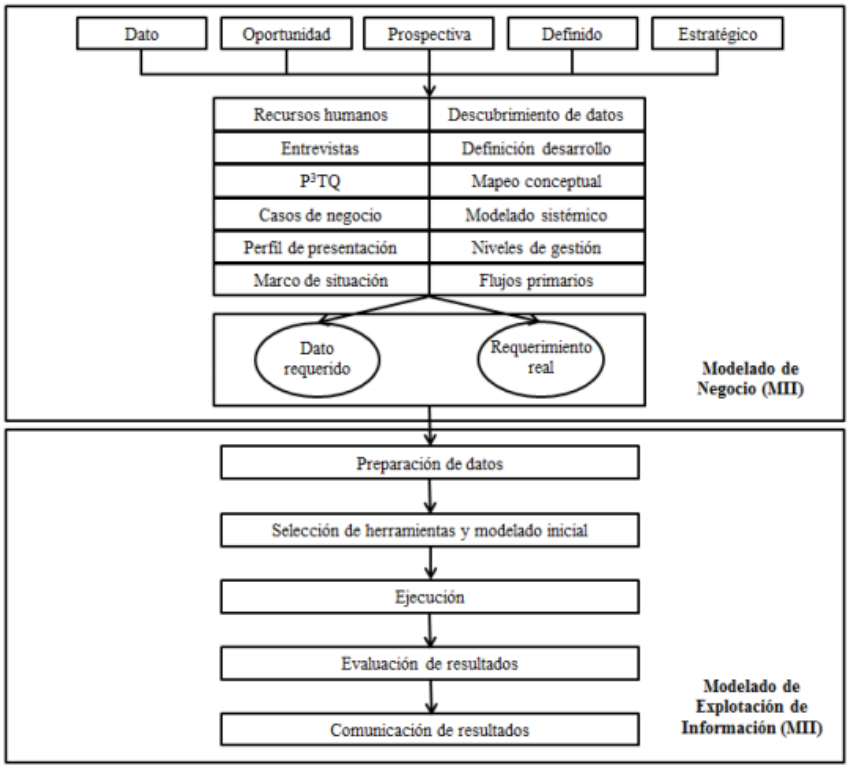


Figura 8. Fases de la metodología P3TQ y sus componentes

6.4.2.3 CRISP-DM (Cross- Industry Standard Process for Data Mining)

La metodología CRISP-DM estructura el ciclo de vida de un proyecto en seis fases (Figura 9); las flechas indican las relaciones más usuales e importantes entre ellas, aunque se pueden establecer distintas relaciones entre las distintas fases componentes. El círculo exterior simboliza la naturaleza cíclica del modelo de proceso de Explotación de Datos propiamente dicho. La secuenciación de fases no es rígida. Las fases definidas para un proyecto de desarrollo de software clásico (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre) claramente difieren de las fases propias de esta metodología [Chapman et al., 2000].

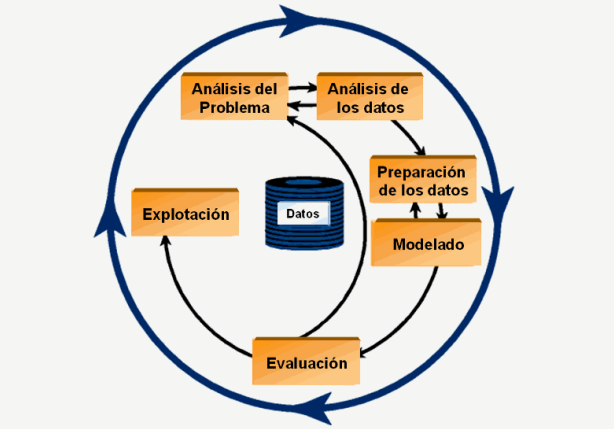


Figura 9. Metodología CRISP-DM

Las fases se definen como sigue a continuación:

- **Comprensión del negocio**: Esta primera fase inicial, se basa en el entendimiento de los objetivos del proyecto y la comprensión de los requerimientos del mismo desde el punto de vista del negocio, a fin de definir el problema a resolver y diseñar una planificación preliminar para el cumplimiento efectivo de los objetivos en cuestión.

- **Comprensión de los datos**: La segunda fase de análisis, comienza con la recolección inicial de los datos con el propósito de familiarizarse con los mismos, identificando problemas de calidad asociados a ellos e información adicional relevante para la formulación de las primeras hipótesis.
- **Preparación de los datos:** La fase de preparación de los datos, abarca todas aquellas actividades destinadas a la construcción del conjunto de datos finales. Las tareas de esta fase pueden ser ejecutadas varias veces, sin un orden predefinido. Las mismas incluyen la selección de tablas, registros y atributos, así como también la transformación y limpieza de datos para que puedan ser tratados por las herramientas de modelado.

- **Modelado:** En esta fase, se seleccionan y aplican las técnicas de modelado más apropiadas para el proyecto en cuestión, calibrando sus parámetros a valores óptimos. Básicamente, existen varias técnicas para un mismo tipo de problemas en proyectos de Explotación de Datos. Algunas de ellas, demandan requerimientos específicos sobre los datos que se van a procesar, por tal motivo muchas veces es necesario volver a la fase de preparación de los datos antes de avanzar con el modelado de los mismos.

- **Evaluación:** Esta fase, involucra la evaluación del modelo y revisión de los pasos ejecutados en relación a los objetivos del negocio, y busca determinar si hay alguna razón de negocio para el cual el modelo es deficiente, asegurándonos de esta forma, alcanzar los objetivos inicialmente propuestos. Al final de esta fase, se debe tener una decisión sobre el uso de los resultados alcanzados.

- **Implementación:** La fase de despliegue o implementación, dependiendo de los requisitos del proyecto, puede ser tan sencilla como la generación de un simple reporte o tan compleja como la implementación de un proceso de Explotación de Datos repetible en toda la empresa.

Las tres metodologías son similares, y sus fases o etapas guardan la relación entre cada una que ellas sugiere. Sin embargo, SEMMA está enfocada netamente en la implementación, CRISP-DM y P3TQ parten de la identificación del problema que se quiere tratar de resolver.

A diferencia de CRISP-DM, las metodologías P3TQ y SEMMA no identifican problemas de inteligencia de negocio (PIN), ni realizan una caracterización abstracta de los mismos (CRISP-DM hace una caracterización parcialmente abstracta de dichos problemas).

P3TQ y SEMMA, tampoco identifican relaciones entre problemas de inteligencia de negocio y técnicas de Explotación de Información, ni procesos de Explotación de Información (CRISP-DM esboza parcialmente los procesos a desarrollar).

En la Tabla 1 se muestran las ligeras diferencias que existen entre las metodologías mencionadas.

Caracteristica / Metodología	CRISP-DM	SEMMA	P^{3}TQ
Identifica problemas de inteligencia de negocio	SI	NO	NO
Identifica una caracterización abstracta de PIN (PIN)	PARCIAL	NO	NO
Identifica técnicas de Explotación de información (TEI) utilizables	SI	SI	SI
Identifica relaciones entre las TEI y los PIN	PARCIAL	NO	NO
Identifica procesos de Explotación de Información	PARCIAL	NO	NO

Tabla 1. Cuadro Comparativo de Metodologías

6.5 Arquitecturas de Software

Para este proyecto es importante conocer los componentes de una arquitectura de software y presentar la importancia de este. El modelo presentado se adapta a las necesidades del desarrollo de este proyecto, en donde será propuesto en el contexto médico y que ya ha sido utilizado en el área educativa por el desarrollo que será mencionado a continuación; además de esta forma se puede encapsular funcionalidades, permitiendo su reusabilidad en otros campos de investigación y reduciendo el impacto ante un cambio tecnológico.

Una arquitectura de software (AS) según (Clements, 1996) es, a grandes rasgos, una vista del sistema que incluye los componentes principales del mismo, la conducta de esos componentes según se la percibe desde el resto del sistema y las formas en que los componentes interactúan y se coordinan para alcanzar la misión del sistema.

Un patrón proporciona una solución común a un problema específico en un contexto dado; esta solución puede ser total, o parcial (una pieza dentro de un conjunto más amplio).

El Patrón Arquitectónico expresa un esquema de organización estructural fundamental para un sistema de software. Provee un conjunto predefinido de subsistemas, especifica sus responsabilidades, e incluye reglas y guías para organizar las relaciones entre sí. Algunos ejemplos de patrones arquitectónicos más utilizados son:

- **Pipes and Filtres:** Los datos se procesan en flujos que pasan de filtro en filtro, de manera que cada filtro representa un avance en el procesamiento.

- **Blackboard:** Aplicación independiente especializada que colabora para derivar la solución, trabajando sobre una estructura de datos común.

- **Modelo Vista Controlador:** Este modelo es un patrón que define la organización independiente del **modelo** (La parte encargada de la obtención, procesamiento, y almacenamiento de los datos según la acción transmitida desde el controlador), la **vista**

(recibe por parte del controlador los nuevos datos a mostrar, y los representa de forma gráfica para mejor entendimiento del usuario y pueda seguir interactuando con la aplicación) y el **controlador** (es el organizador de la aplicación, decide que hacer según interactúe el usuario con la aplicación)[29], esto se aprecia en la Figura 10.

Figura 10. Arquitectura del Modelo vista controlador– MVC

- **Capas o Layers:** Descompone la aplicación en diferentes niveles de abstracción. Las capas van desde capas específicas de aplicación hasta capas de implementación propias de la parte tecnológica.

El patrón de arquitectura de capas, mencionado anteriormente es una de las técnicas más comunes para dividir sistemas de software complicados. Los beneficios de trabajar con este patrón son[30]:

- Se puede entender una capa como un todo, sin considerar las otras.
- Las capas se pueden sustituir con implementaciones alternativas de los mismos servicios básicos.
- Se minimizan dependencias entre capas.
- Las capas posibilitan la estandarización de servicios.
- Luego de tener una capa construida, puede ser utilizada por muchos servicios de mayor nivel.

Teniendo en cuenta el patrón de capas, y las características mencionadas; el **modelo actuador-indicador** se tendrá en cuenta para proponer la arquitectura a utilizar en este trabajo.

Este modelo consiste en una arquitectura con cuatros capas funcionales (Zimmermann, A., Specht, M., & Lorenz, 2005). Estas capas funcionales son sensor, semántica, control e indicador. La capa **sensor** es el responsable de los registros de información (Logs) de las diferentes interacciones. La

[29]http://blog.cubenube.com/2011/11/la-arquitectura-modelo-vista.html
[30]http://arevalomaria.wordpress.com/2010/12/02/introduccion-al-patron-de-arquitectura-por-capas/

capa **semántica** recoge los datos del registro del sistema mediante el uso de agregadores y éste se refiere a cómo se transforma semánticamente los registros. La capa **control** es el encargado de interpretar la respuesta de los agregadores a través de las diferentes estrategias, es decir cuándo y cómo recoger las respuestas del agregador y presentarlos al usuario. Y finalmente la capa **indicador** se encarga de transformar los datos devueltos por la capa control de tal forma que sea interpretada por los usuarios; este modelo se muestra en la Figura 11.

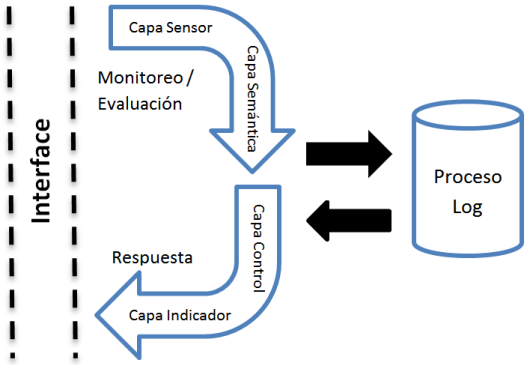


Figura 11. Modelo actuador-indicador. Tomado de (Zimmermann, et al, 2005)

Este modelo es utilizado en un prototipo de un ambiente virtual de aprendizaje – VLE (Florian, Glahn, & Drachsler, 2011), en donde los registros de datos y las interacciones de los estudiantes, son interpretadas y visualizadas en controles de mando según el rendimiento en las actividades de aprendizaje. La implementación de este modelo, como se muestra en la Figura 12, es representado en el área de aprendizaje (Florian, 2013), en donde los datos son analizados y presentados por métodos visuales que permiten la toma de decisiones en el nivel de aprendizaje en los cursos de educación superior.

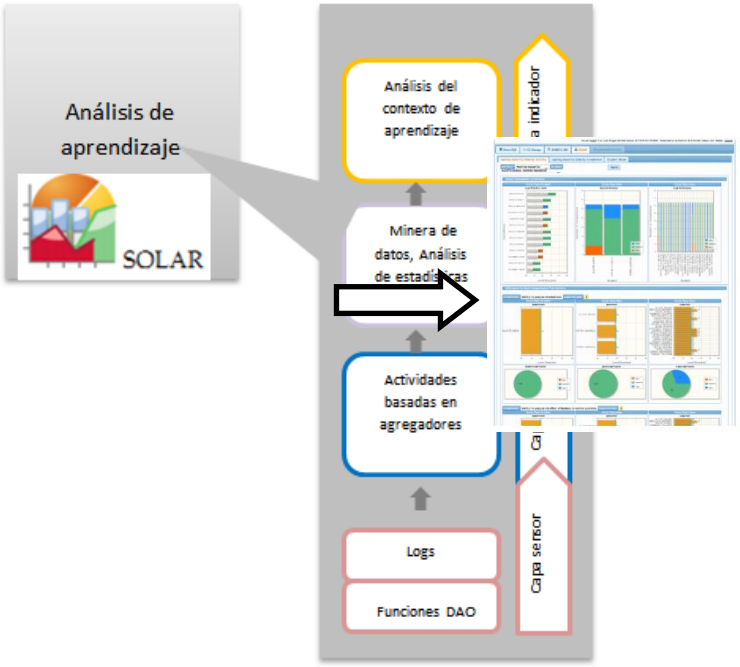


Figura 12. Arquitectura de las Analíticas visuales presentadas en la suite VLE Tomado de (Florian, 2013)

6.6 Grandes Volúmenes de Datos

La información del RPCC es diariamente almacenado en una base de datos relacional mediante el Sistema de Información (SISCAN). Durante 50 años esta información se incrementa anualmente por los casos de cáncer nuevos y algunos tumores manejan información adicional de estudios de investigación; además para completar la información de los casos de cáncer ingresados en el SISCAN se ha vinculado bases de datos de Aseguramiento en salud (SISBEN y Régimen Contributivo) y la Mortalidad de Cali, haciendo que no solamente aumente el volumen por el ingreso de los casos sino también por la actualización de esta información que se realiza aproximadamente cada año. Es de importancia conocer que el volumen de la información manejada en el RPCC es considerado un Big Data por el crecimiento constante de esta, por la variedad de información que se almacena y por la necesidad de analizar los datos de forma óptima.

Los grandes volúmenes de datos (Big Data en Inglés) se refiere a conjuntos de datos que crecen tan rápidamente que no pueden ser manipulados por las herramientas de gestión de bases de datos tradicionales. Sin embargo, el tamaño no es el único problema al que nos enfrentamos si

buscamos una solución: además de almacenarlo, es necesario capturar, consultar, gestionar y analizar toda esta información[31].

Es indispensable que las organizaciones se concentren en el volumen, variedad y velocidad cada vez mayores de la información que conforma el Big data (IDC., n.d.).

- **Volumen.** Existen muchos factores que contribuyen al aumento del volumen de datos: los datos de transacciones almacenados a lo largo de los años, los datos de texto que constantemente generan las redes sociales, la creciente cantidad de datos recopilados de sensores, etc.

- **Variedad.** En la actualidad, los datos se encuentran en todo tipo de formatos: desde las bases de datos tradicionales hasta los almacenes de datos jerárquicos creados por los usuarios finales, pasando por los sistemas OLAP, los documentos de texto, el email, los datos de mediciones, el video, el audio, la información bursátil y las transacciones financieras. Según algunos cálculos, el 80% de los datos de las organizaciones no son numéricos. No obstante, estos también deben incluirse en los análisis y el proceso de toma de decisiones.

- **Velocidad.** Según Gartner(Beyer, n.d.), la velocidad "designa la rapidez con que se generan los datos y con la que deben procesarse para satisfacer la demanda".

Los datos se hallan constantemente en aumento, porque las empresas se mantienen a lo largo del tiempo y los procesos operativos que se relacionan a esta, hacen que el manejo de la información tengan datos estructurados y no estructurados; como lo son documentos, en donde es necesario almacenar o consultar, haciendo que sea difícil de manejarlos para los análisis.

Al tener presente las tres dimensiones mencionadas de un bigdata, existe una cuarta dimensión que puede ser considerada que es la veracidad[32] en donde es importante abordar y gestionar la incertidumbre inherente a algunos tipos de datos.

Teniendo en cuenta dichas características, el RPCC cumple con las dimensiones tanto de **variedad** porque existen datos estructurados, no estructurados, textos; con la **velocidad** que es el tiempo de espera que existe entre el momento en el que se crean los datos, el momento en el que se captan y el momento en el que están accesibles, siendo actualizados constantemente por diferentes usuarios del sistema de información. En cuanto al **volumen,** la información de la base de datos tiene un crecimiento estimado del 20-25% anual; aproximadamente hace 8 años se han vinculado algunos proyectos de investigación que contiene variables complementarias haciendo

[31] http://www.analiticaweb.es/que-es-big-data/
[32] http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf

que se creen nuevas estructuras de tablas para alojar dicha información. Aunque su tamaño hoy en día abarca las **10 gigas**, se debe contemplar dicho crecimiento, dado que el RPCC se ha tenido que adaptar a cambios en su forma de recolección por que los datos en las instituciones manejan en su gran mayoría la historia clinica automatizada de esta forma se da la con la posibilidad de entregar los datos de cáncer en archivos digitales; estos deben ser procesados y anexados masivamente a la base de datos y continuar con los procesos de validación de los mismos.

6.7 Características Generales de Inteligencia de Negocios y Analítica de Negocios.

Para el análisis de los datos es necesario conocer las posibilidades que ofrece la inteligencia de negocios y sus herramientas, así implementarlos con el fin de cambiar la forma de visualizar los datos y tomando decisiones de forma inmediata de acuerdo a la información obtenida.

La inteligencia de negocios (Business intelligent - BI), la podemos definir como la obtención, administración y reporte de los datos orientada a la toma de decisiones, y las técnicas analíticas y procesos computarizados que se usan para el análisis de la misma (Davenport & Harris, 2007).

Cuando hablamos de BI tenemos que considerar los diferentes elementos que la constituyen como se describe en la Figura 13, dentro de los cuales están : la base de datos centralizada, el conjunto de herramientas que utilizará el usuario final como analíticas de negocios (Business analytics BA), las relaciones no conocidas entre las variables, que tienen que descubrirse mediante la minería de datos, y metodologías complementarias como BPM (Business Performance Management), las cuales sirven para monitorear el desempeño y obtener ventaja competitiva (E Turban, Aronson, Liang, & Sharda, 2007).

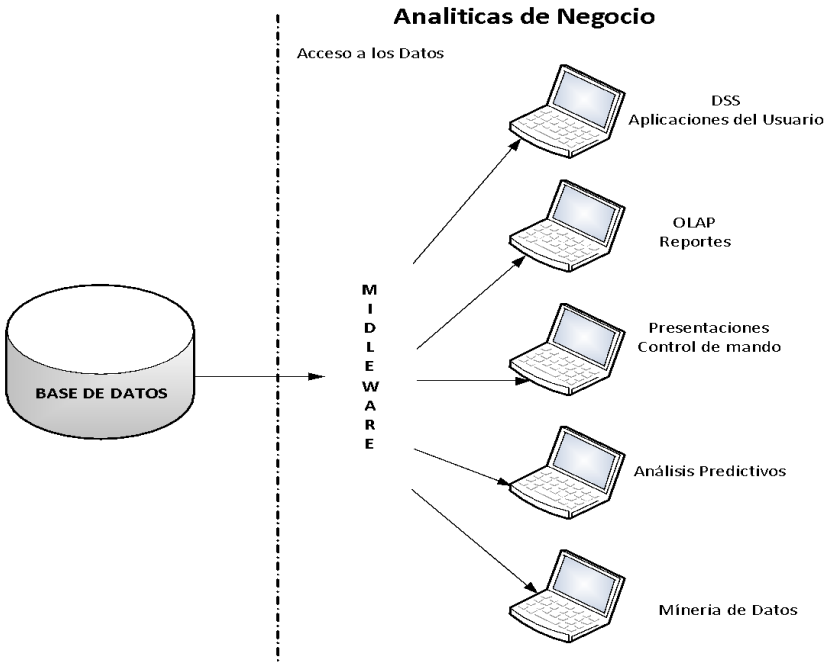


Figura 13. Componentes de la inteligencia de negocio enfocado a analíticas de negocio

Analítica se puede definir como la ciencia del análisis (de datos), y analítica de negocio como el gran conjunto de aplicaciones y técnicas para obtener, almacenar, analizar y permitir acceso a la data que ayude a los usuarios de la empresa a tomar mejores decisiones estratégicas de negocios.

BA puede ser definido como parte de la inteligencia de negocios y está formado por un conjunto de herramientas que permiten el análisis para la toma de decisiones, las cuales incluyen a OLAP, la multidimensionalidad, visualización de datos, Sistemas de información geográficos (geographic information systems - GIS), minería de datos y técnicas analíticas avanzadas. BA le proporciona los modelos y procedimientos analíticos a la Inteligencia de Negocios, de manera de buscar una ventaja competitiva.

El concepto de Inteligencia de negocios se ha ganado la atención masiva entre los profesionales de la salud y los profesionales de la asistencia sanitaria en cuanto a su aplicabilidad en la historia clínica electrónica, en el cual se combina el análisis de conjuntos de datos estructurados y no estructurados que puede resultar fructífera y útil. Existe la necesidad de aplicar tecnologías de minería de datos para extraer datos de calidad con el fin de proporcionar apoyos de decisiones en tiempo real y basada en la evidencia de los médicos y los profesionales de la salud (Bonney, 2013).

(Microstrategy, 2006) por otro lado clasifica las herramientas de BA en cinco categorías:

1) Reportes Empresariales (del tipo estandarizado, lo que tendrán una gran distribución dentro de la organización).
2) Análisis del cubo (para análisis simples del tipo OLAP en cubos multidimensionales)
3) Búsquedas y análisis específicos (para análisis relacionales complejos OLAP con los que buscarán en toda la base de datos y a gran detalle)
4) Estadísticos y de Minería de Datos (para los análisis más complejos mediante el uso de estadística, matemática y minería de datos, los cuales descubrirán correlaciones, relaciones causa-efecto y análisis predictivos y otros), y
5) De envío de reportes y alertas (mediante el uso de motores de distribución de reportes, basados en suscripciones, horarios, límites y alertas).

Las técnicas de BI se pueden aplicar a las organizaciones de salud, los cuales manejan grandes volúmenes de datos y en su mayoría existen inconsistencias o carecen de integridad. Con el fin de hacer que esta información sea más accesible y comprensible es importante implementar visualizaciones de datos que proporciona un mecanismo óptimo para dicho análisis (Wagar Haque, 2014).

Es importante que al contener información, se aplique la inteligencia de negocios para incrementar la eficiencia y analizar el comportamiento de estos datos almacenados, de esta forma tomar decisiones correctas en un corto plazo.

6.8 Analíticas Visuales (Visual analytics – VA)

El principal propósito para utilizar las analíticas visuales es mostrar la información del RPCC por medio de controles de mando implementando gráficos dinámicos y de esta forma entender y analizar la información almacenada durante todo este tiempo.

La Analítica visual es la ciencia del razonamiento analítico con el apoyo de interfaces visuales (Thomas & Cook, 2005). Hoy en día, los datos se producen a un ritmo increíble y la capacidad de recolectar y almacenar los datos está aumentando de forma más rápida que la capacidad de analizarla.

En las últimas décadas, se ha desarrollado un gran número de métodos automáticos de análisis de datos. Sin embargo, una definición más específica sería: “VA es la técnica de combinar análisis automatizado con métodos visuales que permiten el razonamiento y la toma de decisiones de

grandes cantidades de datos almacenados (D. Keim, Kohlhammer, Ellis, & Mannsmann, 2010). En la Figura 14 se muestra las áreas de investigación relacionadas con VA, en donde se refiere a las áreas de la visualización de la información e informática gráfica, y al análisis de los datos, que se beneficia de las metodologías desarrolladas en el ámbito de la recuperación de información, la gestión y el conocimiento de los datos, así como la minería de datos.

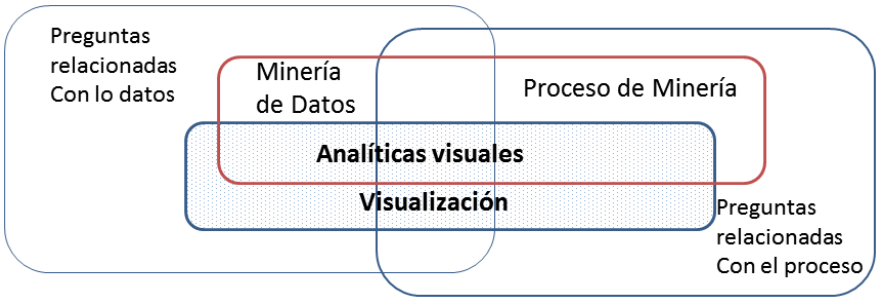


Figura 14. Analíticas visuales combinadas con técnicas de análisis

En muchas áreas de aplicación, el éxito depende de que la información sea disponible en el momento adecuado y cualquier tecnología que pretenda superar estas cantidades de información tiene que dar respuesta a los siguientes problemas:

- ¿Quién o qué define la "relevancia de la información" para una tarea determinada?
- ¿Cómo pueden los procedimientos adecuados apoyar una toma de decisión compleja, cuando el proceso es identificado?

- ¿Cómo puede la información obtenida ser presentada para la toma de decisiones?
- ¿Qué tipo de interacción puede facilitar resolver los problemas tomando decisiones?

En general, la forma de resolver estas problemas es utilizando analíticas visuales, en donde la visión general es convertir de forma transparente los datos y la información para un resultado analítico, proporcionando una forma de divulgación acerca de ellos (D. Keim, Andrienko, Fekete, & Carsten, 2008) como se muestra en la Figura 15.

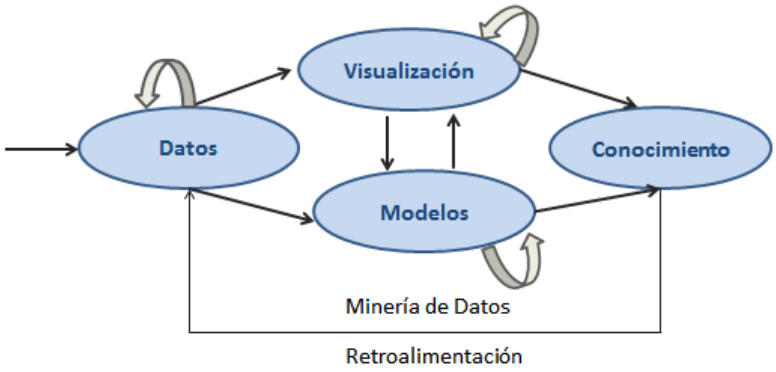


Figura 15. Integración de los métodos visuales y automáticos de análisis de datos.

Algunas de las herramientas que permiten visualizar esta información que se almacena en las bases de datos son:

6.8.1 Sistemas de Soporte a la Decisión (Decision Support Systems– DSS)

Los sistemas de soporte a la decisión, DSS son sistemas de información basados en computadora los cuales combinan modelos y datos para intentar resolver problemas no estructurados utilizando una interfaz amigable para el usuario (Efraim Turban & Aronson, 2001).Asimismo(Olson & Courtney, 1992), dicen que un sistema de soporte a la decisión es el uso de cómputo interactivo para aprender acerca de problemas de decisión, frecuentemente a través de accesos a datos y modelos.

Un **DSS** es una herramienta de inteligencia de negocio (*Business Intelligence*) enfocada al análisis de los datos de una organización. Las principales características son (Efraim Turban & Aronson, 2001):

- **Informes dinámicos, flexibles e interactivos**, de manera que el usuario no tenga que ceñirse a los listados predefinidos que se configuraron en el momento de la implantación, y que no siempre responden a sus dudas reales.
- **No requiere conocimientos técnicos**. Un usuario no técnico puede crear nuevos gráficos e informes y navegar entre ellos, haciendo *drag&drop* o *drillthrough*. Por tanto, para examinar la información disponible o crear nuevas métricas no es imprescindible buscar auxilio en el departamento de informática.
- **Rapidez en el tiempo de respuesta**, Este tipo de bases de datos están optimizadas para el análisis de grandes volúmenes de información (Análisis).

- **Cada usuario dispone de información adecuada a su perfil**. No se trata de que todo el mundo tenga acceso a toda la información, sino de que tenga acceso a la información que necesita para que su trabajo sea lo más eficiente posible.
- **Disponibilidad de información histórica**. En estos sistemas está a la orden del día comparar los datos actuales con información de otros períodos históricos de la organización, con el fin de analizar tendencias, fijar la evolución de parámetros de negocio… etc.

Los DSS permiten a los usuarios fácilmente tomar ventaja de la información que se encuentra previamente almacenada en los repositorios, teniendo posiblemente una vista por medio de algún diagrama, gráfica o algún formato en específico(Cross, 2002). Esto, no se limita a un área específica sino puede abarcar un área geográfica amplia en la organización, lo cual puede brindar todos los requerimientos necesarios.

6.8.2 Control de mandos (Dashboard)

Según (Ballvé, 2007)"El concepto de control de mando (CM), parte de la idea de configurar un cuadro de información cuyo objetivo y utilidad básica es diagnosticar adecuadamente una situación. Se lo define como el conjunto de indicadores cuyo seguimiento periódico permitirá contar con un mayor conocimiento de la situación de la empresa o sector". El CM ha demostrado ser un excelente soporte para la dirección cuando está integrado a un buen sistema interactivo. Algunas de sus características son:

a. Incluye información que cambia de manera constante.
b. Brinda información suficientemente significativa para el logro de las metas.
c. Su información es actualizada constantemente y evaluada de acuerdo su desarrollo.
d. Está diseñado para facilitar el análisis

Un CM es una herramienta de acción a corto plazo de implementación rápida y estrechamente ligado a los puntos clave de decisión y de responsabilidad de la empresa" (López Viñeglas, 1999). En términos generales el Control de Mando es una herramienta de ayuda a la gestión, en sí mismo no es un objetivo, sino un efecto que ha de estar orientado hacia la acción (Norton & Kaplan, 1992).Existen dos tipos de Control de Mando:

1. Analíticos o Dashboard, que permite obtener informes e indicadores clave. Son operativos o tácticos y analizan áreas de negocio no relacionadas entre sí (Figura 16).

Figura 16. Ejemplo de control de mando analítico (Dashboard)

2. Integrales o Balanced Scorecard, se desarrollan a nivel estratégico de toda la organización, así los diferentes niveles de gestión y dirección de la organización disponen de una visión estratégica con un conjunto de objetivos e indicadores que abarcan a toda la organización.

Para diseñar Control de mando, es necesario tener en cuenta 4 principios (Gónzalez, n.d.):

1. **Simplicidad.**
   - Un CM debe ocupar una página y preferiblemente con orientación horizontal.
   - No Incluir demasiadas tablas o listas.
   - No incluir cantidad de objetos innecesarios
2. **Uso adecuado de los espacios.** Un dashboard solo debe aparecer información relevante y que aporte conocimiento para la toma de decisiones.
3. **Intuitivo.** Cualquier persona debería poder entender qué tipo de información ofrece y su contexto.
4. **Navegable.** Contener links y/o botones para su navegación.

Las ventajas de unos CM bien construidos son evidentes[33], como por ejemplo:

- Mejora en la toma de decisiones y el rendimiento: facilidad para identificar y corregir tendencias negativas, posibilidad de llevar a cabo análisis no previstos, mejor toma de decisiones basado en la información recogida en un sistema de BI, etc.

[33] http://www.dataprix.com/blogs/respinosamilla/teoria-cuadros-mando-tarjetas-puntuaci-n-dashboard

- Mejora en la eficiencia de los empleados: incremento productividad, tiempos de análisis más reducidos al fusionar varios informes en uno, reducción de tiempos de aprendizaje, reducción de la necesidad de crear nuevos informes.

- Motivación del empleado: el usuario puede generar nuevos informes siguiendo las nuevas tendencias, es más agradable trabajar con gráficos que con los viejos informes, el usuario utilizará más tiempo en el análisis y menos en la elaboración de la información. Además, pueden ser una herramienta para compartir estrategias, tácticas y datos de los sistemas operacionales que permitan al empleado una mejor comprensión de objetivos y una mejor toma de decisiones.

Poder visualizar los datos almacenados en una base de datos de forma entendible como las gráficas, independientemente de los campos específicos (medicina, geofísica, geografía, educativo, etc.) donde se aplican estas visualizaciones; es una ventaja que puede ayudar a interpretar mejor los resultados y tomar las decisiones en corto plazo.

7 Estado del arte

El RPCC ha utilizado diferente software que han ayudado a almacenar la información siendo de gran ayuda para su progreso. Actualmente el Registro de Cáncer estableció su propio sistema de información (SISCAN) y ha sido utilizado desde el 2009 hasta el momento, a pesar que este sistema ha sido importante avance tecnológico por estar en la Web y ser accedida por el personal del RPCC ingresando los datos directamente desde las instituciones donde recolectan la información; este actual sistema no contiene un manejo de estadísticas por eso el RPCC utilizado algunas herramientas para el manejo de estos datos.

Los sistemas de información son un conjunto de componentes interrelacionados que recolectan (o recuperan), procesan, almacenan y distribuyen información para apoyar la toma de decisiones y el control de una organización. (Laudon & Laudon, 2004)

Los planificadores de salud y los encargados de adoptar decisiones de alto nivel del sector de la salud a nivel nacional necesitan integrar y analizar información y evidencia para apoyar las políticas de salud, la planificación y la toma de decisiones, en una amplia gama de áreas relacionadas con salud pública y sistemas de salud

Existen algunos sistemas para el análisis de datos relacionados con cáncer, los cuales son:

CanReg5[34] es una herramienta de código abierto diseñada para introducir, almacenar, verificar y analizar los datos de los registros de cáncer. Tiene módulos para realizar la introducción de datos, el control de calidad, las comprobaciones de consistencia de los datos y el análisis básico de los datos.

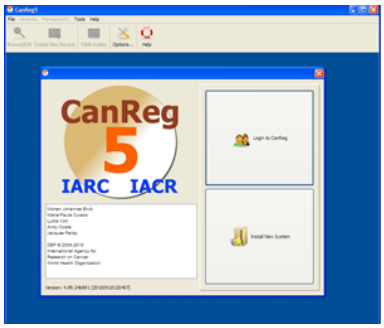


Figura 17.Herramienta para almacenar, verificar y analizar registros de cáncer.

La versión anterior, CanReg4 fue utilizada en el RPCC durante aproximadamente 5 años; pero esta herramienta no suplió las necesidades, siendo un software de escritorio que no se podía trabajar de forma compartida y además no se podía ingresar múltiple información. La versión de CanReg5 fue analizada en el RPCC y los anteriores defectos encontrados en la versión anterior fueron resueltos; aunque actualmente se puede realizar un básico de los datos, no suplió las necesidades de este, porque no se podían vincular tablas de datos externas que proporciona el sistema actual del Registro de Cáncer de Cali.

IARCcrgTools[35] es un paquete que permite a los registros de cáncer validar y explorar sus datos. El programa IARCcrgTools permite realizar conversiones de los códigos, tanto topográficos como morfológicos, entre las distintas ediciones de las Clasificaciones Internacionales de Enfermedades (CIE), las distintas ediciones de las Clasificaciones Internacionales de Enfermedades para Oncología (CIE-O) o entre dichas clasificaciones (CIE y CIE-O).

[34] http://www.iacr.com.fr/canreg5.htm

[35] http://www.iarc.com.fr/software-frame.htm

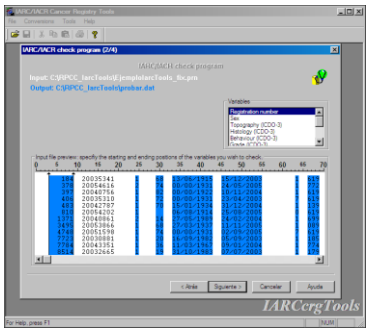


Figura 18. Herramienta verificar los datos de un registros de cáncer

Este paquete se utiliza actualmente en el RPCC, siendo muy útil para la verificación de la información; aunque en el sistema actual se han integrado la mayoría de verificaciones que realiza este software. El inconveniente de este software es que no toma los datos directamente de la base de datos, sino que se deben exportar las variables a un archivo de texto, quedando de alguna forma desactualizado según los procesos del RPCC; además sería de mucha ayuda que mostrará de alguna forma la relevancia de la información como se encuentra distribuida, ya que las variables que se usa en este software son las necesarias para medir este tipo de enfermedad.

Joinpoint Regression Program [36] es un software estadístico para el análisis de las tendencias temporales utilizando modelos joinpoint (de punto de cambio), es decir, modelos que ajustan los datos a una tendencia, seleccionando el modelo más simple que los datos permiten, dónde varios modelos lineales se conectan entre sí en los puntos de intersección o joinpoint, formando splines lineales.

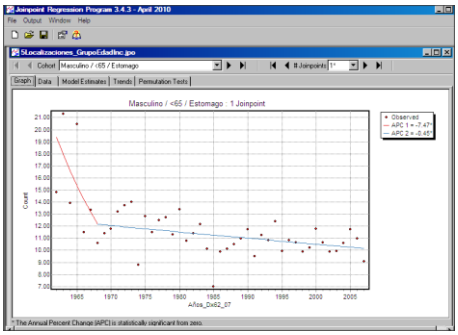


Figura 19. Software estadístico para el análisis de las tendencias temporales

Este software presenta la información de forma visual mediante un gráfico de líneas que estadísticamente muestra una regresión lineal o ajuste lineal por medio de un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente. En

[36] http://surveillance.cancer.gov/joinpoint/

el RPCC se utiliza para encontrar los puntos de cambio que existen durante los diferentes años de diagnóstico para un cáncer determinado.

CanSurv[37] es un software estadístico para analizar datos de supervivencia basados en estudios de base poblacional, que puede ajustar distintos modelos de supervivencia, como por ejemplo, los modelos paramétricos o el modelo de riesgos proporcionales de Cox. Además, los factores pronósticos y las variables demográficas que se relacionan con la supervivencia del paciente con cáncer, pueden utilizarse como variables en el modelo.

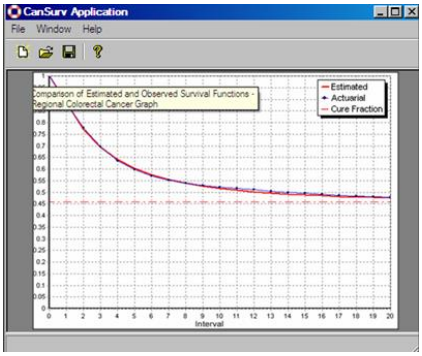


Figura 20. Software estadístico para análisis de supervivencia

Este software presenta la información mediante un gráfico de líneas del comportamiento de supervivencia según el tumor que el usuario defina, esta supervivencia se refiere al porcentaje de pacientes que viven después del diagnóstico de cáncer.

WAERS[38] Web-Assisted Estimation of Relative Survival. Es desarrollado por el Instituto Catalán de Oncología (ICO), es una aplicación web que permite a los registros de cáncer de base poblacional y hospitalaria, y a los registros de enfermedades (en general), estimar la supervivencia relativa de una cohorte de pacientes, seleccionando a la población de referencia que consideren (provincia, comunidad autónoma o país). Este desarrollo no ha sido utilizado por el RPCC, porque se cuenta con análisis propios para dicha estadística.

[37] http:// http://surveillance.cancer.gov/cansurv/
[38] http://rht.iconcologia.catsalut.net/cas/surv.htm

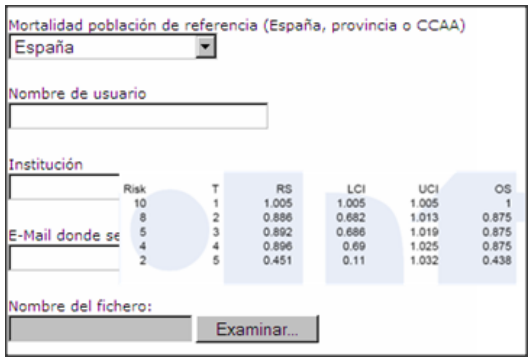


Figura 21. Aplicación Web para registros de cáncer que estima la supervivencia relativa.

Epidat[39](Análisis de Datos Epidemiológicos) es un programa de libre distribución para el manejo de datos tabulados, como respuesta a la necesidad de tener una calculadora para consultas estadísticas y epidemiológicas básicas. El desarrollo del Epidat se integró en el marco de un convenio firmado por la Organización Panamericana de la Salud (OPS) y la Cancillería de Sanidad de la junta de Galicia con el objetivo de colaborar en tareas de investigación en el área de la salud[40]. Este programa no gestiona bases de datos y es orientado al manejo de datos tabulados, ofrece la posibilidad de organizar y resumir un conjunto de datos mediante tablas de frecuencias, cálculo de medidas características y representaciones gráficas, de igual forma realiza diferentes estadísticas como es: pirámides poblacionales, análisis de supervivencia, vigilancia en salud pública entre otras. EpiDat tiene una amplia difusión y es utilizado en más de 40 países con el 98% de usuarios iberoamericanos. Lo anterior lo hace un programa muy utilizado, pero los requerimientos del RPCC van más allá de un análisis que funcione en un entorno local.

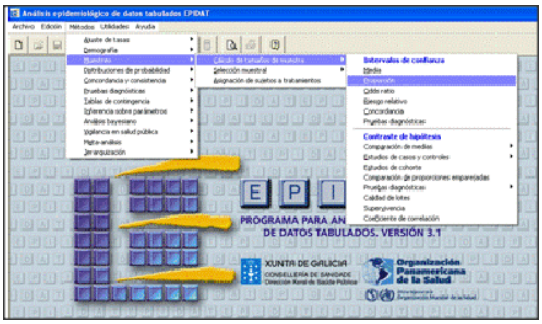


Figura 22. Programa para manejo de datos tabulado

SEER*Stat[41] es un software estadístico que proporciona un mecanismo para el análisis de datos relacionadas con el cáncer. Se trata de una potente herramienta de PC para ver los registros de

[39]http://dxsp.sergas.es
[40]http://www1.paho.org/spanish/sha/epidat.htm
[41]http://seer.cancer.gov/seerstat/

cáncer de las personas y la elaboración de estadísticas para el estudio del impacto del cáncer en una población. Este software presenta datos tabulados y es utilizado en el RPCC para realizar las estadísticas de incidencia y mortalidad.

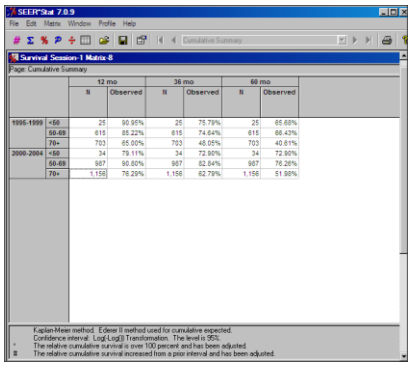


Figura 23. Software estadístico para el análisis de datos relacionados con el cáncer

iSALUS[42]. Es el portal que facilita la gestión diaria de los centros médicos, la generación de negocio entre proveedores y clientes del sector sanitario, y el conocimiento compartido entre profesionales de la medicina de todo el mundo resolviendo online todas las funciones de gestión de los centros médicos de forma eficiente y sin necesidad de instalar ningún tipo de software. Este portal presenta la línea de tiempo de un paciente en información útil, lo que permite un examen simple, pero conciso de la historia clínica del paciente. Mediante la entrega de la información clínica en un formato visual, de esta forma los médicos pueden evaluar rápidamente y comparar los hechos críticos sobre las tendencias cambiantes del paciente.

Figura 24. Software de medición de tendencias del paciente.

[42]http://www.isalushealthcare.com/Resources/InTheNews/tabid/139/EntryId/32/iSALUS-Healthcare-Releases-Next-Generation-of-OfficeEMR.aspx

Este desarrollo está enfocado en la atención del paciente, permitiendo presentar la información inmediata y con ayudas visuales. Este tipo de soluciones son las presentadas actualmente para para mostrar los datos de una forma eficiente y en tiempo real.

Aunque este tipo de aplicativos no hacen parte de los evaluados respecto a Cáncer que se menciona en este documento, se presenta para mostrar los alcances tecnológicos que se pueden implementar en el área de la salud mostrando los alcances a los que se quiere llegar con el desarrollo de este proyecto.

GLOBOCAN[43], proporciona estimaciones contemporáneas de la incidencia, mortalidad y prevalencia de los principales tipos de cáncer a nivel nacional para 184 países del mundo. Las estimaciones GLOBOCAN se presentan para 2008, por separado para cada sexo, están disponibles para la población adulta los datos de prevalencia a 5 años. Estas estimaciones se basan en los datos más recientes disponibles en la Agencia Internacional para la Investigación sobre el Cáncer (IARC) y en la información a disposición del público en Internet, pero las cifras más recientes pueden estar disponibles directamente de fuentes locales.

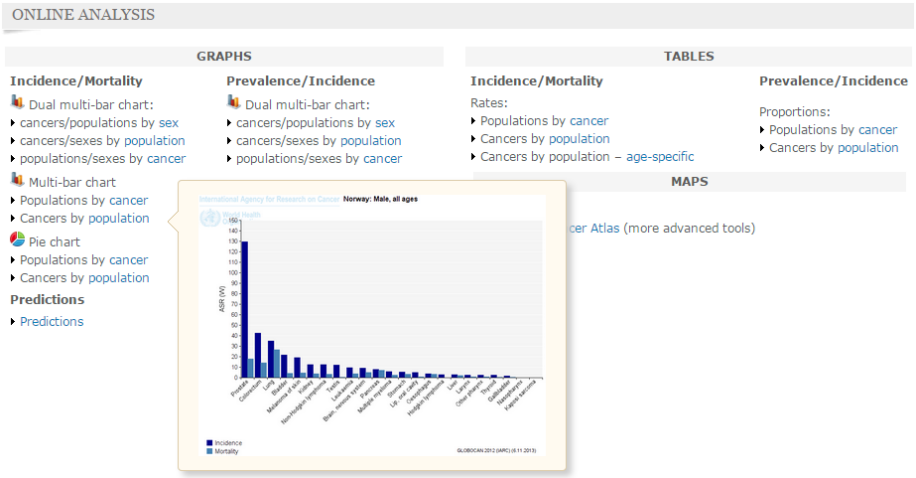


Figura 25. Globocan 2008

SAS® Visual Analytics[44] permite, a todos los perfiles de usuarios, explorar cualquier cantidad datos de forma sencilla y casi instantánea. Descubre correlaciones, predice tendencias, detectar anomalías en tu negocio y genera gráficos e informes que puedes compartir a través de la Web y los dispositivos móviles.

[43]http://globocan.iarc.fr/Default.aspx
[44]http://www.sas.com/offices/latinamerica/mexico/solutions/visual-analytics/overview.html

Este desarrollo sería la implementación necesaria en donde la información es transformada en tiempo real y se presenta visualmente en gráficos desde la web. SAS comercializa paquetes de procedimientos adicionales para el análisis estadístico de los datos.

Figura 26. Visual Analytics SAS

Después de contextualizar algunos softwares para los análisis de la información en su mayoría relacionados con cáncer, se presenta un cuadro comparativo teniendo en cuenta los aspectos principales que se plantearon para evaluar el funcionamiento de las herramientas existentes.

Id	Software	Ventajas	Desventajas
1	CanReg4	Herramiente de codigo abierto. Validaciones de la informacion de Cáncer. Verificacion de duplicados. Filtros según la necesidad del usuario. Analisis estadisticas basicas de la informacion ingresada.	Es un software de escritorio y no se puede utilizar de forma compartida. No tiene una base de datos relacional. No se pueden ingresar tablas anexas de informacion. Informacion de estadistica no publica. No es expandible a mas implementacion de modulos
2	CanReg5	Herramiente de codigo abierto. Validaciones de la informacion de Cáncer. Verificacion de duplicados. Filtros según la necesidad del usuario. Base de datos relacional. Analisis estadisticas basicas de la informacion ingresada. Se puede utilizar en red.	No se pueden ingresar tablas anexas de informacion. Informacion de estadistica no publica. No es expandible a mas implementacion de modulos.
3	IARCcrgTools	Validaciones de la informacion de Cáncer.	Se carga un archivo de texto con los datos a verficar y no s e pudede adaptar directamente al analisis de los datos. No es una herramienta de ingreso de informacion.
Id	**Software**	**Ventajas**	**Desventajas**
4	Joinpoint Regression Program	Software estadístico para el análisis de las tendencias.	Se carga un archivo de texto con los datos a verficar y no s e pudede adaptar directamente al analisis de los datos. No es una herramienta de ingreso de informacion.
5	CanSurv	Software estadístico para analizar datos de supervivencia basados en estudios de base poblacional.	Se carga un archivo de texto con los datos a verficar y no s e pudede adaptar directamente al analisis de los datos. No es una herramienta de ingreso de informacion.
6	WAERS	Aplicación Web que permite estimar la supervivencia relativa.	Se carga un archivo de texto con los datos a verficar y no s e pudede adaptar directamente al analisis de los datos.
7	Epidat	Es una herramienta de manejo sencillo y de utilidad para el análisis estadístico y epidemiológico de datos.	No es una herramienta de ingreso de informacion. Informacion de estadistica no publica.
8	SEER*Stat	Software estadístico para el análisis de las tendencias.	Se carga un archivo de texto con los datos a verficar y no s e pudede adaptar directamente al analisis de los datos. No es una herramienta de ingreso de informacion.
9	iSALUS	Software de medición de tendencias del paciente.	No es precisamente para informacion con cáncer.
10	NTEGRIS Health	Exploración interactiva de datos basada en web y accesible para varios tipos de roles. Capacidades automáticas de visualización y análisis. Facilita el acceso a los datos e informes a todo tipo de usuarios, a través de la Web.	No es una herramienta de codigo abierta, siendo una aplicación de alto costo.

Tabla 2.Ventajas y desventajas de las herramientas para el análisis de información de Cáncer.

Finalmente después de realizar una descripción en el cuadro comparativo de las ventajas y desventajas que tiene cada herramienta de software, se presenta una tabla de caracterización y evaluación de los programas utilizados; de esta forma se hizo uso de un modelo para dicha evaluación de los software teniendo como guía general a la ISO-9126, donde pretende establecer un estándar internacional para la evaluación de la calidad de productos de software[45], el enfoque principal es en la prestación de los datos y los aspectos que puedan tener o no las herramientas para el análisis de información.

[45] http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352009001200003

Caracteristicas / Software	FUNCIONABILIDAD		PORTABILIDAD		PRODUCTIVIDAD		VISUALIZACION		
	Conexión a B.D	Informes, estadisticas	Disponibilidad inmediata	Reemplazabilidad	Esfuerzo del usuario	Costo financiero	Gráfico	Tablas	Datos en Web
CanReg	S	S	N	N	S	N	S	S	N
IARCcrgTools	N	N	N	N	N	N	N	N	N
Joinpoint Regression Program	N	S	N	N	S	N	S	S	N
CanSurv	N	S	N	S	S	N	N	S	N
WAERS	N	S	N	S	N	N	N	S	S
Epidat	S	S	N	S	N	N	S	S	N
SEER*Stat	N	S	N	S	S	N	N	S	N
iSALUS	S	S	S	S	N	S	S	S	S
GLOBOCAN	N	S	S	N	N	N	S	S	S
SAS® Visual Analytics	S	S	S	N	N	S	S	S	S

Tabla 3. Resumen de herramientas de análisis de información de cáncer.

En el resumen presentado en la Tabla 3 de las herramientas de análisis se evaluó dando el enfoque a la parte de análisis de los datos. A continuación se explican las características generales que se tuvieron en cuenta en relación al objetivo del desarrollo del proyecto:

1. **Funcionalidad:** El aplicativo debe tener la capacidad de proveer las funciones que satisfacen las necesidades explícitas e implícitas bajo condiciones específicas. En este caso es importante contar con una conexión directa de los datos, y de esta forma mantener datos en tiempo real, además dicha información debe ser presentada mediante datos estadísticos que representen información necesaria para la toma de decisiones y el entendimiento de los mismos.
2. **Portabilidad:** el aplicativo debe poder transferirse independientemente del entorno de hardware o software y de forma inmediata.
3. **Productividad:** el aplicativo debe permitir a los usuarios emplear cantidades apropiadas de recursos, ser comprendido sin un mayor esfuerzo y que sea al alcance de todos y a un bajo costo.
4. **Visualización:** Finalmente los datos deben permitir ser visualizados mediante la ayuda de tablas y gráficos, así lograr explorar y entender esos datos para convertirlos en conclusiones y conocimiento para una óptima toma de decisiones.

Después de conocer la forma de evaluación que se le realizó a los diferentes aplicativo existentes en la parte del análisis de la información de cáncer, se concluyó que no son del todo completas; porque aunque algunas tienen la opción de presentar los datos de forma estadística pero la mayoría no tiene conexión directa con la base de datos, quedando los datos obsoletos al momento de ser analizados; o estos datos no pueden ser visualizadas directamente en la web y no son adaptadas a los diferentes usuarios.

Al conocer algunas de estas herramientas, se quiere con este proyecto proponer una solución que facilite la visualización de los datos de cáncer de forma inmediata y dinámica. Las características que tendría esa herramienta serían las siguientes:

- Es una herramienta de código abierta.
- Tiene interacción directamente con los datos almacenados en la base de datos del RPCC.
- Información de resultados inmediatos publicados en la Web.
- Puede ser adaptado al Sistema información del RPCC (SISCAN).
- Es expandible a realizar más estadísticas o visualizaciones analíticas.
- Es adaptada a tres usuarios en este desarrollo (Personal del RPCC, Médico especialista "Hemato-oncólogo Pediatra" y usuarios en general).

8 Introducción al Desarrollo del Proyecto

En esta sección se explica la forma en que se representará el desarrollo del proyecto propuesto en este documento y la finalidad lograda en el desarrollo del Aplicativo Web.

El propósito fundamental de este proyecto es presentar la información de los datos de cáncer en cali, mediante analiticas visuales usando controles de mando que presentan los datos mediante gráficos generales como son: barra, torta, líneas etc y gráficos complejos: burbuja, sunburst etc, de esta forma sean útiles al público en general, al personal de Registro de Cáncer y a proyectos de investigación en cáncer que contenga información adicional que pueda ser presentada en dichos gráficos haciendo énfasis en lo visual. Este aplicativo Web llamado "VA_RPCC", es la estrategia para divulgar la información de forma inmediata y en tiempo real.

Inicialmente se realizó una minería de datos para corroborar las localizaciones de los tumores principales y teniendo como guía la información divulgada en el portal Web del RPCC; la metodología para la minería fue CRISP-DM teniendo en cuenta sus respectivas fases que se explica de todo el proceso realizado con estos datos en la sección 9 correspondientemente; en la sección 9.2.2 se presenta la descripción de las variables seleccionadas para dicha minería según el tipo de analitica visual a implementar.

En la sección 10 se presenta la arquitectura de software adaptada y las librerías, API´S de georeferenciación y de redes sociales que ha sido usadas en dicha arquitectura propuesta para la visualización de los datos estadísticos mediante controles de mando especializados para las tres analíticas del RPCC (Generales, Registro de Cáncer y Médico especialista). En este caso para el desarrollo del proyecto se definen solamente tres tipos de usuarios con el fin de presentar la posibilidad de adaptar la información y las visualizaciones a cada uno de ellos según la información disponible que exista. El aplicativo Web inicia ingresando a un navegador y al ingresar utilizando un usuario y contraseña que es según el perfil de usuario, asi se visualiza la información de forma inmediata con los gráficos para cada analitica visual, en la sección 11 se presenta de forma más detallada el funcionamiento del aplicativo y a que corresponde cada analitica y gráfica presentada para cada usuario.

Finalmente en la sección 12 se presenta la evaluación realizada del Aplicativo Web y los resultados que se obtuvieron mediante una encuesta en línea a teniendo en cuenta los tres diferentes tipos de usuarios.

A continuación se presentará una breve descripción correspondiente a la información que se mostrará en las visualizaciones para cada usuario del aplicativo Web. En el caso de uso (Figura 27), se define las actividades que deberán realizarse para llevar a cabo algún proceso según el control de mando a visualizar y el perfil del usuario.

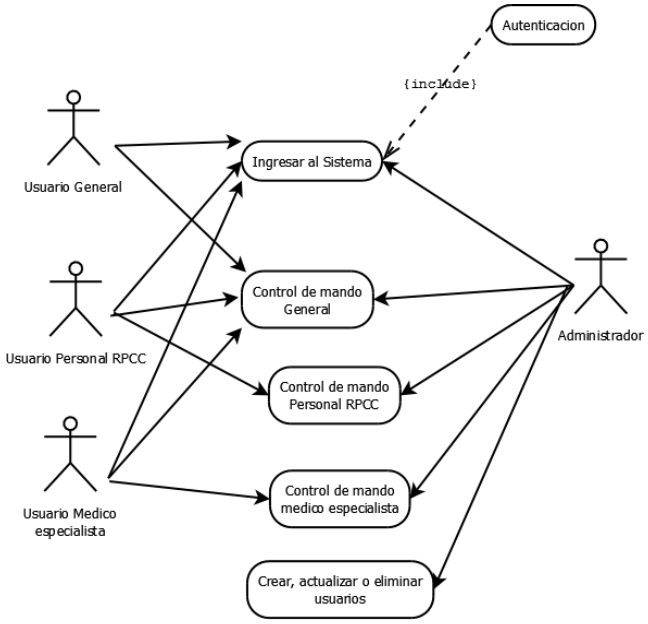


Figura 27. Caso de uso aplicativo Web VA_RPCC

1. **Control de mando General:** Este control es para un tipo de usuario general, el cual se puede visualizar los datos de los casos diagnosticados en Cali durante el periodo de 1962 hasta el 2013. Inicialmente en una tabla se describen los datos quinquenales por sexo según los principales tumores, en la sesión 9 se explica de forma detallada los sitios principales de los tumores; posteriormente se muestran 3 gráficas, una gráfica de barra que representa un conjunto de datos en columnas en los diferentes quinquenios y según el tumor seleccionado; posteriormente se presenta una gráfica de pastel el cual permite comparar los diferentes sitios de los tumores según un periodo de diagnóstico seleccionado, esta gráfica brinda una idea de la distribución de los tumores y se puede

identificar cuales tiene una mayor o menor proporción. Finalmente se presenta la información en una gráfica de líneas para mostrar la tendencia a lo largo de un periodo de tiempo.

2. **Control de mando Personal del RPCC:** En este control se presenta la información general de los datos de todas las residencias mostrando una distribución anual y mediante una grafico de burbuja se representa dicha visualización por el personal del RPCC en donde se identifican las diferentes residencias aparte de Cali, sexo y se muestra otras variables como son el estado vital y estrato socioeconómico. Estos datos se presentan en tres tipos de gráficas (burbuja, distribución de partición y serie de tiempo).

3. **Control de mando médico especialista**: En este control de mando se presenta los datos referente a otra clasificación del Cáncer que es para Cáncer infantil, de igual forma se muestra la información de variables adicionales que hacen parte de este proyecto de investigación como son: seguridad social, afrocolombiano, edad, residencia y estado vital, de igual forma se georeferencia en el mapa de Cali los pertenecientes a esta residencia.

9 Minería de Datos

El propósito de este trabajo es visualizar los datos de forma dinámica, y por eso en necesario identificar la información y la relación entre estos para así adaptar una arquitectura de software que ayudará a la visualización inmediata de los datos. La metodología de minería de datos CRISP-DM. Se ha seleccionado como metodología base para llevar a cabo este trabajo, buscando más que una aplicación rigurosa, una guía de referencia. Está minería será realizada con el fin de demostrar las localizaciones del tumor principales

En esta sección se presentan los resultados de aplicar algunos de los pasos de CRISP-DM en el contexto de la información del RPCC.

9.1 Comprensión del Negocio

Utilizando como referencia CRISP-DM, la “Comprensión del negocio” se basa en los objetivos del proyecto con el fin de definir el problema a resolver y diseñar una propuesta para el cumplimiento de dicho objetivo. En la Tabla *4* , se especifica de manera general las condiciones actuales del proyecto.

Tabla 4. Descripción de la Comprensión del negocio

Dominio de Aplicación	Sector de Salud. Estadísticas de Información de Cáncer en Cali.
Objetivo del negocio	Identificar las variables de interés según tres diferentes tipos de usuarios que puedan tener una mejor visualización de la información de manera inmediata.
Situación Actual	Existe información histórica en una base de datos de pacientes con Cáncer, y para ser visualizada de forma pública y privada se deben realizar varios procesos en diferentes herramientas de software haciendo que se requiera de conocimiento y tiempo para ver estos resultados.
Objetivo de minería de datos	Clasificar la información para visualizar según los diferentes usuarios. Mostrar información de forma inmediata según las variables analizadas.
Plan del proyecto	Explorar las variables para el análisis utilizando la herramienta Weka para el descubrimiento de conocimiento, y usando técnicas de Clusterización.

9.2 Situación Actual de la información

El Registro Poblacional de Cáncer de Cali (RPCC) presenta en su Página Web información correspondiente al periodo desde 1962 a 2008 para incidencia. Estos datos son extraídos de la base de datos cada cuatro años dependiendo el periodo a divulgar la información (Ejemplo: los pacientes diagnosticados con cáncer en el periodo 2009 residentes de Cali), y procesados teniendo en cuenta los estándares de calidad correspondientes para dicha publicación; al procesar la información en los paquetes estadísticos los resultados son adjuntados a otra tabla que contiene los datos de los anteriores periodos presentados en el sitio Web[46], quedando desvinculada de los datos rutinarios donde se almacena la información ingresada diariamente a la base de datos del RPCC. En la Figura 28 se presenta la frecuencia relativa de las diez primeras localizaciones de tumores primarios en Cali durante el periodo 2004-2008 en hombres.

[46] http://rpcc.univalle.edu.co/es/Frecuencias/frames.php

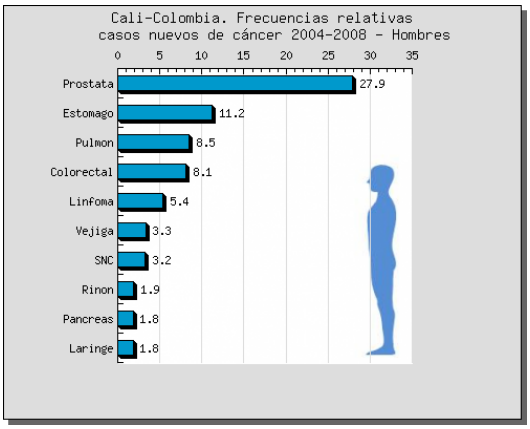


Figura 28. Frecuencia relativa casos nuevos de Cáncer. 2004-2008. Hombres

Si bien esta información es actualizada aproximadamente cada cuatro años en donde se suben los datos anuales después de realizar los procesos de calidad correspondientes para dicha publicación.

Los datos que ya se encuentran registrados en la base de datos hasta la fecha pueden ser de igual forma utilizados para brindar información de forma inmediata mostrando resultados preliminar es y así poder realizar un seguimiento óptimo de la calidad de los datos (Ejemplo: Si se identifica que existen datos codificados como desconocidos en un periodo, es posible poder mejorar esta información de forma inmediata utilizando bases de datos adicionales que se manejan en RPCC).

En la Figura 29 se muestra para el periodo 2013 la información por residencia de los pacientes diagnosticados con Cáncer que se encuentran hasta el momento procesadas y de las que se podría tener una visión preliminar.

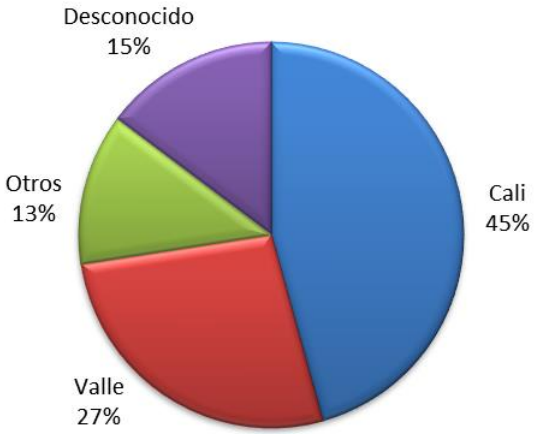


Figura 29. Casos registrados de pacientes con Cáncer durante el periodo 2013 según residencia

En la gráfica anterior se visualiza la información que existe hasta el momento del periodo correspondiente a 2013 con una cantidad de 11,0% de casos registrados y que ya contienen un 15% de residencia desconocida; para completar dicho periodo estaría pendiente por ingresar el 89% de información haciendo que incremente los datos desconocidos de residencia o de otras variables que son necesarias para los análisis.

En la Figura 30, para el periodo 2009 que se considera cubierto a nivel de la recolección de información en donde ya se visitaron las instituciones de salud que brindan dicha información, se muestra que un 1% de los datos contienen residencia desconocida; este periodo ya ha tenido todo el proceso de calidad y verificación de datos desconocidos.

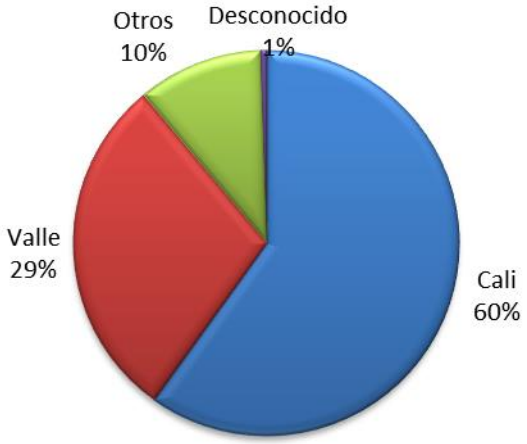


Figura 30. Casos registrados de pacientes con Cáncer durante el periodo 2009 según residencia.

Al comparar de cierta forma dos periodos, el 2009 que ya se puede considerar listo para publicar y el 2013 que contiene una pequeña parte de información recolectada, se muestra que la información puede ser verificada y actualizada sin todo el tiempo como se hace hasta el momento y de esta forma poder tener información preliminar de estos datos tan valiosos.

9.2.1 Metas

Generar visualización de la información inmediata y actualizada permitiendo al RPCC divulgar sus datos y a la sociedad tener conocimiento del Cáncer en Cali, se espera identificar las variables que sean de ayuda para el mejoramiento de la óptima divulgación de estos datos.

9.2.2 Criterio del éxito

El éxito de la implementación de la metodología es definir los datos que ayudan a cumplir el objetivo, esperando que se pueda:

- Visualizar los datos para el público en general según los presentados en la página Web actual.
- Identificar las variables de interés para los tres diferentes usuarios.
- Definir los tipos de gráficas que serán presentados para los diferentes tipos de usuarios según las variables definidas.

9.2.3 Objetivo cumplido

Se determinará que el objetivo se habrá cumplido cuando se implementen los resultados en las gráficas y tablas permitiendo brindar información a los usuarios definidos para una toma de decisiones.

9.2.4 Solución actual

Los datos que se presentan y analizan para estudios específicos son manejados de forma independiente utilizando herramientas que ayudan a visualizar la información en tablas y gráficos, este proceso depende del conocimiento de los datos y cada vez que se hace una análisis los datos son exportados a archivos planos que hacen que la información quede inmediatamente desactualizada.

De igual forma, se puede reconocer que este proceso mejoraría implementando de alguna forma la salida de los datos a gráficos que puedan ser accedidos en cualquier momento y con información actualizada.

9.2.5 Recursos

Se usará la información que se encuentra en la base de datos desde el año 1962 hasta la fecha.

Los datos existentes serán analizados y se implementarán en las gráficas definidas para un mejor entendimiento de los usuarios.

En la parte técnica se cuenta con:

- Servidor de Base de Datos
- Herramienta de pre-procesamiento y generación de modelos (Weka 3.7.4)
- PHP para el desarrollo del aplicativo Web.

9.3 Comprensión de los Datos

Antes de comenzar un proceso de minería de datos es necesario identificar los datos que se van a utilizar de una forma más detallada y se definirá qué datos son relevantes o sirven para dichos análisis.

Se dispone de un conjunto de tablas relacionales en una base de datos en Postgres, correspondientes desde 1962 hasta la fecha. Esta base de datos consta de 14 Tablas principales y 99 tablas que corresponden a diccionarios pre-codificados.

Según las analíticas y el desarrollo que se implementa existen tres diferentes usuarios para la visualización de la información, pero solo en dos de estos usuarios sería necesario aplicar un análisis de minería de datos, ya que estos podrían contener patrones que no son evidentes para la toma de decisiones.

9.3.1 Recolectar datos iniciales

Los recursos disponibles de información para este proyecto son los datos recolectados por el personal del Registro de Cáncer y las fuentes de datos son los siguientes:

- **Encuesta de Morbilidad de Cáncer.**
  Es una herramienta para la recolección de la información sobre incidentes de neoplasias malignas en los residentes de la ciudad de Cali, mediante la visita de los recolectores a las diferentes instituciones de salud, tales como, hospitales, clínicas, centros médicos, centros de tratamiento, EPS, IPS, laboratorios de patología y médicos especialistas.

- **Aplicativo Web – SISCAN**
  Desarrollo web el cual puede ser accedido desde cualquier parte del mundo mediante la dirección: rpcc.univalle.edu.co/siscan, y por medio de éste se puede acceder a la base de datos del Registro Poblacional de Cáncer de Cali, de acuerdo al perfil de usuario.

- **Bases de datos externas:**
  Estas bases de datos contienen información que ayudan a completar datos faltantes que no se han captado en la recolección con la encuesta de morbilidad por cáncer.

    - **Notificaciones Hospitalarias:** Es toda la información que es suministrada por las instituciones de salud en medio magnético de los pacientes hospitalizados, consulta externa, urgencias y unidad oncológica, de acuerdo a las variables solicitadas por el registro Poblacional de Cáncer de Cali en el Manual para las Instituciones.

    - **Aseguramiento en Salud**
      Es la información suministrada en medio magnético de las personas que se encuentran afiliadas a los aseguramientos (Régimen contributivo y Sisben).

    - **Mortalidad**
      Es toda la información suministrada en medio magnético por la Secretaría de Salud Pública Municipal (SSPM) de las personas que han fallecido en Cali.

9.3.2 Descripción de los datos

En esta sección se muestra una descripción detallada de los datos, descripción del tipo de dato y su valor desconocido. La tabla principal de la BD del Registro de Cánceres ***Pacientes*** el cual contiene 199.667 registros de los datos demográficos, del diagnóstico del tumor, entre otras de personas con Cáncer registrados durante el periodo mencionado.

1. **Demográfica**
   Contiene los atributos de una persona. A continuación se describen las variables.

2. **Tumor**
   Contiene los atributos que define un caso de tumor. A continuación se describen las variables:

Nombre	Descripción	Valor	Desconocido	Tipo de dato
Fecha de diagnóstico		día-mes-año	-	Fecha
Localización del tumor (CIEO-3[42])		000-809	999	Texto
Clasificación ICCC[43]		I-XII		Numérico

3. Información complementaria

Existen variables que complementan datos específicos demográficos y/o del tumor. A continuación se describen las variables:

Nombre	Descripción	Valor	Desconocido	Tipo de dato
Afrocolombiano		1:Si 2:No	9	Numérico
Seguridad Social		1:POS 2:POSs 3:Subsidiado 4:Otros	9	Numérico

9.3.3 Exploración de los datos

En la exploración de los datos se analizará un conjunto de información mediante gráficos y tablas, utilizando dos herramientas, la primera STATA 10.0 para obtener estadísticas descriptivas y la segunda WEKA para el pre procesamiento y filtrado de datos.

Utilizando WEKA se aplicará la técnica de clusterización utilizando las variables correspondientes a cada usuario. El algoritmo definido para realizar la clusterización es el algoritmo de k-medias, el cual entregará los k segmentos según un k que recibe como parámetro; para encontrar ese k óptimo por segmento se aplicará otro algoritmo de clusterización, llamado Expectation Maximization (EM), este algoritmo entregará el clúster óptimo, buscando los clúster más probables dados los datos.

Cada modelo de clustering será calculado sobre los datos históricos del RPCC siendo tres diferentes conjunto de datos para el procesamiento de la información

	Entrada	Salida
	Año de diagnóstico (Categorías) Localización del tumor Sexo (Valores categóricos)	Agrupamiento de elementos con similitud entre sí o diferencias.
	Año de diagnóstico Sexo (Valor categórico) Localización del tumor Estatro Socioeconomico residentes Cali (Variable Categórica) Estado vital (Variable categórica) Residencia(Valores categóricos)	Agrupamiento de elementos con similitud entre sí o diferencias.
	Clasificación del tumor (Variable categórica) ICCC (Variable categórica) Edad (Variable categórica) Afrocolombiano (Variable categórica) Seguridad social (Variable Categórica) Estatro Socioeconomico residentes Cali (Variable Categórica)	Agrupamiento de elementos con similitud entre sí o diferencias.

Tabla 5. Entrada de datos según analítica visual

Los datos que serán analizados del conjunto 1 y 2 para realizar la minería de datos, será una muestra de los casos incidentes de pacientes con cáncer ingresados en la base de datos desde el año 1998 hasta la fecha, siendo aproximadamente 15 años de información; de esta forma se quiere demostrar las localizaciones principales de los tumores según el sexo (Tabla 6) teniendo como referencia la información publicada en el sitio Web del RPCC[47] correspondiente a 2003-2007. Esta sería la base de la información que se pretende descubrir al realizar la minería de datos y lograr coincidir con los datos presentados en el RPCC.

Sitio del tumor	Frecuencia relativa	
	Hombres	Mujeres
Prostata	28,3	-
Mama	-	23,7
estomago	11,4	7,2
Cervix	-	10,3
Pulmon	8,5	4,5
Colon y recto	7,2	7,2
Tiroides	-	5,6
Linfomas	5,5	4,1
Ovario	-	4,2
Leucemias	3,6	3,2

Tabla 6. Frecuencia relativa casos nuevos de cáncer en el RPCC. 2003-2007.

[47] Sitio Web Registro de Cáncer de Cali: rpcc.univalle.edu.co

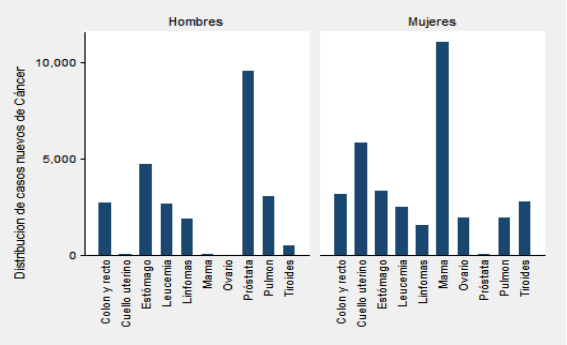
Por último el conjunto 3 son los casos incidentes de pacientes con cáncer infantil menores de 20 años ingresados en la base de datos desde el año 2009 hasta la fecha.

9.3.3.1 Exploración de variables conjunto de datos 1 y 2

Para estos dos conjuntos de datos se manejan tres variables iguales (año de diagnóstico, localización del tumor y sexo), pero en el segundo se anexan otras variables que son identificadas en el RPCC que sirven para verificar la calidad de la información. La distribución de los datos para el análisis es la siguiente:

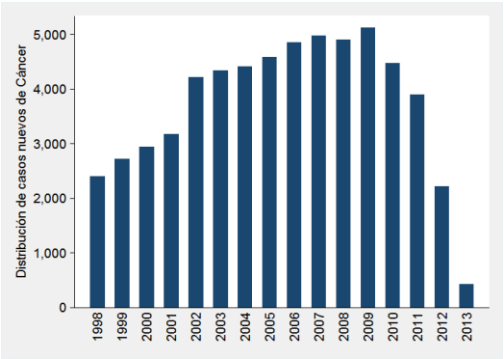
Cali, Colombia. Distribución de casos de Cáncer del RPCC según localización del tumor y sexo.

Localizacion Tumor	Sexo			Total
	Hombres	Mujeres	Desc.	
Colon y recto	2,77	3,201	2	5,973
Cuello uterino	2	5,854	1	5,857
Estómago	4,747	3,357	4	8,108
Leucemia	2,691	2,506	1	5,198
Linfomas	1,889	1,593	1	3,483
Mama	88	11,107	5	11,2
Ovario	0	1,943	1	1,944
Próstata	9,605	18	1	9,624
Pulmon	3,069	1,95	1	5,02
Tiroides	517	2,799	1	3,317
Total	**25,378**	**34,328**	**18**	**59,724**



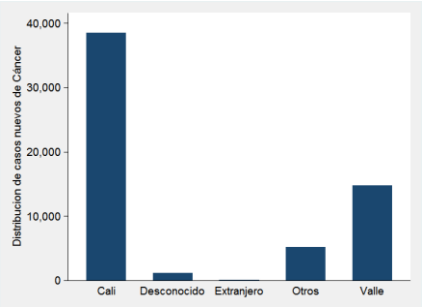
Cali, Colombia. Distribución de casos de cáncer en el RPCC según periodo de diagnóstico.

Periodo	n	%
1998	2,407	4.03
1999	2,716	4.55
2000	2,946	4.93
2001	3,18	5.32
2002	4,223	7.07
2003	4,346	7.28
2004	4,412	7.39
2005	4,586	7.68
2006	4,858	8.13
2007	4,986	8.35
2008	4,912	8.22
2009	5,125	8.58
2010	4,472	7.49
2011	3,904	6.54
2012	2,218	3.71
2013	433	0.73
Total	**59,724**	**100.00**



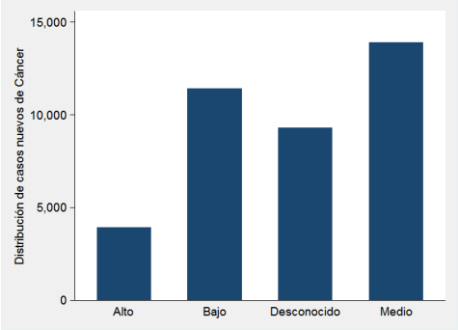
Cali, Colombia. Distribución de casos de cáncer en el RPCC según residencia.

Residencia	n	%
Cali	38,532	64.52
Valle	14,785	24.76
Otros	5,201	8.71
Extranjero	78	0.13
Desconocido	1,128	1.89
Total	**59,724**	**100.00**



Cali, Colombia. Distribución de casos de cáncer en el RPCC según estrato socioeconómico

Residencia	Estrato socioeconómico				Total
	Alto	Bajo	Desc.	Medio	
Cali	3,922	11,419	9,297	13,894	38,532
Valle	4	18	14,727	36	14,785
Otros	1	3	5,19	7	5,201
Extranjero	0	0	78	0	78
Desconocido	1	1	1,125	1	1,128
Total	**3,928**	**11,441**	**30,417**	**13,938**	**59,724**



Cali, Colombia. Distribución de casos de cáncer en el RPCC según estado vital

Estado vital	n	%
Vivo	34,364	57.54
Muerto	22,692	37.99
Desconocido	2,668	4.47
Total	**59,724**	**100.00**

Figura 31. Distribución de casos de Cáncer del RPCC para el conjunto de datos 2.

Las estadísticas descriptivas se muestran en la detallan el comportamiento de las variables, donde se presentan las frecuencias relativas de la información. Seguidamente se exponen los resultados estadísticos obtenidos de cada una de las variables numéricas (Tabla 7).

Variables	Periodo	Sitio del tumor	Residencia	Sexo	Estrato	Estado Vital
No. de valores utilizados	59724	59724	59724	59724	59724	59724
Media	2.005.477	4.525.683	1.576.736	1.577.188	5.993.236	1.932.757
Desviacion Standard	393.811	1.890.473	1.218.116	.5108072	3.174.051	1.601.376
No. de Valor Minimo	1998	16	1	1	1	1
No. de Valor Máximo	2013	77	9	9	9	9

Tabla 7. Estadísticas descriptivas del conjunto de datos 1 y 2

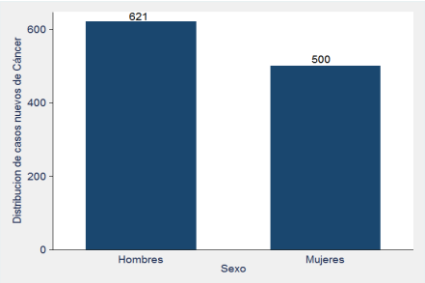
Para la descripción de los datos de la Tabla 6 se utilizaron los valores pre-codificados utilizados en la base datos del RPCC.

9.3.3.2 Exploración de variables conjunto de datos 3

En este conjunto de datos se analizaran la información de cáncer infantil específicamente desde el periodo 2009, y existen variables complementarias que son de gran importancia para este análisis de datos.

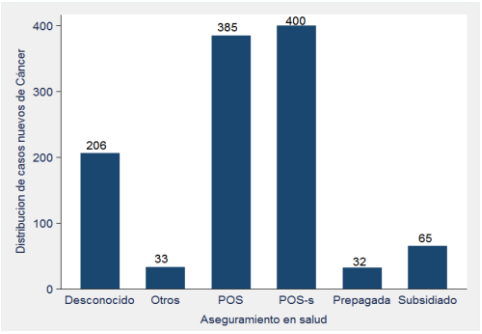
Cali, Colombia. Distribución de casos de cáncer infantil en el RPCC según sexo. 2009-2014

Sexo	n	%
Hombres	621	55.4
Mujeres	500	44.6
Total	1,121	100.0



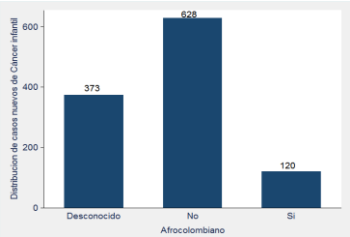
Cali, Colombia. Distribución de casos de cáncer infantil en el RPCC - Vigicáncer según aseguramiento. 2009-2014

Seguridad social	n	%
POS	385	34.3
POS-s	400	35.7
Prepagada	32	2.9
Subsidiado	65	5.8
Desconocido	206	18.4
Otros	33	2.9
Total	1121	100.0



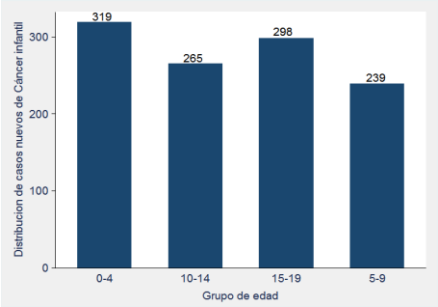
Cali, Colombia. Distribución de casos de cáncer infantil en el RPCC - Vigicáncer según Afro Colombiano. 2009-2014

Afro colombiano	n	%
Si	120	10.7
No	628	56.0
Desconocido	373	33.3
Total	1121	100.0

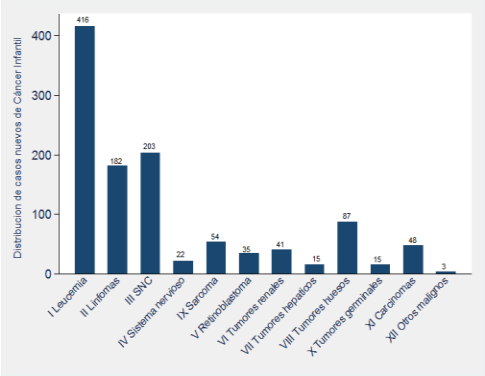


Cali, Colombia. Distribución de casos de cáncer infantil en el RPCC - Vigicáncer según edad. 2009-2014 .

Edad(años)	n	%
0-4	319	28.5
5-9	239	21.3
10-14	265	23.6
15-19	298	26.6
Total	1121	100.0



Cali, Colombia. Distribución de casos de cáncer infantil en el RPCC - Vigicáncer Según Clasificación Internacional de Cáncer Infantil (ICCC) . 2009-2014 .



Clasificacion de Cáncer infantil	n	%
I Leucemias	416	37.1
II Linfomas	182	16.2
III SNC	203	18.1
IV	22	2.0
V Retinoblastoma	35	3.1
VI Tumores renales	41	3.7
VII Tumores hepaticos	15	1.3
VIIII Tumores oseos	87	7.8
IX Sarcomas	54	4.8
X Tumores germinales	15	1.3
XI Carcinomas	48	4.3
XII Otros malignos	3	0.3
Total	1121	100.0

Figura 32. Resumen de variables para conjunto de datos 3

Las estadísticas descriptivas se muestran en la detallan el comportamiento de las variables. Seguidamente se exponen los resultados estadísticos obtenidos de cada una de las variables numéricas (Tabla 8).

Variables	No. de valores utilizados	Media	Desviacion Standard	No. de Valor Minimo	No. de Valor Máximo
Sexo	1121	144.603	.4973006	1	2
Edad	1121	9.490.633	5.830.657	1	19
Clasificacion del tumor	1121	3.478.145	3.066.192	1	12
Seguridad social	1121	3.175.736	2.915.228	1	9
Afrocolombiano	1121	4.222.123	3.388.754	1	9
Resiencia	1121	2.040.143	1.402.547	1	9
Estado vital	1121	1.651.204	.4768019	1	2

Tabla 8. Estadísticas descriptivas conjunto de datos 3

9.3.4 Evaluación inicial de los datos

En la selección de los datos se pudo observar problemas en la calidad de los mismos, y se encontraron campos que presentaban datos nulos y/o desconocidos. En general se han considerado los datos con valores y se han descartado solamente los valores nulos.

Los datos explorados brindan información sobre el comportamiento de cada uno de ellos pero existen variables que deben ser agrupadas y vistas en conjunto, de esta forma mostrar nuevo conocimiento sobre qué ocurre.

Algunos grupos que podemos identificar para el conjunto de datos 2 son:

- El número de casos que afecta alguna localización del tumor dependiendo del sexo y en cual estrato socioeconómico se presenta más casos de cáncer en Cali.
- El aumento o disminución de alguna localización del tumor según el tiempo o donde se ha presentado algún aumento.

Para el conjunto de datos 3 podemos agrupar:

- Cantidad de niños con cáncer con alguna localización del tumor, cuales son los más afectados según el aseguramiento.
- El número de casos de fallecidos según una localización del tumor y grupo de edad.

Los datos están bastante depurados, ya que se tienen en cuenta las variables que son consideradas como principales para la recolección de esta información. Aun así, existen datos que no tiene su respectiva codificación creando valores nulos.

9.4 Preparación de Datos

En la preparación de los datos se cubre las actividades para la construcción del conjunto de datos que serán utilizadas para el modelado. Esta preparación consta de:

- Eliminar valores con inconsistencias, valores nulos, etc.
- Seleccionar los datos con los que se trabajara
- Construcción de nuevos datos.

9.4.1 Selección y limpieza de datos

Para la selección y limpieza de la información se crearon consultas SQL teniendo en cuenta los datos que deberían tener un valor según la información que mostraría en las diferentes analíticas. Se aplicaron los siguientes ajustes a los datos:

1. En el Estrato socioeconómico solamente se aplicaría para los casos de residentes en Cali, porque para las otras residencias el valor es desconocido y no se tiene en cuenta para codificación respectivamente.
2. Las localizaciones de los tumores que son pertenecientes a cierto sexo solamente serán evaluadas para este como lo es: Próstata en hombres y Cuello uterino en Mujeres.

Con la herramienta WEKA al cargar los archivos de datos se realiza de forma automática la verificación de cada fila del archivo y que esta debe contener un valor para cada atributo, si los datos no se encuentran con valor este no la carga y por ende no se realiza e indica la fila que presenta problemas de inconsistencia de dato. De esta forma se realizó pre-procesamiento del archivo de datos y se descartaron las filas con valores nulos o valores inconsistentes.

9.4.2 Construcción de Nuevos Datos.

Se contó con suficientes datos por lo que no fue necesaria la construcción de nuevas estructura de datos.

9.4.3 Formateo de los Datos.

Los datos son inicialmente pre codificados de forma numérica para ser almacenados en la base de datos y contienen sus respectivos diccionarios indicando cuál es su valor real, de esta forma al crear las consultas para procesar los datos en la herramienta WEKA se tiene en cuenta los valores descriptivos.

9.5 Modelado

9.5.1 Modelo Conjunto de datos 2

Según las técnicas de minería de datos mencionadas en la sesión 6.2.1, se presenta el modelado de la información utilizando técnicas no supervisadas en donde se hacen agrupamientos (Clúster) para estas analíticas visuales.

Modelo de Clusterización:

Entradas:

Sexo (Valores Categóricos)
Sitio del tumor
Residencia (Valores Categóricos)
Estado Vital (Valores Categórico)
Estrato Socioeconómico E.S.E (Valor categórico definidas)

Salida:

Agrupamiento de registros basada en su similitud.

La exploración de estas variables serán analizadas en periodos quinquenales tal como es analizada y divulgada en el sitio Web del RPCC; los tres periodos publicados son: 1998-2002, 2003-2007 y 2004-2008, y se anexa al análisis un periodo que no se encuentra en la divulgación que es 2009-2013, de esta manera sería interesante ver los resultados para este último periodo porque por ahora se encuentra en recolección de información y podríamos identificar si podría presentar un comportamiento parecido a los tres quinquenios anteriores en cuanto a los sitios de los tumores principales. La información que analizada y divulgada por el RPCC son solamente los residentes de Cali porque el objetivo principal es captar la información de la zona urbana de Cali, igualmente el análisis presenta la información teniendo en cuenta otras residencias diferentes a Cali y se realizó el clúster para dos casos: **1**. Residentes general que incluye todas las residencias y **2**. Residentes de Cali, de esta forma se analiza la poca información que se pueda tener de otras residencias.

Usando WEKA para el primer quinquenio **1998-2002** y residentes en general (*Tabla 9*), generó un clúster usando el algoritmo K-MEANS[48]; para calcular la cantidad de clúster se aplicó el algoritmo EM identificando las similitudes de las variables obteniendo 14 clústeres y será realizado como

[48]Algoritmo de Clusterización K-MEANS: Es un algoritmo usado en minería de datos y descubrimiento de conocimiento no supervisado usado para agrupar registros dado el número de clúster K.

mínimo con 4, ya que los valores no presentan otros tumores diferentes en los conjuntos de datos de los clusteres presentados.

Tabla 9. Resultado Clúster quinquenio 1998-2002 Todas la residencias - WEKA.

Atributo	Cluster General	Cluster 1 45%	Cluster 2 30%	Cluster 3 16%	Cluster 4 9%
Sexo	Mujeres	Hombres	Mujeres	Mujeres	Hombres
Sitio del tumor	Mama	Próstata	Mama	Mama	Pulmon
Residencia	Cali	Cali	Cali	Valle	Cali

En el clúster General se muestra que los pacientes con Cáncer de Mama en mujeres es el principal tumor para el periodo 1998-2002, esto mismo se puede observar en la información divulgada por el RPCC en ese mismo periodo. De igual forma en los clúster 2 y 3, en donde son dos agrupamientos de datos diferentes siendo el 45% de los datos analizados y el 30% de los datos respectivamente, la similitud que se presenta en este conjunto de datos presenta que el sitio primario en este porcentaje de muestras se sigue presentando que el sitio primario de mama; los otros tumores que correspondiente a principales son próstata y pulmón, que son considerados tumores principales según las estadísticas del RPCC.

Al realizar este análisis solamente en los residentes de Cali, y teniendo como variable complementaria el E.S.E se obtiene una nueva localización del tumor en el clúster 5, siendo cuello uterino en mujeres en una muestra del 11% de la información analizada; y con respecto al estrato se presentan más para los estratos medio y bajo.

Atributo	Cluster General	Cluster 1 27%	Cluster 2 18%	Cluster 3 30%	Cluster 4 14%	Cluster 5 11%
Sexo	Mujeres	Hombres	Hombres	Mujeres	Mujeres	Mujeres
Sitio del tumor	Mama	Próstata	Próstata	Mama	Mama	Cuello Uterino
Estado vital	Muerto	Vivo	Muerto	Muerto	Vivo	Muerto
E.S.E	Medio	Bajo	Medio	Medio	Desconocido	Bajo

Tabla 10. Resultado Clúster quinquenio 1998-2002 residentes en Cali - WEKA.

Los resultados para el segundo quinquenio **2003-2007** con todas las residencias (Tabla 11) se obtiene que el clúster General muestra que los pacientes con Cáncer de Mama en mujeres es el principal tumor para este periodo, e igualmente se sigue presentando en los clúster 2,4 y 5. En el

cluster 1 encontramos el sitio del tumor en cuello uterino y en el cluster 6 se presenta un nuevo sitio primario que es estómago en esta agrupación de datos que no se presentó en el periodo 1998-2002 analizado anteriormente.

Tabla 11. Resultado Clúster quinquenio 2003-2007 todas la residencias - WEKA.

Atributo	Cluster General	Cluster 1 23%	Cluster 2 20%	Cluster 3 29%	Cluster 4 10%	Cluster 5 8%	Cluster 6 9%
Sexo	Mujeres	Mujeres	Mujeres	Hombres	Mujeres	Mujeres	Mujeres
Sitio del tumor	Mama	Cuello uterir	Mama	Próstata	Mama	Mama	Estómago
Residencia	Cali	Cali	Valle	Cali	Cali	Cali	Cali
Estado vital	Vivo	Vivo	Vivo	Muerto	Vivo	Vivo	Muerto
E.S.E	Desconocido	Desconocidc	Desconocidc	Desconocido	Medio	Bajo	Medio

Teniendo en cuenta lo mencionado de la residencia se realiza el análisis teniendo en cuenta sólo los residentes de Cali (Tabla *12*) y se obtienen los siguientes resultados:

Tabla 12. Resultado Clúster quinquenio 2003-2007 residentes de Cali - WEKA

Atributo	Cluster General	Cluster 1 43%	Cluster 2 19%	Cluster 3 10%	Cluster 4 25%	Cluster 5 3%
Sexo	Mujeres	Mujeres	Mujeres	Mujeres	Hombres	Mujeres
Sitio del tumor	Mama	Mama	Estómago	Cuello uterin	Próstata	Leucemia
Estado vital	Vivo	Vivo	Muerto	Muerto	Muerto	Muerto
E.S.E	Medio	Medio	Medio	Bajo	Bajo	Medio

En el clúster General se presenta que los pacientes con Cáncer de Mama en mujeres es el principal tumor para el periodo 2003-2007 de los residentes de Cali, en este caso los resultados son igual que el clúster aplicado a todas las residencias. En el cluster 5 se presenta el sitio del tumor Leucemias, siendo un sitio del tumor principal que es presentado en la divulgación de la información

Los resultados para el tercer quinquenio **2004-2008** (Tabla *13*) se obtiene que el clúster General muestra que los pacientes con Cáncer de Mama en mujeres es el principal tumor para este periodo, e igualmente se sigue presentando en los clúster 1.

En los clúster 2 y 5 los sitios primarios en hombres son en próstata y estómago; en el clúster 3 en las mujeres se presenta más frecuente en pulmón y en el cluster 4 en leucemias.

Tabla 13. Resultado Clúster quinquenio 2004-2008 todas los residentes de Cali - WEKA.

Atributo	Cluster General	Cluster 1 39%	Cluster 2 28%	Cluster 3 15%	Cluster 4 9%	Cluster 5 9%
Sexo	Mujeres	Mujeres	Hombres	Mujeres	Mujeres	Hombres
Sitio del tumor	Mama	Mama	Próstata	Pulmon	Leucemia	Estómago
Residencia	Cali	Cali	Cali	Cali	Valle	Cali
Estado vital	Vivo	Vivo	Vivo	Muerto	Muerto	Muerto
E.S.E	Desconocido	Desconocido	Desconocido	Medio	Desconocido	Bajo

En la Tabla *14* los resultados obtenidos para los residentes en Cali durante el periodo 2004-2008 muestra que los pacientes con Cáncer de Mama en mujeres es el principal tumor para este periodo y que el E.S.E se presenta en medio, e igualmente se sigue presentando en los clúster 5 con los mismo resultados.

En los clúster 1 y 2 para hombres se presenta que el sitio primario más frecuente es próstata y estómago y que E.S.E se presenta en estrato medio.

En el clúster 3 se presenta que para el sitio primario de cuello uterino el estrato socioeconómico más frecuente es bajo.

Tabla 14. Resultado Clúster quinquenio 2004-2008 residentes en Cali - WEKA.

Atributo	Cluster General	Cluster 1 18%	Cluster 2 29%	Cluster 3 25%	Cluster 4 9%	Cluster 5 19%
Sexo	Mujeres	Hombres	Hombres	Mujeres	Hombres	Mujeres
Sitio del tumor	Mama	Próstata	Estómago	Cuello uterino	Próstata	Mama
Estado vital	Vivo	Vivo	Muerto	Vivo	Vivo	Vivo
E.S.E	Medio	Medio	Medio	Bajo	Desconocido	Medio

Finalmente el último quinquenio 2009-20013 para analizar, es el periodo que se encuentra en proceso de recolección en el RPCC, pero se tiene datos procesados en el sistema de información utilizado (Tabla 15).

Los resultados indican que hasta el momento las localizaciones más frecuentes mencionadas en los anteriores periodos analizados siguen siendo mama y cuello uterino en mujeres, próstata en hombres y estómago.

TABLA 15. RESULTADO CLÚSTER QUINQUENIO 2009-2013 TODAS LAS RESIDENCIAS - WEKA.

Atributo	Cluster General	Cluster 1 31%	Cluster 2 29%	Cluster 3 8%	Cluster 4 23%	Cluster 5 10%
Sexo	Mujeres	Mujeres	Hombres	Mujeres	Mujeres	Mujeres
Sitio del tumor	Mama	Cuello uterinc	Próstata	Estómago	Mama	Estómago
Residencia	Cali	Valle	Cali	Otros	Cali	Cali
Estado vital	Vivo	Vivo	Vivo	Vivo	Vivo	Muerto
E.S.E	Desconocido	Desconocido	Desconocido	Desconocido	Desconocido	Medio

En conclusión con respecto a la definición de los principales tumores que se van a tener en cuenta para realizar las analíticas visuales después de realizar el análisis de minería de datos a los cuatro periodos comprendidos entre 1998 a 2013, se encontraron que las localizaciones de los tumores principales son: mama, próstata, pulmón, cuello uterino, estómago y leucemias.

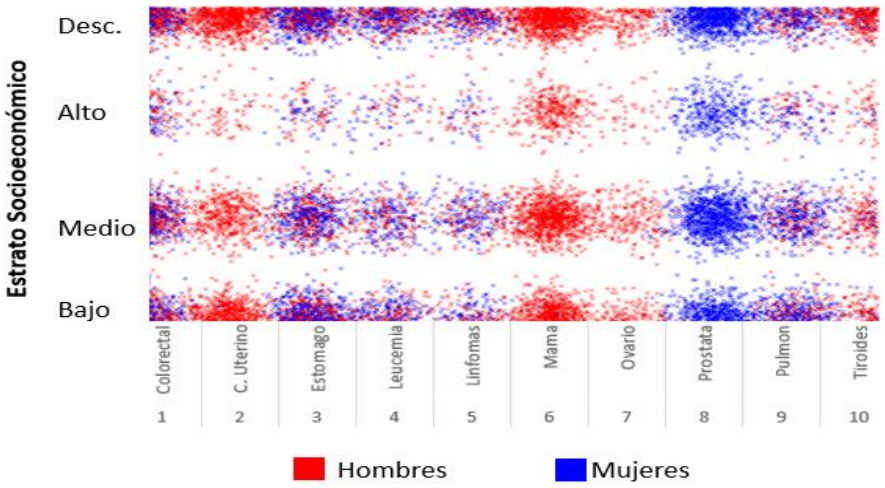
Existen 3 localizaciones adicionales que la minería de datos no logro encontrar en las agrupaciones de los diferentes cluster para los periodos, y son principales según la información presentada por el RPCC: colorectal, linfoma y tiroides; de igual forma estos sitios se tendrán en cuenta al presentar las analíticas.

Esta minería de datos evidencia la información divulgada por el RPCC, en donde se pretendía definir los sitios de los tumores con mayor frecuencia partiendo de una muestra de un conjunto de datos históricos de información de pacientes con cáncer registrados en el registro de cáncer durante los últimos de 15 años.

Después de definir los sitios de tumores principales y teniendo en cuenta dos variables adicionales como son: estado vital y estrato socioeconómico, se realiza un descubrimiento de los datos al lograr combinar un conjunto variables, de esta forma se tendrá como base la información para presentar en las analíticas visuales.

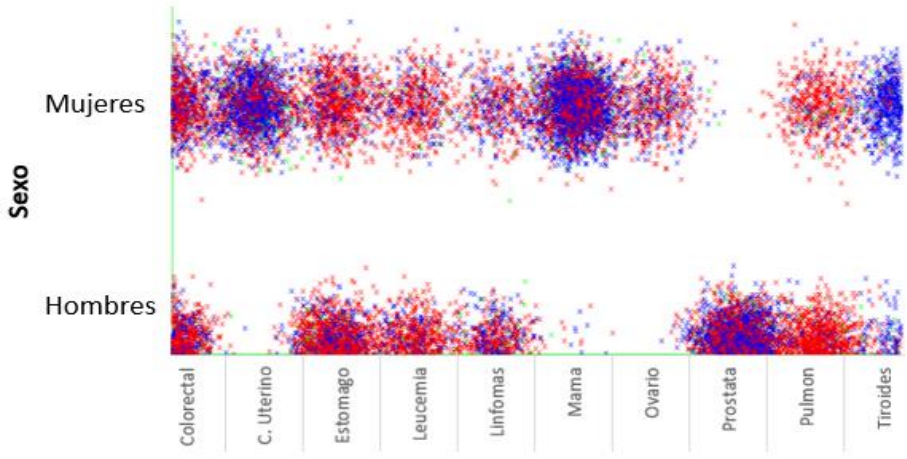
En la Figura 33 se observa que para el periodo 1998-2002 los estratos socioeconómicos más afectados por el cáncer son bajo y medio, y que las localizaciones que presentan la mayoría de los casos es para los mencionados en la tabla 10 que son: cuello uterino, estómago, mama, pulmón y próstata.

Figura 33. Gráfico de dispersión WEKA Clúster estrato socioeconómico/sitio del tumor. 1998-2002



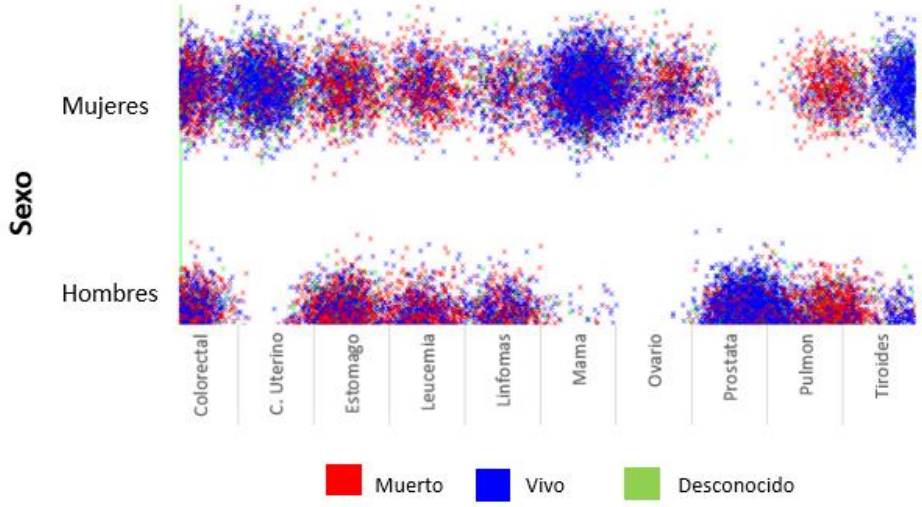
En la Figura 34 se puede especificar como es el comportamiento del estado vital según los sitios de tumor descritos y se puede observar que para estómago en hombres y mujeres, leucemia y pulmón en mujeres la cantidad de fallecidos es más alta en estos tumores; en cambio para cuello uterino, mama, próstata y tiroides tiene mejor supervivencia.

Figura 34. Gráfico de dispersión WEKA Clúster sexo/sitio del tumor. 2000-2004



En el gráfico de dispersión en la Figura 35 para el periodo 2003-2007 se observa que los sitios de tumor en donde existe mayor supervivencia son: cuello uterino, mama, próstata y tiroides; esto indica que su comportamiento es igual que el periodo 1998-2002; mientras que estómago en mujeres, pulmón y leucemias sigue siendo los sitios del tumor con mayor muerte.

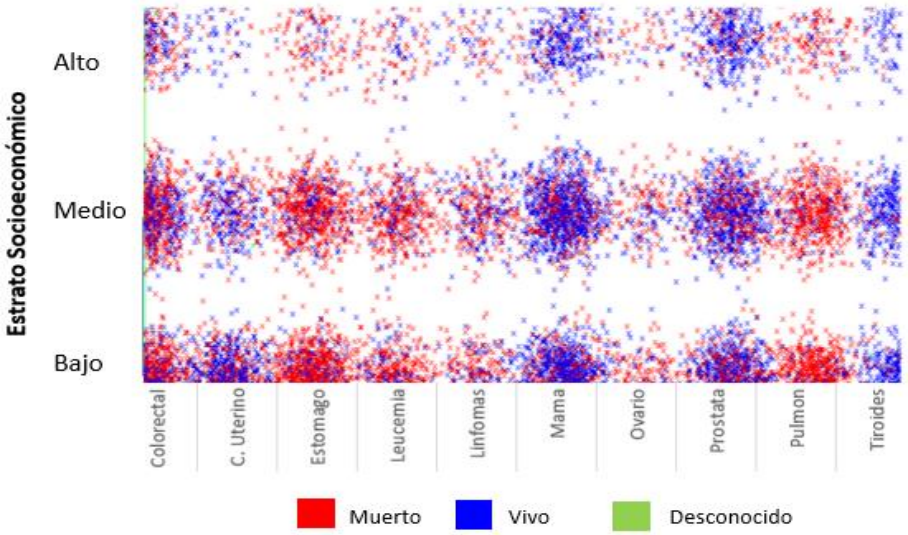
Figura 35. Figura 35. Gráfico de dispersión WEKA Clúster sexo/sitio del tumor según estado vital. 2003-2007.



En la Figura 36 se expone la distribución del estrato socioeconómico frente el sitio del tumor durante el periodo 2003-2007, en donde se observa que los tumores se presentan con mayor frecuencia en los E.S.E bajo y medio.

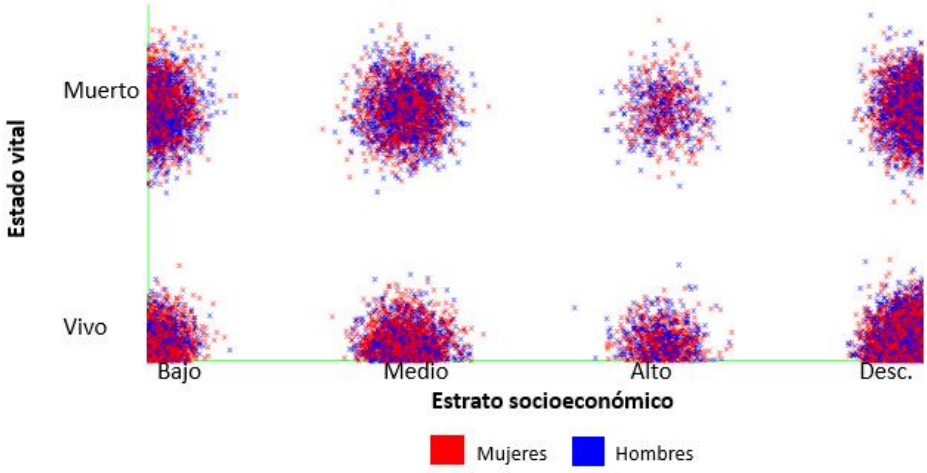
Para el E.S.E bajo y medio se muestra que los sitios de tumor con estado vital fallecido ocurren con más frecuencia en estómago y pulmón. Para el E.S.E alto se presenta que los sitios del tumor con más frecuencias son en mama y próstata.

Figura 36. Gráfico de dispersión WEKA Clúster E.S.E/sitio del tumor según estado vital. 2003-2007.



En la Figura 37 el gráfico de dispersión para el periodo 2004-2008 presenta el comportamiento del estado vital según el estrato socioeconómico, se puede observar que en los E.S.E bajo y medio la tendencia es que fallecen a comparación del E.S.E alto.

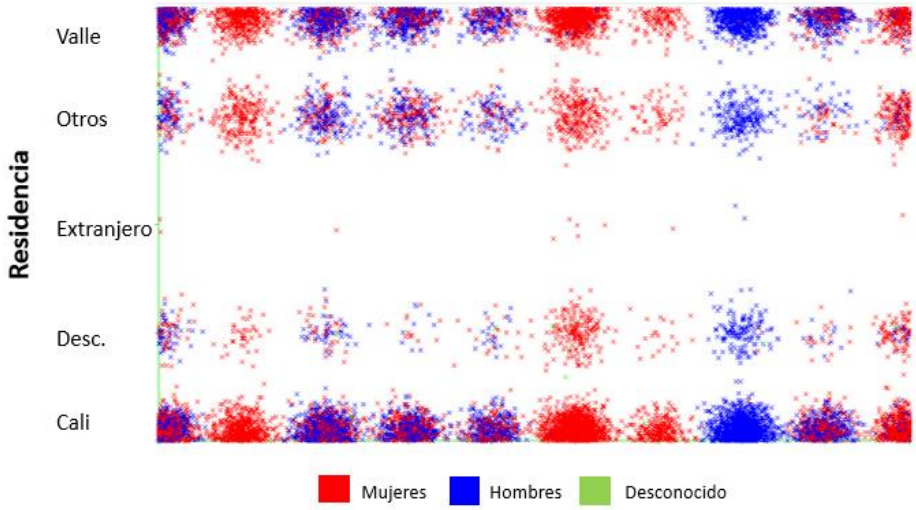
Figura 37. Gráfico de dispersión WEKA Clúster E.S.E/Estado vital según sexo. 2004-2008.



En el último periodo analizado 2009-2013, se puede mostrar información que representa lo que se encuentra procesado hasta el momento en el sistema de información del RPCC; en la Figura 38 se presenta que los datos que se han ingresado corresponden a su mayoría a los residentes de Cali y

Valle; y los sitios del tumor para hombres se presenta en los sitios primarios del tumor como son: próstata, estómago, leucemia, linfomas y pulmón para la ciudad de Cali.

Figura 38. Gráfico de dispersión WEKA Clúster residencia/Sitio del tumor según sexo. 2009-2013



La información presentada en los gráficos de dispersión de los datos analizados mediante el aplicativo WEKA, es la forma de tener un conocimiento previo respecto a la información que se mostrarán en las visualizaciones analíticas a implementar en el Aplicativo Web, siendo de un mayor conocimiento respecto al presentado en la actualidad d en el portal web del RPCC.

9.5.2 Modelado conjunto de datos 3

Modelo de Clusterización:

Entradas:

Sexo (Valores Categóricos)
Edad (Valores Categórico)
Clasificación internacional de Cáncer infantil – ICCC (Valores Categórico)
Residencia (Valores Categóricos)
Estado Vital (Valores Categórico)
Seguridad social (Valores Categórico)
Afrocolombiano (Valores Categórico)

Salida:

Agrupamiento de registros basada en su similitud.

AL CALCULAR LA CANTIDAD DE CLÚSTER SEGÚN EL ALGORITMO EM SE IDENTIFICAN 3 CLÚSTERES. LOS RESULTADOS QUE QUE MUESTRA ESTE ANÁLISIS (

Tabla 16) es el siguiente:

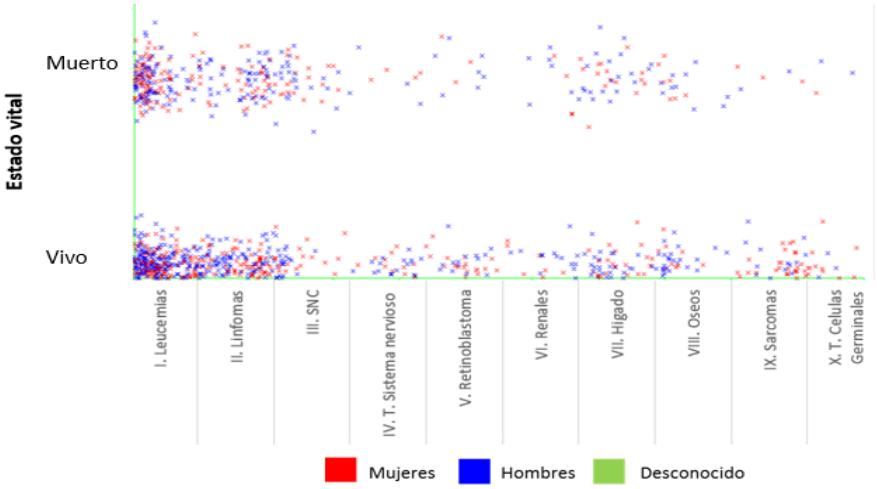
- En los clúster 1 y 3 muestra que el riesgo de cáncer es mayor en hombres con residencia en Cali.
- En los clúster 1 y 2 muestra el grupo de edad para estos tumores es en menores de 4 años de edad.
- En los clúster 2 y 3 se presenta que para los tumores del SNC y los linfomas su estado vital vivo. En el clúster 2 se muestra que para el tumor de SNC tienen residencia de Valle.

TABLA 16. RESULTADO CLÚSTER MÉDICO ESPECIALISTA 2009-2013- WEKA

Atributo	Clúster 1 44%	Clúster 2 31%	Clúster 3 25%
Seguridad Social	POS-s	POS	Desconocidc
Afrocolombiano	No	No	Desconocidc
Sexo	Hombres	Mujeres	Hombres
Edad	0-4	0-4	15-19
Residencia	Cali	Valle	Cali
ICCC	Linfoma	SNC	Linfomas
Estado vital	Muerto	Vivo	Vivo

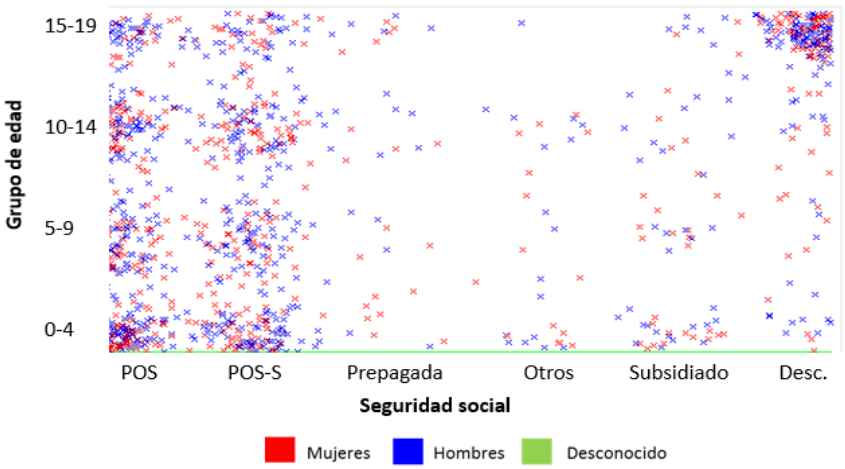
En la Figura 39 se muestra como es el comportamiento del estado vital según la clasificación internacional de cáncer infantil (ICCC), y se puede observar que se encuentran vivos principalmente los sitios de tumor como son: leucemia, linfomas y Neoplasia del sistema nervioso central.

Figura 39. Gráfico de dispersión WEKA Clúster estado vital/ICCC según sexo. 2009-2013



También se presenta la distribución de los casos de cáncer infantil por sexo según el grupo de edad y el aseguramiento de salud al que estén afiliados, se observa que principalmente el aseguramiento está entre POS y POS-s y en su mayoría están agrupados en los grupos de edad de 0-4 años de edad (Figura 40).

Figura 40. Gráfico de dispersión WEKA Clúster edad/seguridad social según sexo. 2009-2013



Los datos modelados brindan un mejor conocimiento de los datos del cáncer dependiendo de las variables que puedan correlacionarse en comparación con la exploración de estos mismos, como por ejemplo se pueden determinar las siguientes observaciones para estudios de investigación:

- Existen sitios del tumor que tiene una mejor supervivencia
- El estrato socioeconómico se comporta de forma diferentes según el sitio del tumor
- El aseguramiento de salud para cáncer infantil hace que exista una mejor supervivencia
- Los sitios de tumor son más frecuentes en un determinado sexo
- Que sitios del tumor son más frecuentes en otras residencias diferente de Cali

9.6 EVALUACIÓN

Los resultados obtenidos de los datos procesados muestran que existen localizaciones del tumor que se mantienen en el tiempo; según la información del conjunto de datos 2 se visualiza que durante los cuatro periodos comparados los sitios más comunes de Cáncer es igual que el presentado en el sitio Web del RPCC publicado, y que existen variables que pueden ser incluidas para la divulgación de la información.

En esta sección se presentan las localizaciones de los tumores por periodos de diagnóstico de pacientes residentes en Cali según los resultados de los clúster. De igual forma las variables de estado vital y estrato socioeconómico incorporadas para el descubrimiento de los datos presentan información valiosa para cada tumor. Teniendo en cuenta el estado vital, se presenta que algunos tumores son más letales que otros, por ejemplo para el cáncer de estómago es más probable que cause la muerte en comparación con los otros. Según el estrato socioeconómico existe una mayor cantidad de cáncer para los e.s.e medio y bajo para los diferentes tumores presentados (Figura 41):

Sitio del tumor: Próstata				
Atributo	Periodo			
	1998-2002	2003-2007	2004-2008	2009-2013
	Clúster 1 (27%)	Clúster 4 (25%)	Clúster 1 (18%)	Clúster 4 (19%)
Sexo	Hombres	Hombres	Hombres	Hombres
Estado vital	Vivo	Muerto	Vivo	Vivo
E.S.E	Bajo	Bajo	Medio	Desconocido
Sitio del tumor: Mama				
	Clúster 3 (30%)	Clúster 1 (43%)	Clúster 5 (9%)	Clúster 1 (40%)
Sexo	Mujeres	Mujeres	Mujeres	Mujeres
Estado vital	Muerto	Vivo	Vivo	Vivo
E.S.E	Medio	Medio	Medio	Medio
Sitio del tumor: Cuello uterino				
	Clúster 5 (11%)	Clúster 3 (10%)	Clúster 3 (25%)	
Sexo	Mujeres	Mujeres	Mujeres	-
Estado vital	Muerto	Muerto	Vivo	-
E.S.E	Bajo	Bajo	Bajo	-
Sitio del tumor: Estómago				
		Clúster 2 (19%)	Clúster 2 (29%)	Clúster 2 (23%)
Sexo	-	Mujeres	Hombres	Hombres
Estado vital	-	Muerto	Muerto	Muerto
E.S.E	-	Medio	Medio	Bajo

Figura 41. Comparación de los clúster del conjunto de datos 2

Esta evaluación permitió mostrar que la minería de datos realizada y los datos publicados en el portal Web del RPCC presenta las diez primeras causas de morbilidad por cáncer en Cali siendo la base para definir en las visualizaciones los sitios de los tumores a presentar en el aplicativo Web "VA_RPCC" a desarrollar.

10 Arquitectura de Software

En el Aplicativo Web VA_RPCC se ha utilizado un arquitectura de software por capas y se ha escogido actuador-indicador siendo una capa con cuatros capas funcionales y este modelo es utilizado en un prototipo de un ambiente virtual de aprendizaje mencionado en la sesión 6.1.

Esta arquitectura se implementa mediante el uso de bibliotecas JavaScript utilizando jQuery principalmente, y haciendo uso de librerías para representar las gráficas dinámicas como lo son: Jplot, D3 y Protovis, APIS de Google maps y Facebook. El esquema de la arquitectura se describe (Figura 41)

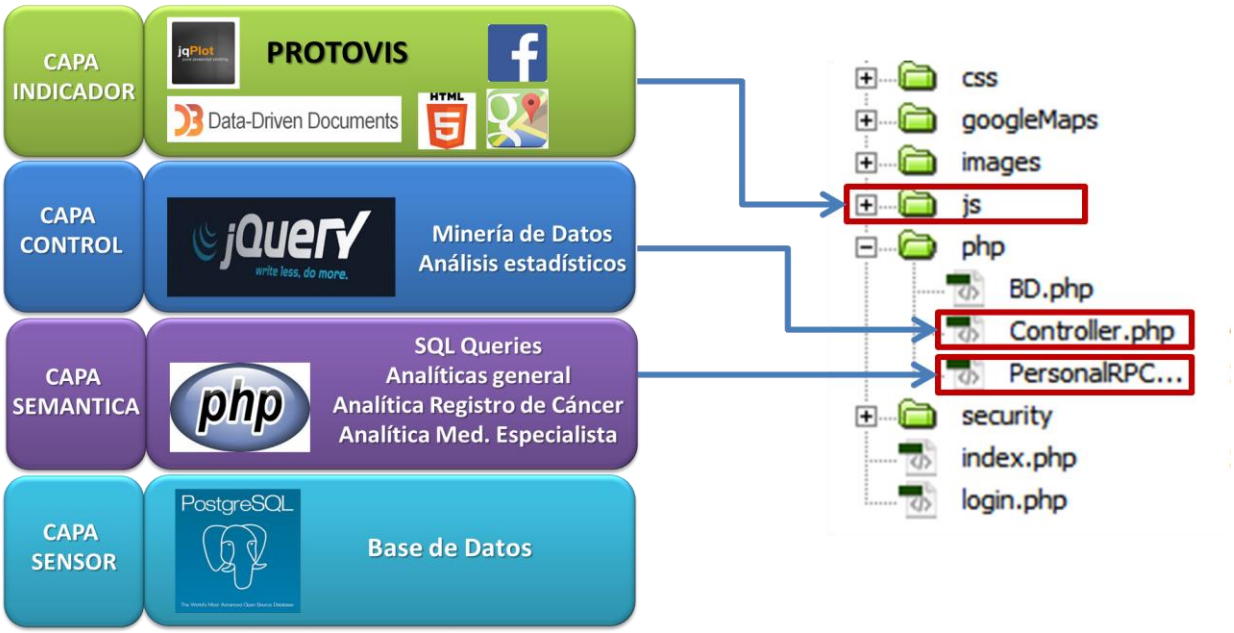


Figura 42. Arquitectura implementada VA_RPCC

Por este motivo se define una arquitectura la cual en su centro se encuentra dicho framework de desarrollo, esquematizando así el siguiente esquema o ciclo:

Se realizan peticiones por medio del cliente el cual interactúa directamente con la web(HTML) (***Capa Semántica***), dichas peticiones son interpretadas por javascript usualmente generando un objeto json las cuales son enviadas a través de peticiones ajax a un controlador escrito en php

(***Capa Control***), el cual a su vez accede a una capa inferior la cual realiza la conexión a la base de datos (***Capa sensor***) y recibe una respuesta se la retorna a la capa control y luego regresa a la petición ajax esperando su respuesta y cuya respuesta es un objeto con estructura json, una vez recibida esta respuesta y de acuerdo a la solicitud que se realizó y la necesidad, se procede a llamar a librerías las cuales nos permitan dibujar interfaces con html5, gráficos los cuales son implementaciones de javascript (***Capa indicador***) y uso los jQuery para lograr sus resultados finales.

A continuación se presenta cada capa implementada y su correspondencia a las visualizaciones implementadas para las tres analíticas presentadas, de esta forma se presenta un ejemplo de la Analítica Registro de Cáncer y la visualización en un gráfico de burbuja que se explica en la sesión 11.1.2 Analíticas del Registro de Cáncer; se presenta el código fuente para un mejor conocimiento de la arquitectura implementada.

10.1 Capa semántica

En la capa semántica se refiere a la capa que traduce la estructura física de los datos (con sus tablas, campos y relaciones) en la terminología de negocio.

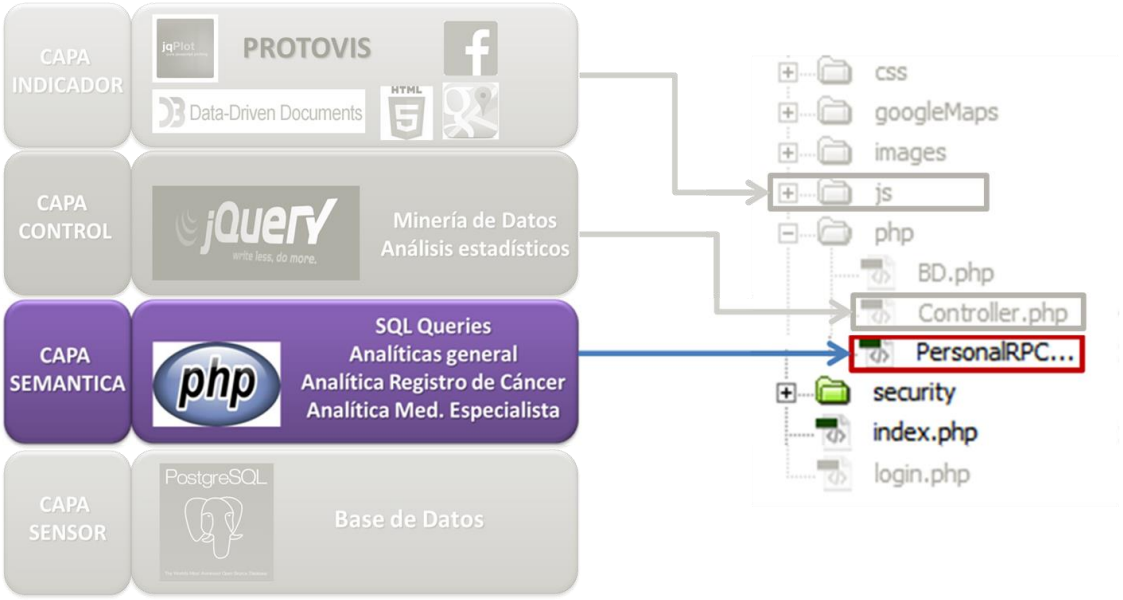


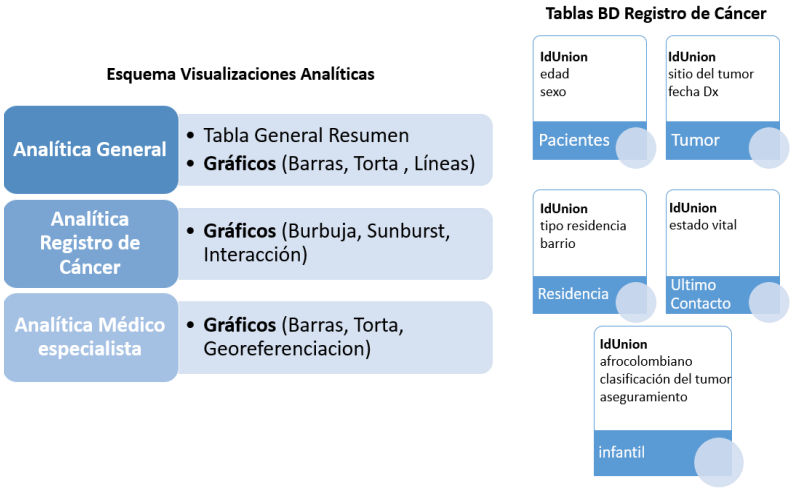
Figura 43. Capa Semántica- Arquitectura de Software VA_RPCC

Para la capa semántica se realizan los query correspondiente a las gráficas según cada analítica, en el caso del grafico de burbuja que corresponde a la visualización de la analítica registro de cáncer el query corresponde al siguiente:

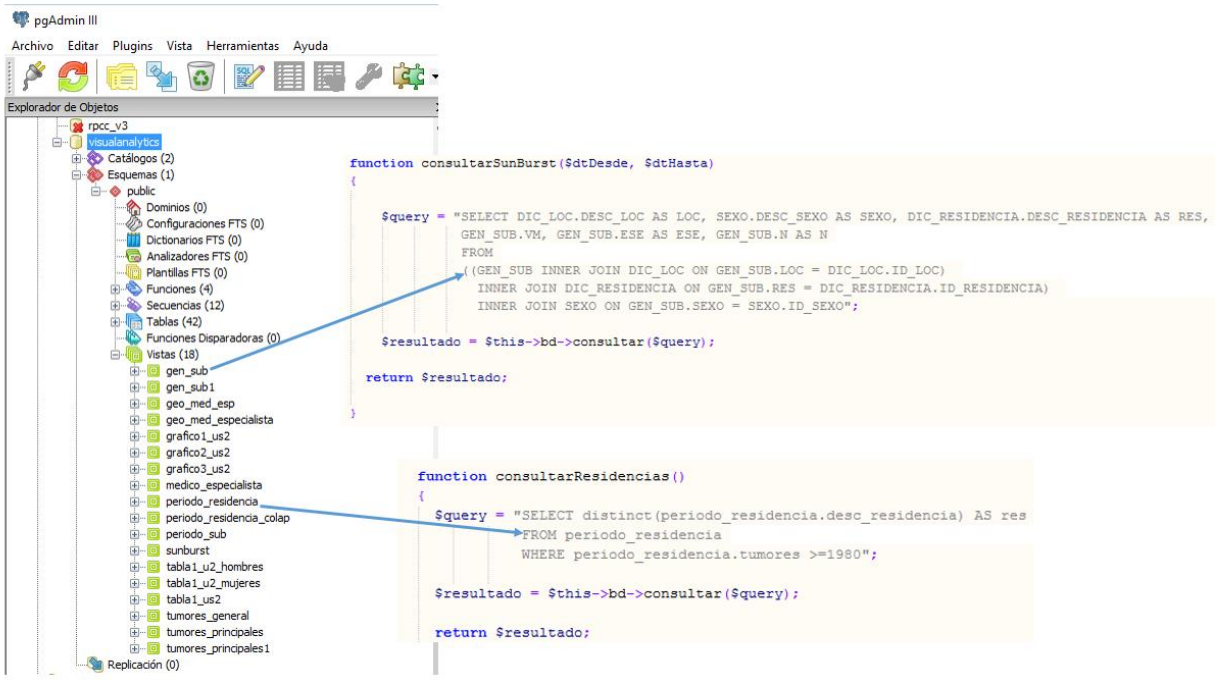
```
function consultarPeriodoRes()
{
  $query = "SELECT periodo_residencia.desc_residencia AS res, periodo_residencia.tumores AS periodo, Count(periodo_residencia.tumores) AS n
            FROM periodo_residencia
            WHERE
            periodo_residencia.tumores >=2000 and periodo_residencia.tumores<=2012 and desc_residencia ='Cali' or
            periodo_residencia.tumores >=2000 and periodo_residencia.tumores<=2012 and desc_residencia ='Valle' or
            periodo_residencia.tumores >=2000 and periodo_residencia.tumores<=2012 and desc_residencia ='Desconocido'
            GROUP BY periodo_residencia.desc_residencia, periodo_residencia.tumores;";

  $resultado = $this->bd->consultar($query);

  return $resultado;

}
```

Inicialmente se realizó un esquema de las interfaces gráficas que se querían implementar mediante los gráficos para cada uno de los controles de mando o analíticas visuales para cada usuario; de esta forma se tenían en cuenta que variables eran necesarias; estas variables fueron usadas en la minería de datos y las tablas donde se encuentra dicho información. Posteriormente se creó una base de datos que contenía la muestra de la información con la que se iba a trabajar teniendo en cuenta la confidencialidad de estos y no se usó ninguna información de identidad de la persona; con esto se fueron creando las consultas y vistas en postgres sql necesarias que se utilizarían en los gráficos.

De forma general se presentará el esquema de las diferentes analíticas y los gráficos que fueron seleccionados para presentar en cada uno de ellos y las tablas principales de la BD del Registro de Cáncer que contaba con dicha información; de igual forma se utilizaron tablas que contenían diccionarios, ya que la información del RPCC se encuentra precodificada.



Finalmente se crearon las siguientes vistas que agilizarán el proceso de carga de información en los gráficos, y se utilizan en la capa semántica.



10.2 Capa Control

La capa controlador se encarga de la abstracción de lógica relacionado con los datos, haciendo las peticiones en JavaScript usando las librerías de JQuery.

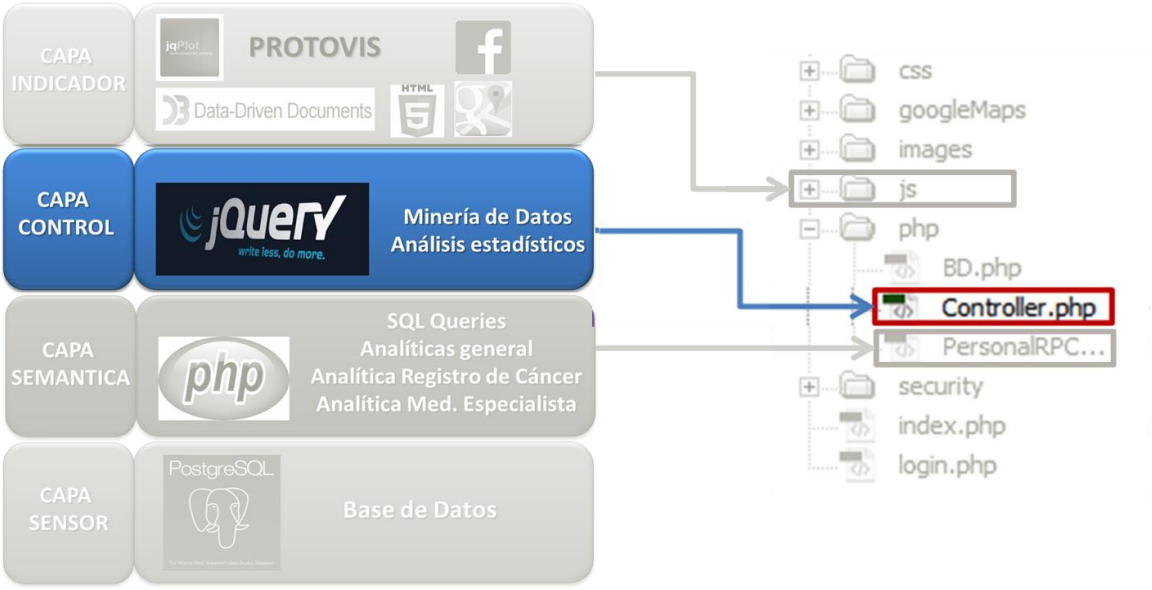


Figura 44. Capa Control- Arquitectura de Software VA_RPCC

Para el desarrollo e implementación de la arquitectura de software se utilizaron diferentes tecnologías Web, por un lado para el desarrollo funcional se utilizó JavaScript, junto a su librería JQuery, para la interfaz y el diseño se utilizó HTML.

HTML (Hyper Text Markup Language): Un estándar para la estructura y presentación de contenidos en la Web. El objetivo de utilizar esta tecnología en este desarrollo es permitir dibujar libremente sobre un espacio en la página sin necesidad de usar complementos y tecnologías propietarias como Adobe Flash.

JavaScript: Es un lenguaje interpretado, utilizado para acceder a objetos en aplicaciones. Se utiliza integrando el navegador web permitiendo el desarrollo de interfaces de usuario y páginas Web dinámicas. Tradicionalmente se ha utilizado en páginas web HTML para realizar operaciones y en el marco de la aplicación cliente, sin acceso a funciones del servidor. JavaScript se ejecuta en el cliente (Navegador Web), al mismo tiempo que las sentencias van descargando junto con el código HTML.

Javascript se utilizó en el proyecto por ser un código "interpretado" por el cliente, un lenguaje abierto, un código integrado a HTML, además se utilizan en el proyecto dos librerías que están escritas en JavaScript: JQuery y Jplot.

JQuery: Es una librería de JavaScript liviana e interoperable entre navegadores web diseñada para simplificar el scripting de HTML del lado del cliente. Es la librería de JavaScript más popular de las usadas hoy en día.

JQuery se utiliza en el proyecto porque implementa una serie de clases (de programación orientada a objetos) que permite programar sin preocupación en la compatibilidad con los navegadores, ya que funciona de exacta forma en todas las plataformas más habituales.

Siguiendo con el ejemplo de la visualización de analíticas de Registro Cáncer para el grafico de burbuja se presenta el código fuente implementado en ***controller.php,*** en donde dicha capa responde a los eventos que son las solicitudes del usuario cuando selecciona un una analítica que información desea visualizar en el gráfico según la selección que realice haciendo que este sea el intermediario entre la capa de semántica y la capa indicador logrando el funcionamiento del desarrollo implementado:

```
case 'getBubbleData':
    $resRes = $obj_PersonalRPCC->consultarResidencias();
    $resPerRes = $obj_PersonalRPCC->consultarPeriodoRes();
    $arrayLoc = array();
    $arrayPerRes = array();
    $arrayData = array();
    if($resRes){
        for($i=0;$i<pg_NumRows($resRes);$i++){
            array_push($arrayLoc,pg_fetch_array($resRes, NULL, PGSQL_NUM));
        }
    }
    if($resPerRes){
        for($i=0;$i<pg_NumRows($resPerRes);$i++){
            array_push($arrayPerRes,pg_fetch_array($resPerRes, NULL, PGSQL_NUM));
        }
    }
    for($j=0;$j<count($arrayLoc);$j++){
        $arrayClidren = array();
        for($i=0;$i<count($arrayPerRes);$i++){
            if($arrayLoc[$j][0] === $arrayPerRes[$i][0]){
                $children = array('name' => $arrayPerRes[$i][0].":".$arrayPerRes[$i][1], 'size' => $arrayPerRes[$i][2]*1);
                array_push($arrayClidren,$children);
            }
        }
        for($i=0;$i<count($arrayClidren);$i++){
            $arrayClidren[$i]["size"] = $arrayClidren[$i]["size"]*1;
        }
        $array = array('name' => $arrayLoc[$j][0], 'children' => $arrayClidren);
        array_push($arrayData,$array);
    }
    $array = array('name' => 'main', 'children' => $arrayData);
    $arrayResponse = array(
        'RESPUESTA' => $RESPUESTA,
        'DATA' => $array,
        'TITLE' => "Grafico de Burbuja",
        'MSG' => $MSG
    );
break;
```

En este ejemplo se hace un llamado para presentar la información en un gráfico de burbuja, utilizando como parametros los datos de residencia y los diferentes periodos de diagnósticos, en la sección 11.2 se explica con más detalle el resultante de dicha visualización .

10.3 Capa indicador

Finalmente la capa **indicador** se encarga de transformar los datos devueltos por la capa control de tal forma que sea interpretada por los usuarios, de esta forma se visualizan las diferentes analíticas y se presentan los gráficos para las analíticas correspondiente.

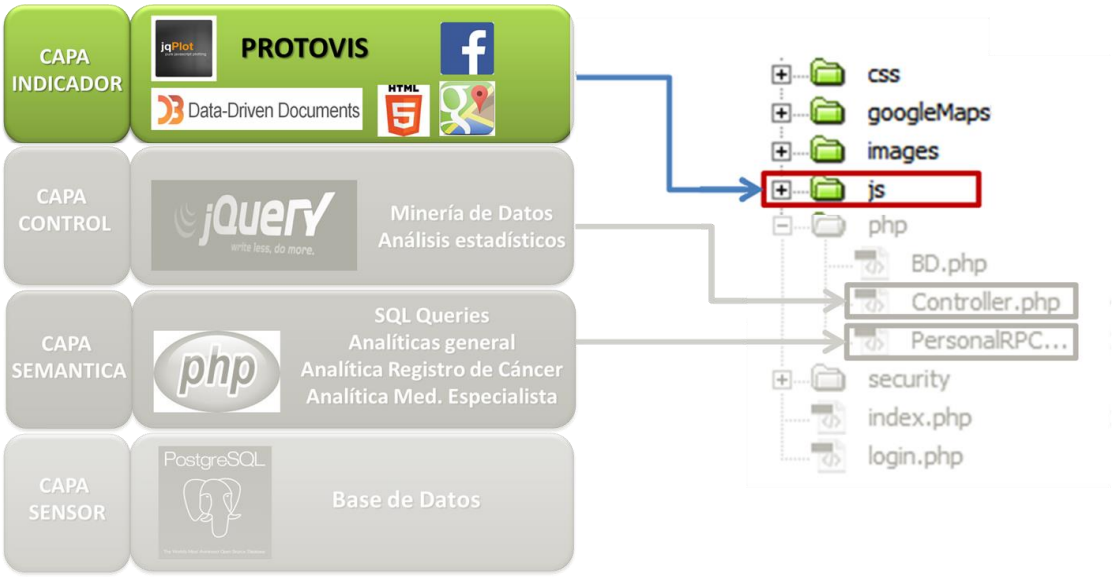


Figura 45.Capa Indicador- Arquitectura de Software VA_RPCC

En el siguiente código fuente se presenta el proceso de transformación de los datos enviado por la capa controlador siguiendo con el ejemplo del grafico de burbuja explicado en las diferentes capas:

Ejemplo - Analítica Registro de Cáncer - Gráfico de burbuja.

Mostrando el ejemplo relacionado se presenta el código fuente en *jquery.jqplot.custom* donde se realiza el llamado de la presentación de los datos al usuario para la gráfica de burbuja.

```
function loadPerRPCC(){

    $("#divBubbleLoad").html("");
    $("#divBubble").html("");
    $("#divBubbleLoad").empty();
    $("#divBubble").empty();
    $("#divBubbleLoad").html("<center><table><tr><td><img src='./css/012.gif'></td></tr></table></center>");

    loadBubbleData();
```

```
function loadBubbleData(){
    var arrayData = "";
    arrayData += "&accion=getBubbleData";
    $.ajax({
        url: './php/Controller.php',
        type: 'POST',
        data: arrayData,
        dataType: 'json',
        success: function(data) {
            if(data.RESPUESTA === true){
                diameter = 900,
                format = d3.format(",d"),
                color = d3.scale.category20c();
                bubble = d3.layout.pack()
                .sort(null)
                .size([diameter, diameter])
                .padding(1.5);
                svg = d3.select("#divBubble").append("svg")
                .attr("width", diameter)
                .attr("height", diameter)
                .attr("class", "bubble");
                loadBubbleChart(data.DATA);
                d3.select(self.frameElement).style("height", diameter + "px");

                $("#divBubbleLoad").html("");
                $("#divBubbleLoad").empty("");

            }else{
                jQueryMessage("Error","<H3>Ocurrio un error </H3>","Error");
            }
```

A continuación se presentan tres librerías (Jplot, D3 y Protovis) utilizadas en el desarrollo de este aplicativo para la generación de gráficos estadísticos; de igual forma se muestran dos API (Google maps y Facebook) implementadas en el desarrollo propuesto.

10.3.1 JPlot

Jplot está escrito en JQuery y JavaScript, utiliza el elemento *Canvas* para renderizar del lado del cliente gráficos dinámicos por medio de programación. Jplot se destaca sobre todo por sus posibilidades y por ser el que da las opciones más avanzadas para generar gráficas y de modificación dinámica.

En la Figura 46 se muestra las gráficas de barra, líneas y pie que han sido implementadas con Jplot, estas se visualizan en el esquema representando un conjunto de datos provenientes del archivo Json.

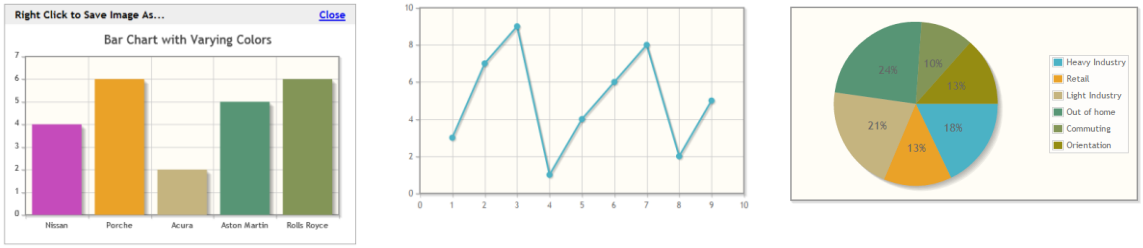


Figura 46. Visualización de Gráficos dinámicos librería JPLOT

10.3.2 D3 (Data-Driven Documents)

D3.js es una librería hecha en JAVASCRIPT que permite enlazar datos al DOM de la página y formar gráficos, tablas o listas con ellos. D3[49] manipula eficientemente los documentos sobre una base de datos establecidos, integrándose completamente con Css3, HTML y SVG disminuyendo coste de tiempo en representaciones propias. Es muy rápido y puede trabajar con grandes cantidades de datos, además de permitir reutilización de código a través de módulos.

D3 reduce el código empleado para la visualización de datos con respecto a otros como DOM API, además de ser compatible con todos los navegadores actuales. Esta librería es la versión mejorada de otra librería utilizada para gráficos llamada Protovis, que permitía hacer visualizaciones de marcos estáticos y por eso su desarrollador pensó en armar toda esta librería y hacerla dinámica.

D3 cuenta con un buen número de tipos de gráficos empezando por tablas, listas, gráficos tipo diagrama, barras apiladas, diagrama de dispersión, pirámide poblacional, mapas, calendario, Árbol de nodos, gráficos de burbujas, círculo de embalaje y diagramas de caja, en la se presenta el grafico de burbujas (Figura 47) que ha sido implementado en este desarrollo y se presenta en las analíticas del Registro de Cáncer en donde se muestra la distribución de las residencias en los diferentes periodos y en la Figura 48 en el gráfico de distribución de partición (Sunburst) en donde

[49]Recopilación información de:mbostock.github.com/d3/

se pueda navegar por las diferentes categorías y se van mostrando los resultados más específicamente.

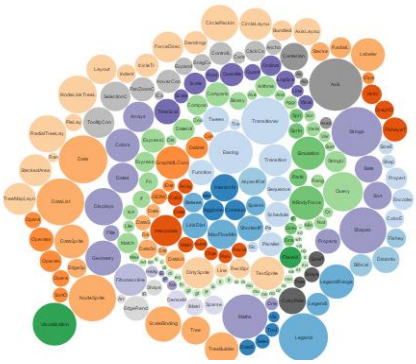


Figura 47. Gráfica de Burbuja librería D3 para las analíticas del RPCC

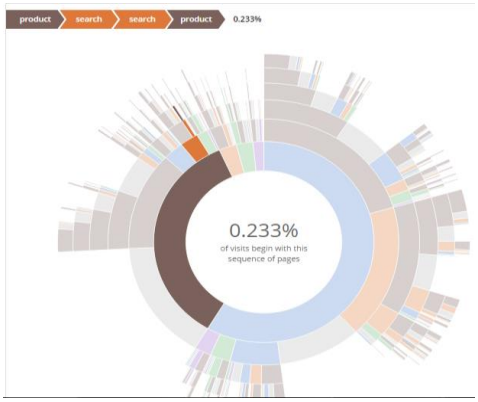


Figura 48. Gráfico sunburst librería D3 para las analíticas del RPCC

10.3.3 Protovis

Protovis[50] es un lenguaje de código abierto para desarrollar visualizaciones. Se distribuye bajo licencia BSD, internamente utiliza *javascript y* SVG para visualizaciones para la web. No es necesario tener un *plugin* en la maquina cliente para correr las visualizaciones desde una página web y es muy sencillo de aprender. Protovis permite, a través de una serie de objetos de datos, construir cualquier gráfico con nodos o barras, de una forma rápida y ligera. Además, esta librería es open-source, por lo que no requiere gasto alguno para el proyecto.

En la Figura 49 se muestra una gráfica de interacción (Serie de tiempo) que ha sido implementada en las analíticas del RPCC y se presenta el comportamiento de la información en los diferentes períodos evaluados.

[50] http://mbostock.github.io/protovis/

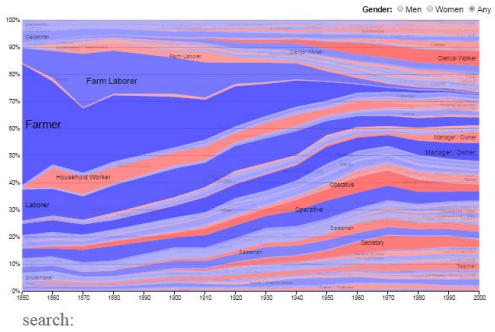


Figura 49. Gráfico de series de datos implementado con Protovis para las analíticas del RPCC

10.3.4 Visualización de datos con Google Maps

Uno más de los productos llevados al mercado por el gigante Google Inc., Se trata de un servidor de aplicaciones de mapas en la Web (Internet), con la capacidad de hacer acercamientos o alejamientos (Zoom) al mapa, controlando con el mouse o las teclas de dirección los movimientos para encontrar la ubicación que se desee; además los usuarios pueden ingresar una dirección, una intersección o un área en general para buscar en el mapa y encontrar los resultados.

Google Maps puso a disposición de los desarrolladores sus códigos fuentes llamados APIS, los mismos que permiten introducir los mapas de Google Maps en cualquier aplicación con el uso de su codificación y con ello se pueden aplicar nuevas formas de ver el mundo.

En la Figura 50 se muestra el API usado en el desarrollo de las Analíticas Visuales del Registro de Cáncer de Cali el cual se presentan en las analíticas del Médico especialista.

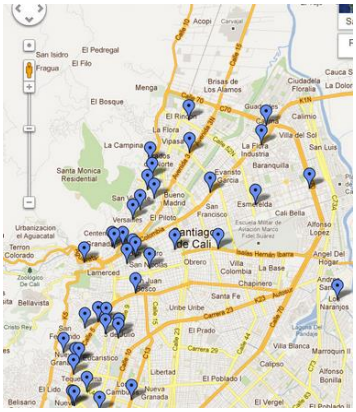


Figura 50. API Google Maps implementado en VA_RPCC

1.1.1 Visualización de datos con Facebook

Facebook es la red social por antonomasia. Creada a principios de 2004 por el estudiante de la Universidad de Harvard, Mark Zuckerberg, esta red social se ha convertido en una de las webs más importantes en todo el mundo, debido a su gran volumen de visitas al día –según el sitio web alexa.com, que muestra información acerca de las visitas de los usuarios en los diferentes sitios de Internet, Facebook es la segunda web con mayor número de visitas al día en todo el mundo, sólo por detrás de Google (Alexa, 2012)- y su gran número de usuarios registrados, siendo líder en este aspecto entre todas las redes sociales del mundo.

Esta red social es la que más facilidades da a la hora de crear aplicaciones. Tiene un apartado dentro de la web dedicada al desarrollo de aplicaciones (Facebook, 2012), en la que se puede obtener ayuda para insertar elementos de Facebook en una página web o aplicaciones de dispositivos móviles, o para diseñar y construir aplicaciones para Facebook. De hecho, tiene una API con la que se puede obtener cualquier dato necesario de una persona para utilizarla en la aplicación a crear.

Los plugins[51] sociales de Facebook son herramientas que te ayudan a integrar un sitio web o blog con Facebook, ofrecen funcionales para compartir contenido e interactuar con tu comunidad sin que los lectores dejen el sitio web.

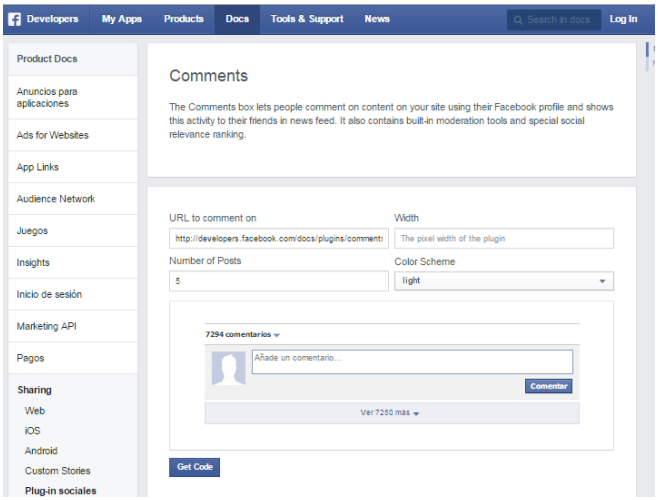


Figura 51. API Facebook usado en VA_RPCC

10.4 **Capa sensor**

La implementación de este trabajo de grado se desarrolló utilizando este sistema de gestión de bases de datos Postgres, porque actualmente es utilizado en el RPCC para el almacenamiento de la información que se alimenta mediante el Sistema de información SISCAN, de igual forma se tomó una muestra de lo datos en donde se mantuvo la confidencialidad de la información utilizando en la muestra solo la información de las variables para los análisis.

[51] http://developers.facebook.com/docs/plugins/

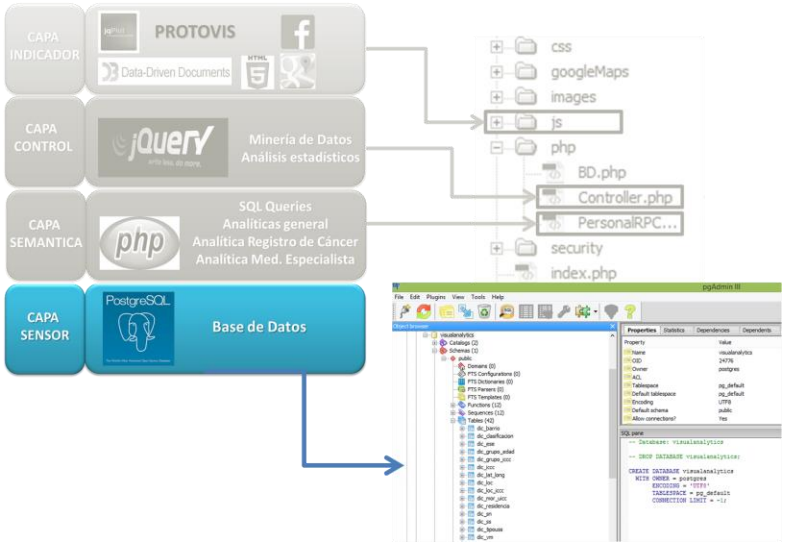


Figura 52. Capa Sensor - Arquitectura de Software VA_RPCC

PostgresQL : Una Base de Datos (BD) es un modelo que representa algún aspecto del mundo real. Es un conjunto de datos coherentes, con cierto significado inherente. Por ser una abstracción del mundo real, toda base de datos se diseña, construye y llena con datos para un propósito específico y está dirigida a un grupo de usuarios o aplicaciones. La base de datos son sistemas que han evolucionado mucho y actualmente ellas son creada con un sistema manejador de base de datos.

El modelo representado en una BD generalmente es un modelo de datos llamado entidad-relación (E-R), el cual está basado en una percepción del mundo real que consta de un conjunto de objetos básicos llamados entidades y de relaciones entre estos objetos. El modelo de datos E-R es uno de los diferentes modelos de datos semánticos; el aspecto semántico del modelo yace en el intento de representar el significado de los datos. El modelo E-R es útil para hacer corresponder los significados e interacciones de los desarrollos del mundo real con un esquema conceptual.

Principalmente se configura la conexión con la base de datos en ***BD.php***:

```
function BD()
{
    $this->servidor="localhost";
    $this->usuario="postgres";
    $this->password="postgres";
    $this->db="visualanalytics";
    $this->puerto="5432";
}

function conexion()
{
// se definen los datos del servidor de base de datos
$this->conexion = pg_connect("host=".$this->servidor."
                             port=".$this->puerto."
                             dbname=".$this->db."
                             user=".$this->usuario."
                             password=".$this->password);
```

11 Aplicativo Web de visualización de analíticas VA_RPCC

VA_RPCC es una aplicación construida en ambiente web, en el cual se visualizan tres controles de mando de analíticas visuales (Analíticas generales, Analíticas Registro de Cáncer y Analíticas Médico especialista); dependiente el tipo de usuario se incluye tablas, gráficas dinámicas y API's de georeferenciación y para todos los usuarios existe el API de redes sociales el cual permite hacer comentarios al respecto de la información presentada; para estas analíticas se presentan las frecuencias o porcentajes de los principales tumores según la información que tiene el Registro de Cáncer de Cali.

El aplicativo tiene un inicio de sesión para los diferentes tipos de usuarios, aunque para el usuario en general se presenta la información requerida para su ingreso y de esta forma visualizar dicha información. Para los otros dos tipos de usuario se utilizan datos de inicio de sesión que deben ser solicitados.

A continuación se presenta cada analítica con las respectivas gráficas propuestas haciendo una breve explicación de cada una.

1.1.1. Control de mando analíticas generales

En este tipo de analítica se presenta para un usuario general (Figura 53); la información de los datos con cáncer se muestra mediante frecuencias o datos porcentuales que se representan en diferentes gráficos.

Inicialmente se presenta la información en una tabla mostrando las frecuencias de los sitios primarios del tumor dependiendo del sexo seleccionado; posteriormente la información de la tabla de frecuencias se puede visualizar en tres tipos de gráficas: barras, torta y línea las diferentes localizaciones de los tumores principales durante el tiempo (Figura 53).

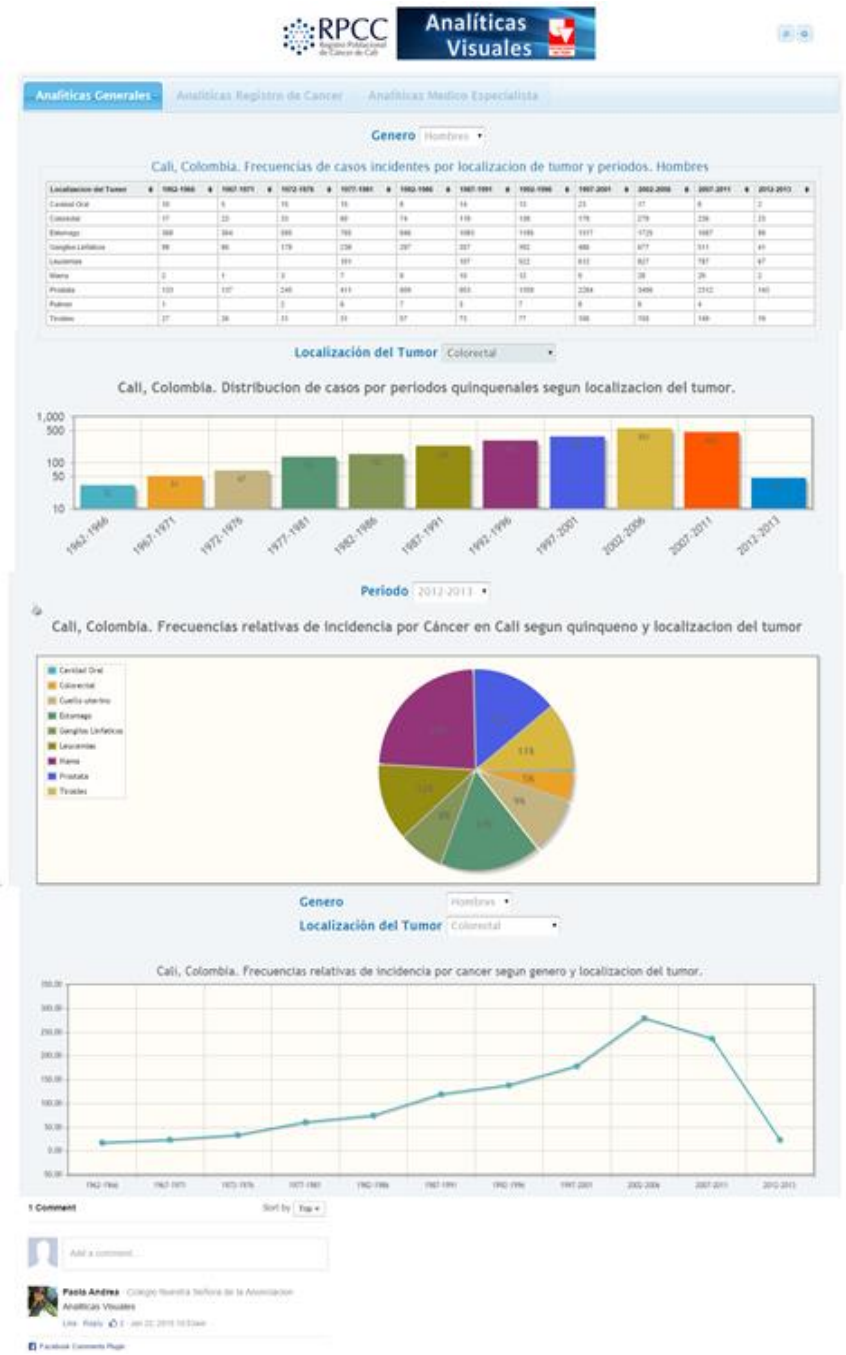


Figura 53. Control de mando analítica general aplicativa Web VA_RPCC

En la gráfica de barras se presenta el comportamiento dependiendo de la localización del tumor seleccionada, esta gráfica muestra las frecuencias en los diferentes quinquenios para ambos sexos, mediante esta gráfica se puede comparar dos o más períodos según la altura de las barras representa la magnitud de los valores, de esa forma es más fácil visualizar dichos datos porque el tamaño de la barra representa la cantidad de los registros que se existen en ese periodo.

En la gráfica de torta se presenta el comportamiento de las diferentes localizaciones de los tumores según un periodo quinquenal seleccionado, de esta forma se representa la proporción de cada uno de los valores de la variable y se logra visualizar un tamaño para cada tumor.

En la gráfica de líneas se muestra la tendencia según la localización del tumor y el sexo seleccionado mediante un conjunto de puntos conectados por una línea que representa los datos continuos en el tiempo. De esta forma se visualiza si ocurre de determinado tumor de un periodo a otro.

1.1.2 Control de mando analíticas Registro de Cáncer

En este tipo de analítica se presentan gráficas un poco más avanzadas que presentan el comportamiento de los datos a través del tiempo que se han ingresados en el sistema información del Registro de Cáncer, actualmente es SISCAN. En la Figura 53 se muestran tres tipos de gráficas: burbuja, distribución de particiones y series de tiempo implementadas en su orden correspondiente.

En la gráfica de burbujas se presenta la dispersión de puntos según el periodo de diagnósticos de los datos por las diferentes residencias de los casos registrados en la base de datos (Figura 54). En este tipo de gráfica el tamaño de la burbuja indica la dimensión de los datos, es decir que entre más grande sea la burbuja significa que contiene más información para ese periodo y la residencia. En este gráfico se pretende dar una vista general de la cantidad de casos registrados, de esta forma se puede definir como se encuentra la recolección de la información y poder realizar un seguimiento de ello.

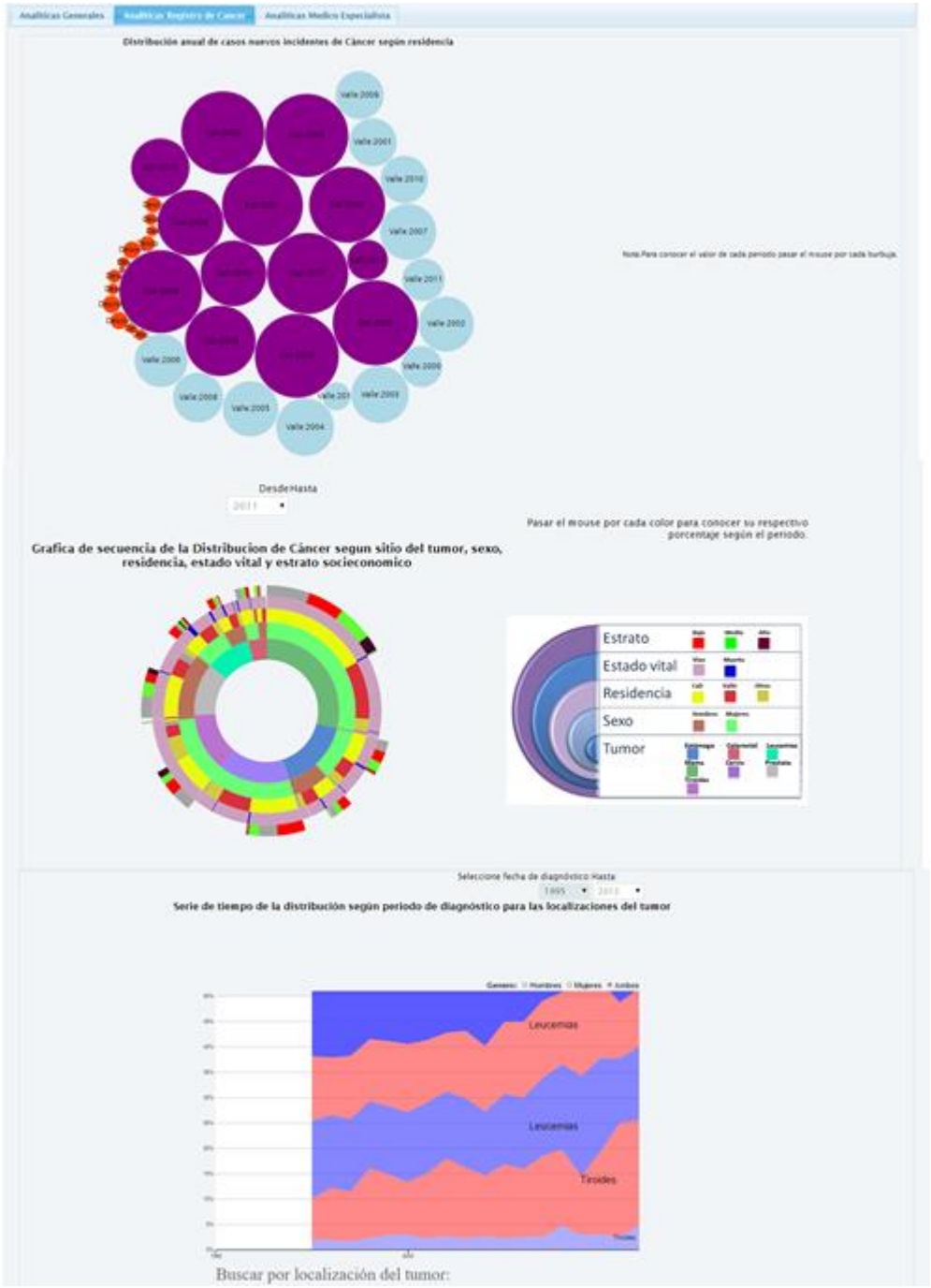


Figura 54. Control de mando analíticas Registro de Cáncer aplicativo Web VA_RPCC

Seguidamente se presenta el grafico de distribución de partición (Sunburst), en este caso los datos se representan por categorías siendo una gráfica de introspección en el cual se parte del mayor detalle hasta el menor (Figura 55). La información mostrada se divide en 5 aros con el siguiente orden: localización del tumor, sexo, residencia, estado vital y estrato socioeconómico, mostrando de forma porcentual los valores cada vez que se va pasando el clic sobre cada aro. Este gráfico sería un resumen en gran parte de las variables principales que se manejan en el RPCC y ayuda a identificar el comportamiento de los datos, siendo la forma más óptima para el análisis de múltiples variables, y logrando identificar cómo se encuentra la actualización de dichos datos para un periodo seleccionado.

Finalmente en la parte inferior se muestra una gráfico de serie de tiempo o gráfico cronológico, en el cual se visualizan las localizaciones de los tumores que ocurren en ambos sexos mediante un área de color azul (hombres) y rosado (mujeres); y según el tamaño este representa si existe mayor o menor cantidad de casos para cada sexo. Los datos visualizados en este tipo de gráfico determinan las tendencias durante el periodo de diagnóstico seleccionado.

Este control de mando o visualización está dirigido al personal del registro y de esta forma conocer cómo se encuentra en tiempo real las variables principales que son utilizadas para presentar y divulgar las estadísticas de cáncer en la ciudad de Cali.

1.1.3 Control de mando analíticas médico especialista

En este tipo de analítica, se ha tenido en cuenta un proyecto de investigación sobre cáncer infantil (Sistema de vigilancia epidemiológica de Cáncer pediátrico para la ciudad de Cali)[52]. Este sistema parte de la información del RPCC, utilizando su sistema de información SISCAN, pero adicionalmente cuenta con variables complementarias que el Registro de cáncer no capta para cáncer en adultos y es vigilada por monitores clínicos y los hemato-oncólogos que participan en dicho proyecto y son encargados de captar o definir dicha información, con un inicio de proyecto desde el año 2009.

[52] http://rpcc.univalle.edu.co/uicc/

En estas analíticas se tiene en cuenta algunas de las variables complementarias, presentando en dos tipos de gráficas (torta y barras), y aparte se implementa la georreferenciación utilizando Google Maps para identificar la ubicación geográfica en el mapa de Cali para los residentes de este (Figura 55).

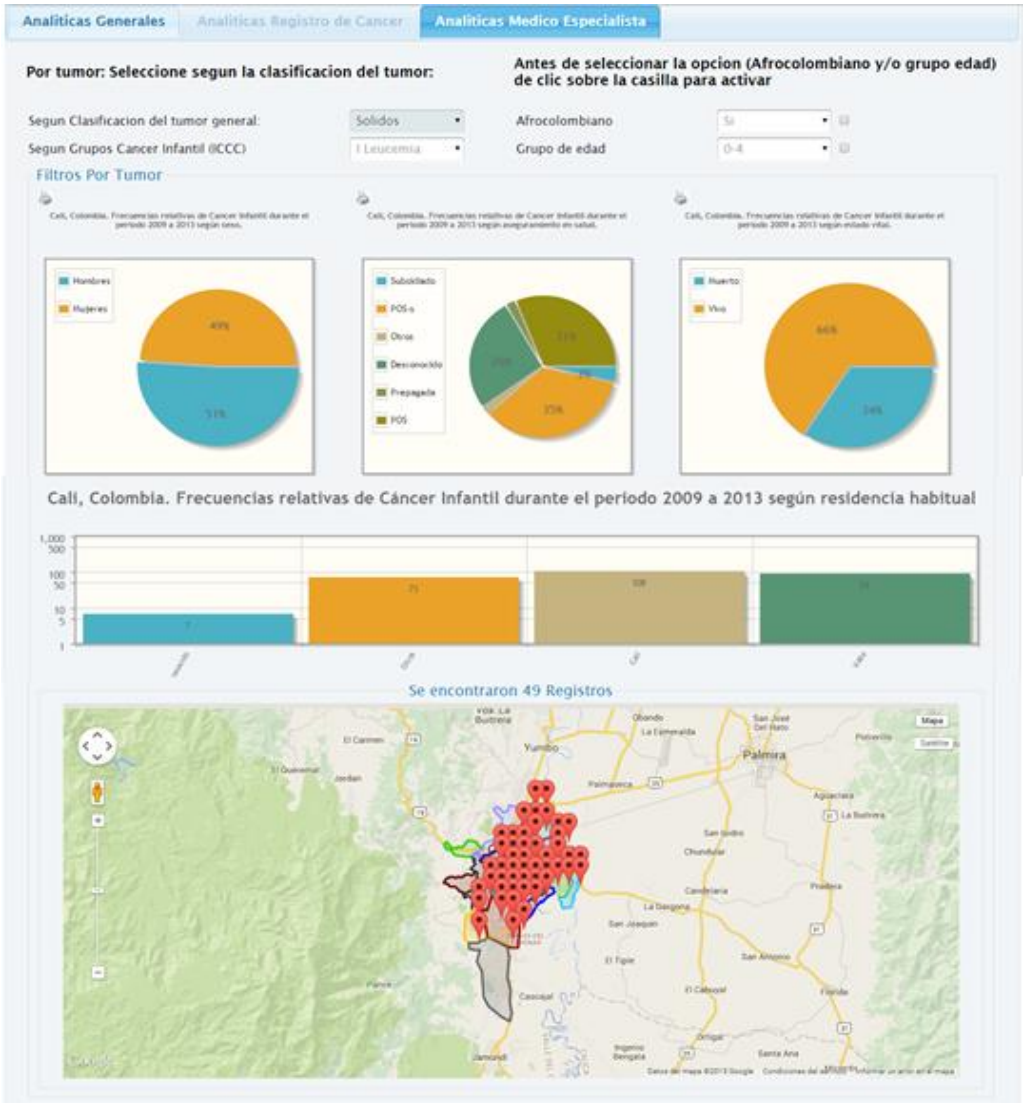


Figura 55. Control de mando analítica visual médicos especialista aplicativo Web VA_RPCC

El inicio de este control de mando depende de la selección(es) de la opción(es) que se presentan, de esta forma los gráficos se generan dependiendo de estos filtros.

En comparación con las dos analíticas presentadas anteriormente (General y Registro de cáncer), todos los gráficos se ajustan según la selección de las diferentes variables. Existen tres gráficos de torta, el primero presenta los datos por sexo de cáncer infantil (menores de 20 años), en el segundo la información según el régimen de aseguramiento y el tercero el estado vital; todas estas gráficas muestran un resumen de dichas variables, siendo importante para el proyecto e identificando su proporción según el tamaño de distribución.

La siguiente gráfica es representada por barras y muestra la distribución según la residencia, a los residentes de Cali se realizó la georeferenciación usando Google Maps, visualizando la cantidad de casos en las diferentes comunas.

Finalmente existe la parte de redes sociales que se presenta al final de cada analítica que fue implementada mediante el API de Facebook, en este se puede realizar los comentarios que ayudan a interactuar entre los diferentes usuarios y a brindar una mejor orientación de las dudas que puedan tener o sugerencias para el mejoramiento de la información presentada.

12 Evaluación del Aplicativo Web VA_RPCC

Para realizar la prueba del Aplicativo Web VA_RPCC desarrollado, se ha construido una encuesta en línea y ha sido aplicada a tres diferentes tipos de usuarios según el perfil de la visualización del control de mando de la analítica. Estas observaciones medirán el funcionamiento y la interpretación de la información presentadas en las diferentes mandos de control.

Para cada analítica se realizó una encuesta en línea diferente siendo de esta forma anónimo y se indicó en la encuesta un enlace para responder la encuesta y visualizar el aplicativo, se usó la escala de Likert (A. Oppenheim, 1992) para estas respuestas. Se tuvo en cuenta el personal que tienen relación con el Registro de Cáncer de Cali (Para el control de mando de las analíticas Registro de Cáncer y Analíticas Médico especialista) y tanto el personal relacionado y el no relacionado para el control de mando de las analíticas generales.

1. **Resultados a usuarios a visualización control de mando analítica general**

   Se obtuvieron 20 respuestas a ocho preguntas diferentes (Figura 56) de p1 a p7 en donde 1 correspondía malo y 4 a Óptimo y para la p8 de tipo abierta para comentarios y observaciones (Figura 57).

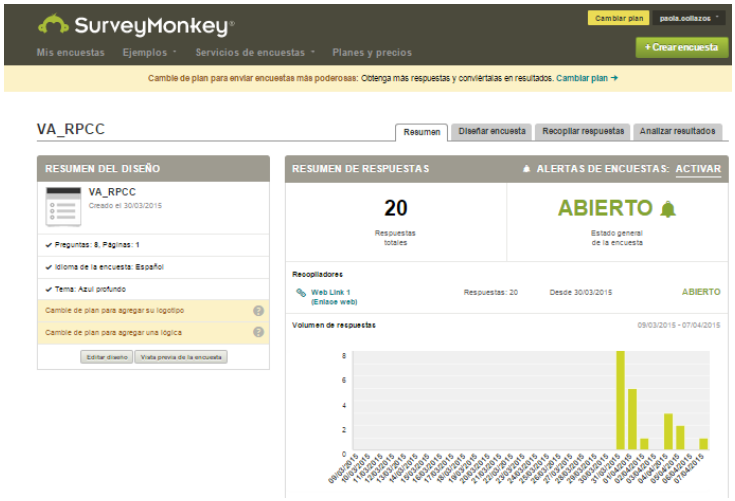
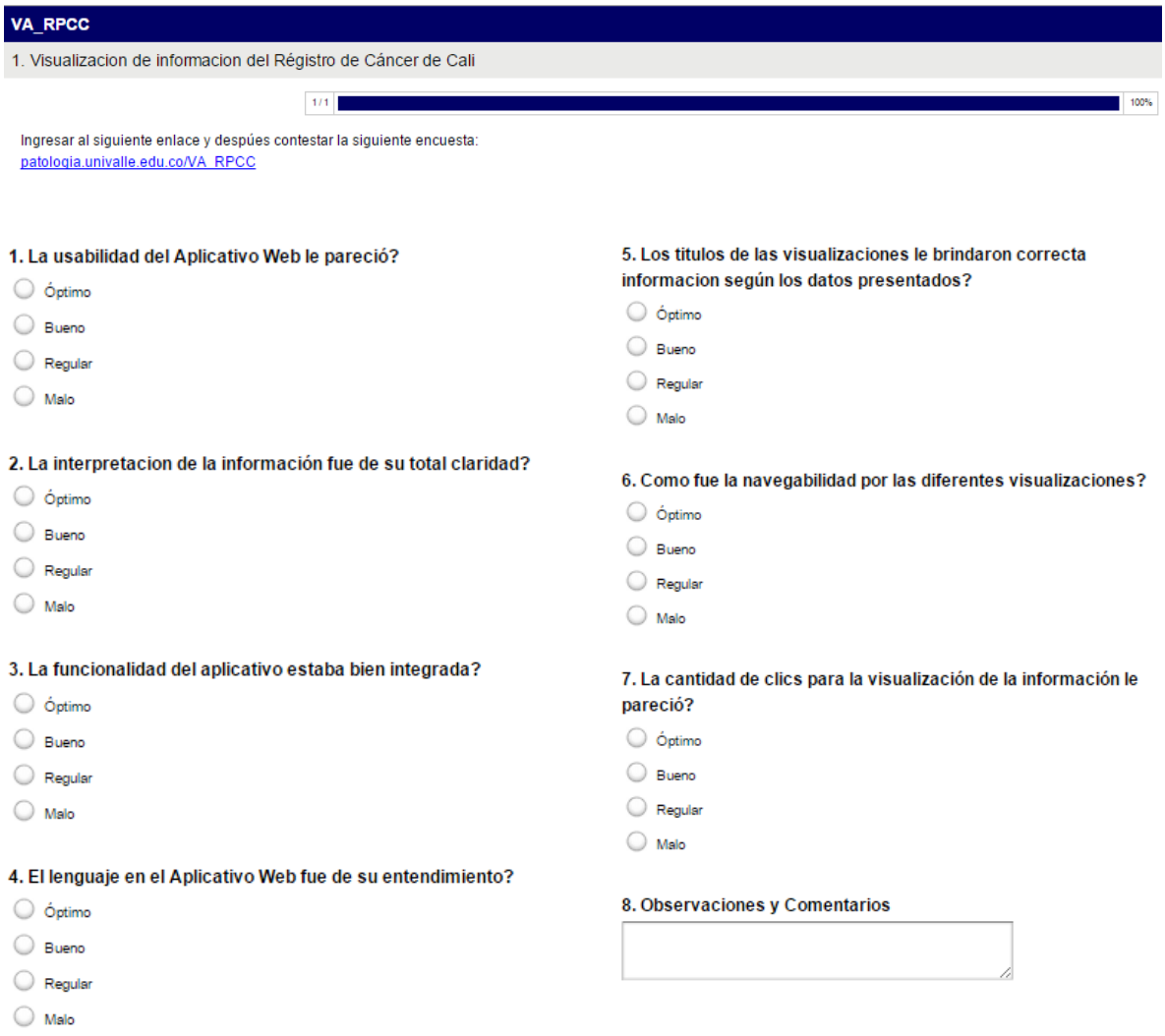


Figura 56. Resumen de preguntas y respuestas para el usuario de Analítica General



VA_RPCC

1. Visualizacion de informacion del Régistro de Cáncer de Cali

1 / 1 100%

Ingresar al siguiente enlace y despúes contestar la siguiente encuesta:
patologia.univalle.edu.co/VA_RPCC

1. La usabilidad del Aplicativo Web le pareció?
- Óptimo
- Bueno
- Regular
- Malo

2. La interpretacion de la información fue de su total claridad?
- Óptimo
- Bueno
- Regular
- Malo

3. La funcionalidad del aplicativo estaba bien integrada?
- Óptimo
- Bueno
- Regular
- Malo

4. El lenguaje en el Aplicativo Web fue de su entendimiento?
- Óptimo
- Bueno
- Regular
- Malo

5. Los titulos de las visualizaciones le brindaron correcta informacion según los datos presentados?
- Óptimo
- Bueno
- Regular
- Malo

6. Como fue la navegabilidad por las diferentes visualizaciones?
- Óptimo
- Bueno
- Regular
- Malo

7. La cantidad de clics para la visualización de la información le pareció?
- Óptimo
- Bueno
- Regular
- Malo

8. Observaciones y Comentarios

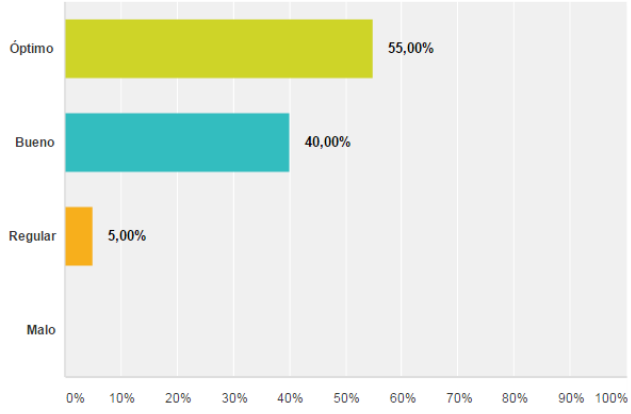
Figura 57. Preguntas de Encuesta aplicada a usuarios Analíticas generales

Los resultados obtenidos referentes a la encuesta mencionada en general fueron óptimos el cual obtuvo (65%), buenos (33.6%), regular (1.4%) y malo (0%).

No. Pregunta	Respuestas				Total
	Optimo (1)	Bueno (2)	Regular (3)	Malo (4)	
P1	11	8	1	0	20
P2	13	6	1	0	20
P3	13	7	0	0	20
P4	15	5	0	0	20
P5	10	10	0	0	20
P6	14	6	0	0	20
P7	15	5	0	0	20
Total	**91**	**47**	**2**	**0**	**140**
%	**65**	**33.6**	**1.4**	**0**	**100**

A continuación se muestran los resultados por cada pregunta evaluada:

P1. ¿La usabilidad del Aplicativo Web le pareció?

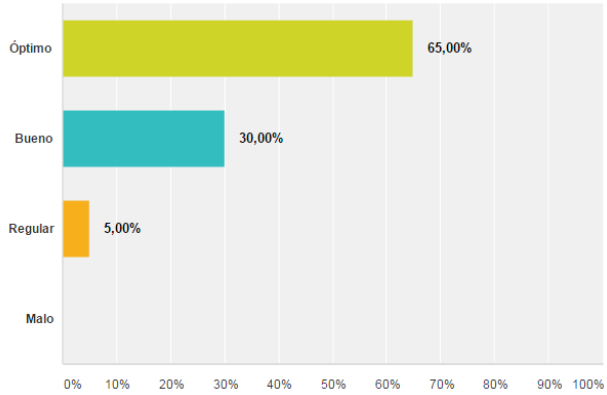


Opciones de respuesta	Respuestas	
Óptimo (1)	55,00%	11
Bueno (2)	40,00%	8
Regular (3)	5,00%	1
Malo (4)	0,00%	0
Total		20

Estadísticas básicas

Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	3,00	1,00	1,50	0,59

P2. ¿La interpretación de la información fue de su total claridad?

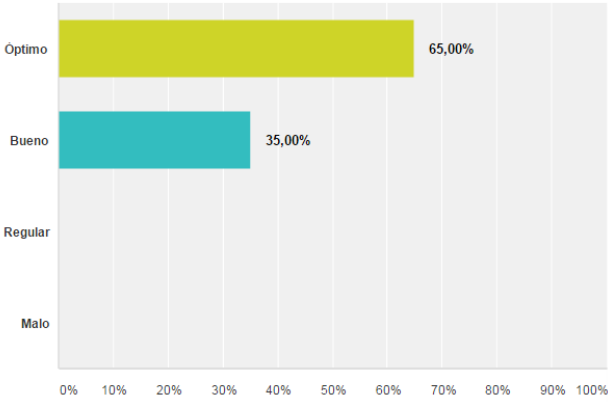


Opciones de respuesta	Respuestas	
Óptimo (1)	65,00%	13
Bueno (2)	30,00%	6
Regular (3)	5,00%	1
Malo (4)	0,00%	0
Total		20

Estadísticas básicas

Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	3,00	1,00	1,40	0,58

P3. ¿La funcionalidad del aplicativo estaba bien integrada?

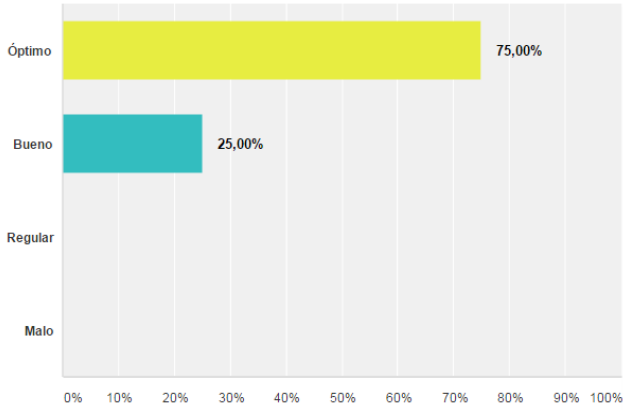


Opciones de respuesta	Respuestas	
Óptimo (1)	65,00%	13
Bueno (2)	35,00%	7
Regular (3)	0,00%	0
Malo (4)	0,00%	0
Total		20

Estadísticas básicas

Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	2,00	1,00	1,35	0,48

P4. ¿El lenguaje en el Aplicativo Web fue de su entendimiento?

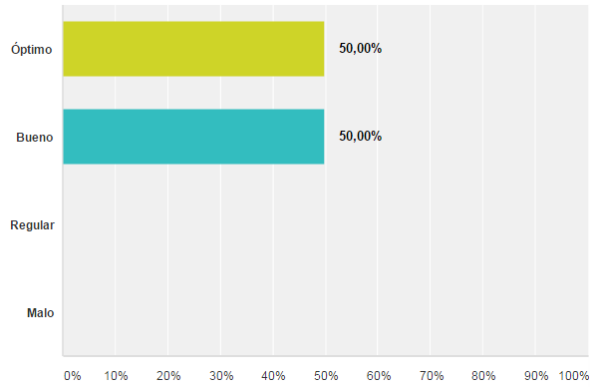


Opciones de respuesta	Respuestas	
Óptimo (1)	75,00%	15
Bueno (2)	25,00%	5
Regular (3)	0,00%	0
Malo (4)	0,00%	0
Total		20

Estadísticas básicas

Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	2,00	1,00	1,25	0,43

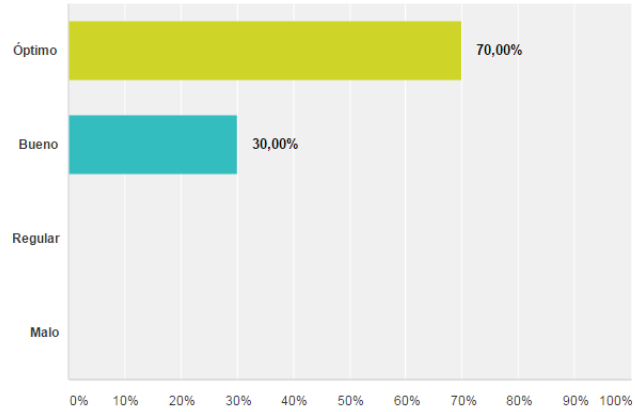
P5. ¿Los títulos de las visualizaciones le brindaron correcta información según los datos presentados?



Opciones de respuesta	Respuestas	
Óptimo (1)	50,00%	10
Bueno (2)	50,00%	10
Regular (3)	0,00%	0
Malo (4)	0,00%	0
Total		20

Estadísticas básicas				
Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	2,00	1,50	1,50	0,50

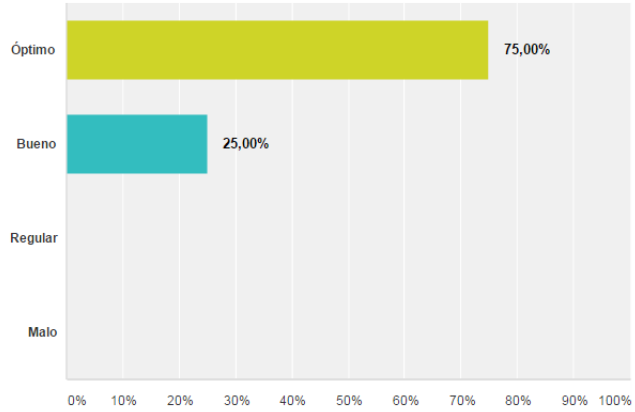
P6. ¿Cómo fue la navegabilidad por las diferentes visualizaciones?



Opciones de respuesta	Respuestas	
Óptimo (1)	70,00%	14
Bueno (2)	30,00%	6
Regular (3)	0,00%	0
Malo (4)	0,00%	0
Total		20

Estadísticas básicas				
Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	2,00	1,00	1,30	0,46

P7. ¿La cantidad de clics para la visualización de la información le pareció?



Opciones de respuesta	Respuestas	
Óptimo (1)	75,00%	15
Bueno (2)	25,00%	5
Regular (3)	0,00%	0
Malo (4)	0,00%	0
Total		20

Estadísticas básicas				
Mínimo	Máximo	Mediana	Media	Desviación estándar
1,00	2,00	1,00	1,25	0,43

P8. Observaciones y Comentarios

1. Excelente aplicativo, aunque en la localización del tumor pienso que deberían poner una opción donde se puede especificar otros. he escuchado de cáncer en un pie o en los ojos,, aunque sean pocos frecuentes y si en Cali no se ha registrado sería como una base para un posible estudio

2. En el primer gráfico: El tamaño de los números podría ser mayor. En algunos casos como en los colores azul 1987-1991 y 1972-1976, y morado 2012-2013 el valor numérico no se distingue, porque el contraste de color es muy bajo. En el segundo gráfico: en algunos casos como en el color azul, y el valor numérico no se distingue, porque el contraste de color es muy bajo. ¿Qué se supone que hacen los iconos en la parte superior-derecha? Sugiero colocar iconos más visibles (color y tamaño) y más conocidos.

3. Buena forma de mostrar la información mediante gráficas

4. En el primer ejercicio la información fue muy clara, el orden o distribución de las tablas hace que se fácil percibir la información ya que se empieza por grupos grandes y poco a poco se va desglosando la información para así obtener un resumen de aquellas variables a evaluar; y en el segundo se logra entender pero si requiere de más observación en especial el círculo de colores, quizás se debería ver otras opciones para plasmar la idea en una gráfica diferente

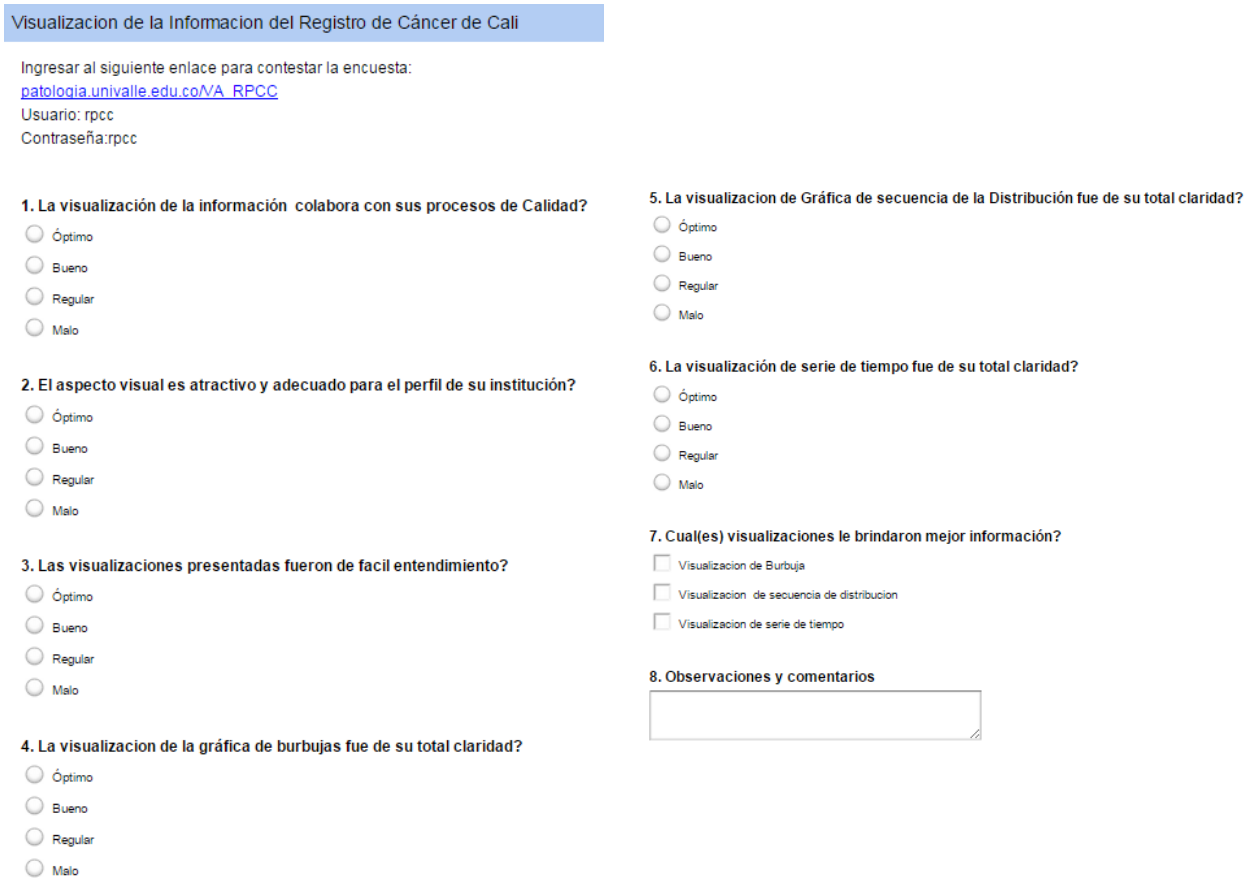
5. Sería bueno tener la opción de exportar esta información a un formato estándar, como XML o csv, para que se pueda analizar por otras aplicaciones como spss, stata o Excel, etc.

6. A primera vista debería de aparecer el rango de edades de las personas que padecen cáncer para estudios posteriores y planes de prevención. Pero en si es una excelente aplicación

De esta forma se logró evaluar la visualización implementada para este tipo de usuario y se puede observar según los resultados de la encuesta que cumplió en un 98% el buen entendimiento de este.

1.2 Resultados a usuarios visualización a control de mando analíticas Registro de Cáncer

Se obtuvieron 3 respuestas a ocho preguntas diferentes, de la pregunta p1 a p6 las posibles respuestas correspondía a malo dando un puntaje de 1 y 4 a Óptimo, en la pregunta p7 las respuesta con la posibilidad de marcar múltiples y finalmente la última pregunta de tipo abierta para comentarios y observaciones (Figura 58).


Visualizacion de la Informacion del Registro de Cáncer de Cali

Ingresar al siguiente enlace para contestar la encuesta:
patologia.univalle.edu.co/VA_RPCC
Usuario: rpcc
Contraseña:rpcc

1. La visualización de la información colabora con sus procesos de Calidad?
- Óptimo
- Bueno
- Regular
- Malo

2. El aspecto visual es atractivo y adecuado para el perfil de su institución?
- Óptimo
- Bueno
- Regular
- Malo

3. Las visualizaciones presentadas fueron de facil entendimiento?
- Óptimo
- Bueno
- Regular
- Malo

4. La visualizacion de la gráfica de burbujas fue de su total claridad?
- Óptimo
- Bueno
- Regular
- Malo

5. La visualizacion de Gráfica de secuencia de la Distribución fue de su total claridad?
- Óptimo
- Bueno
- Regular
- Malo

6. La visualización de serie de tiempo fue de su total claridad?
- Óptimo
- Bueno
- Regular
- Malo

7. Cual(es) visualizaciones le brindaron mejor información?
- Visualizacion de Burbuja
- Visualizacion de secuencia de distribucion
- Visualizacion de serie de tiempo

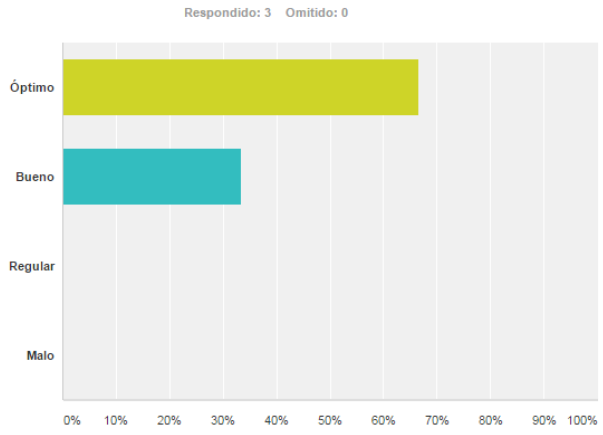
8. Observaciones y comentarios

Figura 58. Preguntas de Encuesta aplicada a usuarios Analíticas Registro de Cáncer

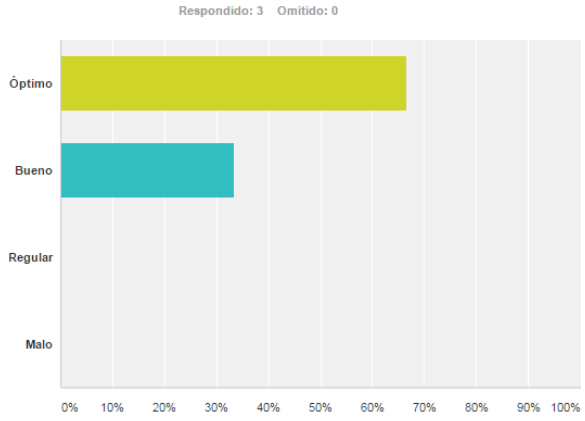
Los resultados obtenidos referentes a la encuesta mencionada en general fueron óptimos el cual obtuvo (66.7%), buenos (33.3%), regular (0%) y malo (0%).

A continuación se muestran los resultados por cada pregunta evaluada:

P1. ¿La visualización de la información colabora con sus procesos de Calidad?

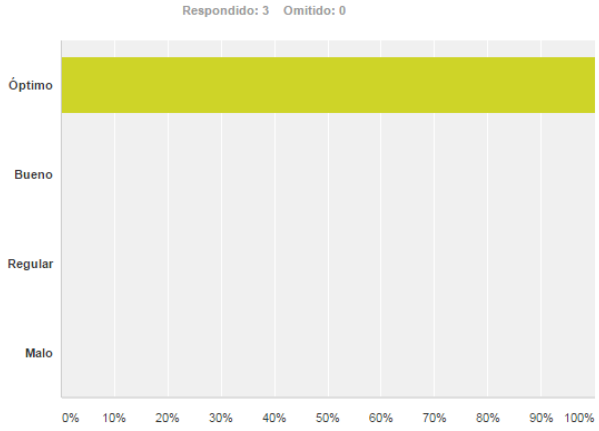


P2. ¿El aspecto visual es atractivo y adecuado para el perfil de su institución?



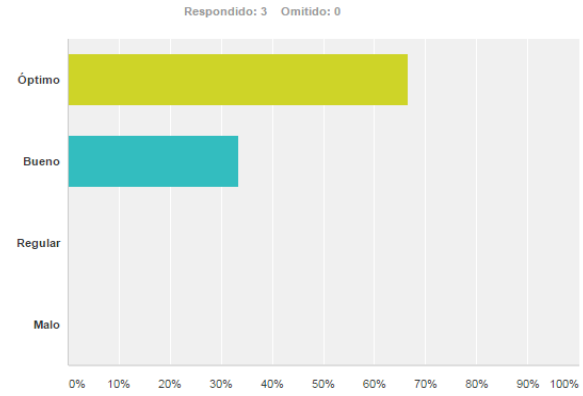
Opciones de respuesta	Respuestas	
Óptimo	66,67%	2
Bueno	33,33%	1
Regular	0,00%	0
Malo	0,00%	0
Total		3

P3. ¿Las visualizaciones presentadas fueron de fácil entendimiento?

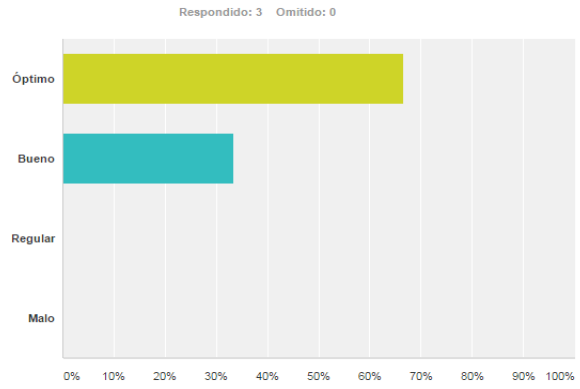


Opciones de respuesta	Respuestas	
Óptimo	100,00%	3
Bueno	0,00%	0
Regular	0,00%	0
Malo	0,00%	0
Total		3

P4. ¿La visualización de la gráfica de burbujas fue de su total claridad?

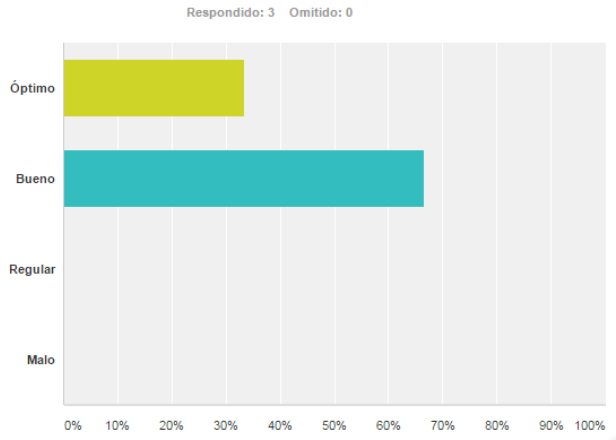


P5. ¿La visualización de la gráfica de secuencia de la distribución fue de su total claridad?



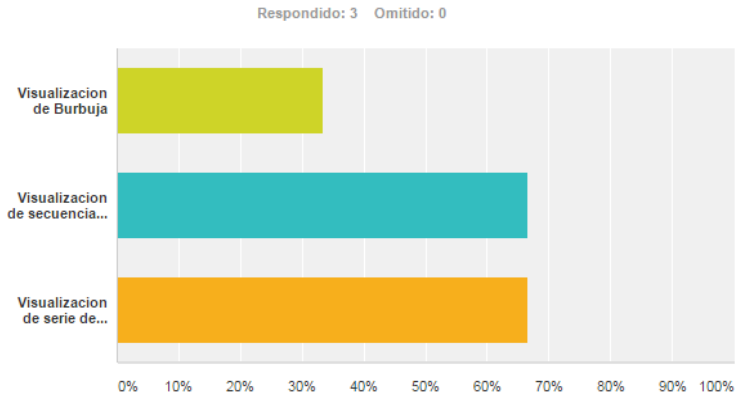
Opciones de respuesta	Respuestas	
Óptimo	66,67%	2
Bueno	33,33%	1
Regular	0,00%	0
Malo	0,00%	0
Total		3

P6. La visualización de serie de tiempo fue de su total claridad?



Opciones de respuesta	Respuestas	
Óptimo	33,33%	1
Bueno	66,67%	2
Regular	0,00%	0
Malo	0,00%	0
Total		3

P7. ¿Cuál(es) visualizaciones le brindaron mejor información?



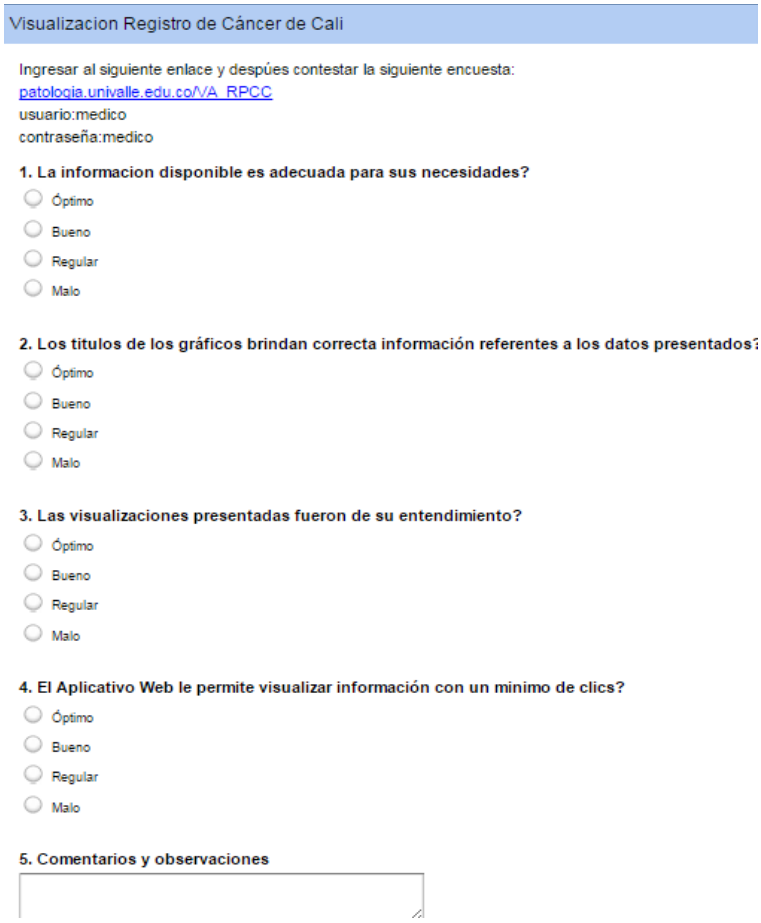
Opciones de respuesta	Respuestas	
Visualizacion de Burbuja	33,33%	1
Visualizacion de secuencia de distribucion	66,67%	2
Visualizacion de serie de tiempo	66,67%	2
Total de encuestados: 3		

P8. Observaciones y comentarios.

La información presentada fue muy clara.

1.3 Resultados a usuarios visualización a control de mando médico especialista

Se obtuvieron 3 respuestas a cinco preguntas diferentes, de la pregunta p1 a p4 las posibles respuestas correspondía a malo dando un puntaje de 1 y 4 a Óptimo, en la pregunta p5 de tipo abierta para comentarios y observaciones (Figura 59).


Visualizacion Registro de Cáncer de Cali

Ingresar al siguiente enlace y despúes contestar la siguiente encuesta:
patologia.univalle.edu.co/VA_RPCC
usuario:medico
contraseña:medico

1. La informacion disponible es adecuada para sus necesidades?
- Óptimo
- Bueno
- Regular
- Malo

2. Los titulos de los gráficos brindan correcta información referentes a los datos presentados?
- Óptimo
- Bueno
- Regular
- Malo

3. Las visualizaciones presentadas fueron de su entendimiento?
- Óptimo
- Bueno
- Regular
- Malo

4. El Aplicativo Web le permite visualizar información con un minimo de clics?
- Óptimo
- Bueno
- Regular
- Malo

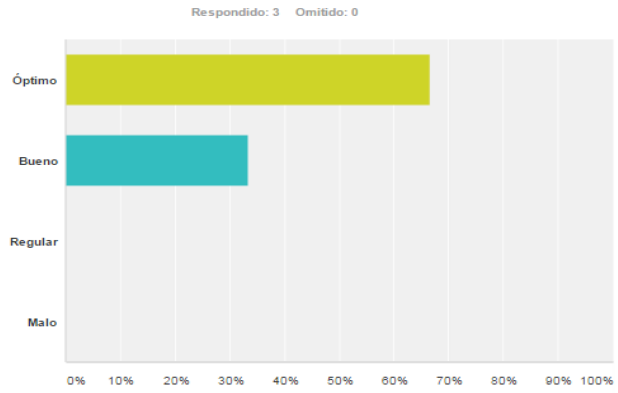
5. Comentarios y observaciones

Figura 59. Preguntas de Encuesta aplicada a usuarios Analíticas Médicos especialista

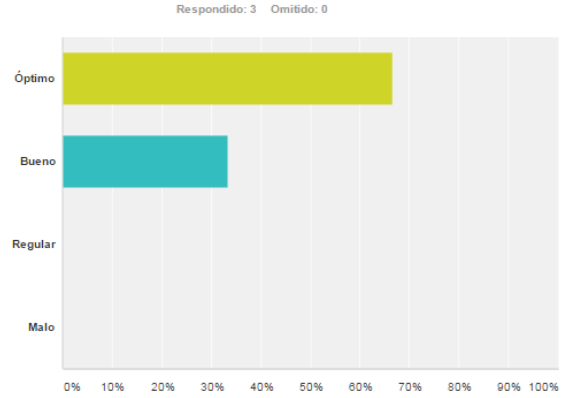
Los resultados obtenidos referentes a la encuesta mencionada en general fueron óptimos el cual obtuvo (75%), buenos (25%), regular (0%) y malo (0%).

A continuación se muestran los resultados por cada pregunta evaluada:

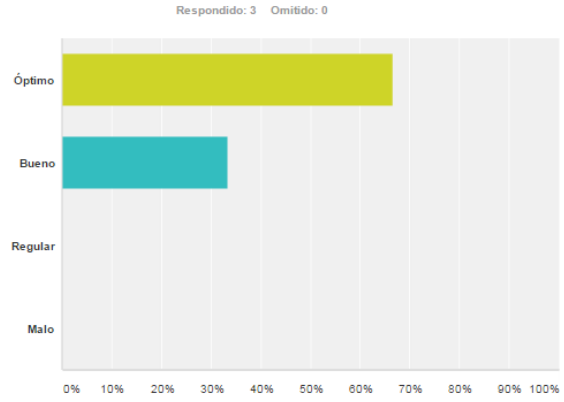
P1. ¿La información disponible es adecuada para sus necesidades?



P3. ¿Las visualizaciones presentadas fueron de su entendimiento?



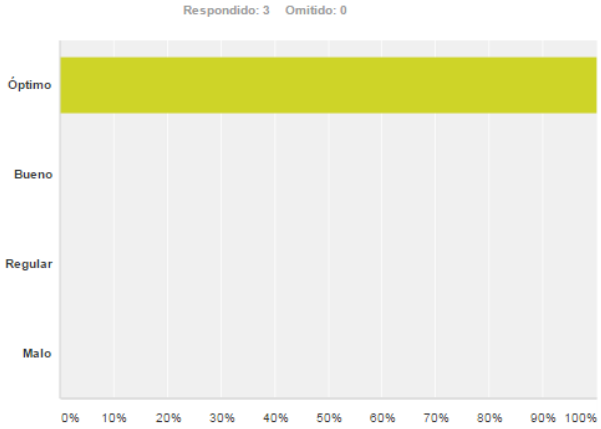
P4. ¿El aplicativo web le permite visualizar información con un mínimo de clics?



Tabulación Respuesta P1, P3 y P4.

Opciones de respuesta	Respuestas	
Óptimo	66,67%	2
Bueno	33,33%	1
Regular	0,00%	0
Malo	0,00%	0
Total		3

P2. ¿Los títulos de los gráficos brindan correcta información referentes a los datos presentados?



Opciones de respuesta	Respuestas	
Óptimo	100,00%	3
Bueno	0,00%	0
Regular	0,00%	0
Malo	0,00%	0
Total		3

Según los resultados anteriores realizados a los diferentes usuarios del Aplicativo Web se puede establecer que los evaluadores encuentran de manera significativamente óptima la visualización de la información y que ha sido de un mejor entendimiento la manera de presentar los datos del Registro de Cáncer de Cali.

13 Conclusiones

El propósito fundamental de este trabajo era adaptar una arquitectura de software enfocada en el área de la salud para la visualización de la información del RPCC con el fin de mostrarla de forma dinámica utilizando controles de mando. La adaptación de este desarrollo surgió de una arquitectura utilizada anteriormente en el área educativa, en donde se presentaba y se comparaba el nivel de aprendizaje que adquiere un estudiante de nivel superior que utiliza un ambiente virtual, haciendo uso del marco Europeo de Cualificaciones para el aprendizaje (EQF), el cual permite percibir lo que una persona en proceso de aprendizaje sabe, comprende y es capaz de hacer al culminar. De esta forma se muestra que dicha arquitectura propuesta para este proyecto es una arquitectura multidisciplinaria que colabora a otra forma dinámica de visualizar la información.

El proceso de análisis y preparación de datos fue sencillo por haber tenido la oportunidad de trabajar en el Registro de Cáncer de Cali y de participar en la creación de la base de datos y administración de esta, además el conocimiento adquirido hasta la actualidad fue un fuerte para el desarrollo del proyecto de grado.

La minería de datos ha tenido una reciente inclusión debido a la enorme preocupación de las empresas por conocer más allá de los datos que estos manejan, siendo de fácil uso y utilizando distintos tipos de algoritmos que brindan óptimos resultados. Es importante resaltar que es necesario dedicar suficiente tiempo para conocer e identificar lo que se requiere y así mismo reducir la cantidad de datos quedándonos solo con la información mínima necesaria para la disminución del esfuerzo computacional y humano.

Las técnicas de análisis basadas en la minería de datos son herramientas poderosas para la búsqueda en diferentes áreas del conocimiento. De hecho superan a las pruebas de estadística tradicional. La minería de datos es una técnica matemática útil y fiable para el análisis de datos en investigación biomédica.

La metodología CRISP-DM utilizada permitió la ejecución de este proyecto para realizar la minería de datos de forma exitosa, así se pudo definir los pasos a seguir de una manera clara y poder decidir cuáles actividades incluir de acuerdo al alcance del proyecto, además indica claramente las salidas de cada actividad y hace recomendaciones puntuales para estas.

Se ha demostrado que para realizar una minería de datos es necesario conocer un poco la información a la que nos enfrentamos y de esta forma poder aprovechar los datos para identificar la importancia de estos, además es necesario contar con una serie de mecanismos y de ayudas

tecnológicas (Weka, Excel, Stata, R) que faciliten y permiten realizar una validación de los análisis de resultados más completo y fiable.

Con esta arquitectura de software implementada se demuestra que existen múltiples librerías para implementar analíticas visuales que pueden seguir complementando la aplicación Web realizada, en donde los datos pueden ser presentados de forma más fácil y adecuada para un mejor entendimiento para diferentes tipos de usuarios.

Este proyecto permite establecer un marco teórico de referencia para nuevas investigaciones dentro de la línea de la aplicación de las tecnologías Web e implementar nuevamente dicha arquitectura Web mostrando de forma gráfica los datos con la utilización de controles de mando.

En este aplicativo se implementaron diferentes tecnologías Web utilizando principalmente el lenguaje de programación JavaScript y se utilizan las librerías (Jplot, D3,Protovis) para mostrar los diferentes gráficos y mediante los API´s se integra a Google Maps para georeferenciar los datos de las analíticas médico especialista y Facebook que es utilizada como red social para interacción con los usuarios.

Este análisis para la visualización de los datos según cada usuario es una aproximación a las características que describen la información a presentar en donde se identifican las variables de interés y de esta forma realizar la minería de datos.

Se debe dejar claro que este campo de las analíticas visuales se ha convertido en una forma útil y dinámica de divulgar la información ayudando a la toma de decisiones de forma más fácil que los resultados que son presentados en tablas.

Estudiar las diferentes tecnologías de visualización existentes nos permitió tener un panorama amplio de cuales podíamos usar para desarrollar el proyecto y la utilización de dicha arquitectura brindó flexibilidad facilitó el desarrollo del proyecto.

Para la creación de las visualizaciones gráficas se ha hecho uso de la librería jqPlot, D3 y Protovis desarrollada a partir de jQuery y que al ser de libre distribución se puede utilizar sin coste. De este modo, se permite al usuario tener un mejor entendimiento de los datos mostrándolos en gráficos y permitiendo que pueda seleccionar la información que quiera visualizar.

Es posible considerar un Big data para este aplicativo desarrollado, porque podría considerarse útil para ser adaptado al Sistema Nacional de información en Cáncer, quienes son los encargado de unificar la información de todo el país en referente a esta enfermedad y seria de mucha utilidad después de dicha integración de la información procedente de las diferentes fuentes, visualizar la información mediante estos controles de mando siendo oportuna para la toma de decisiones para los entes de salud.

En ese orden de ideas, se puede concluir que gracias al presente trabajo, se logra aportar al mejoramiento de la divulgación de la información en el Registro de Cáncer de Cali, el cual sigue mostrando lo valiosa que es esta información y que es de mucha importancia mostrar un panorama más cercano sobre el esta enfermedad.

Una vez terminado el desarrollo de la aplicación en sí, se pueden dejar planteadas varias mejoras para una futura versión de la aplicación. Entre todas las ideas, la principal es poder utilizar otras visualizaciones de las librerías para presentar las gráficas dinámicas (Jplot, D3 y Protovis) u otras que existan.

Durante el desarrollo del Aplicativo Web, fue necesario hacer uso de las diversas herramientas computacionales que apoyan el proceso de desarrollo de dicho proyecto. Inicialmente se utilizó una metodología para el proceso de minería de datos, en este caso CRISP-DM, de esta forma caracterizar la información a presentar, permitiendo una visión de los datos y detectando subconjuntos interesantes para formar las hipótesis de información oculta; para este descubrimiento del conocimiento se utilizó la herramienta WEKA, que permitió que el trabajo fuera más fácil y con resultados óptimos. Posteriormente se analizó la arquitectura de software a adaptar y las diferentes librerías que permiten la visualización de la información teniendo en cuenta diversos niveles de análisis, y finalmente realizar la evaluación del aplicativo para medir su funcionamiento e interpretación de los datos presentados. Lo anterior explica el aporte informativo que implica la elaboración de un desarrollo web, los conocimientos que se deben abordar correspondientes al área computacional independientemente del área de datos a presentar.

Finalmente, otra opción sería la posibilidad de desarrollar esta aplicación para dispositivos móviles, para que de esta manera, se pudiera buscar a la gente en cualquier momento de necesidad.

Bibliografía

A. Oppenheim (1992), Questionnaire Design, Interviewing and Attitude Measurement. Pinter, London.

Ballvé, A. M. (2007). Tablero de Control, Información para crear valor, (Emece – Planeta, ISBN Tablero de Control.).

Berry, M. J. A., & Linoff, G. . (2000). Mastering Data Mining, The Art and Science of Customer Relationship Management., (New York: John Wiley & Sons, Inc).

Beyer, M. (n.d.). Gartner Says Solving "Big Data" Challenge Involves More Than Just Managing Volumes of Data.

Bravo, L. E., Collazos, T., Collazos, P., García, L. S., & Correa, P. (2012). Trends of cancer incidence and mortality in Cali, Colombia. 50 years experience. *Colombia Médica*, *43*.

Brito, P. (2008). Procesos de explotación de información basados en sistemas inteligentes, Universidad Nacional de La Plata, Argentina.

Bonney, W. (2013). Applicability of Business Intelligence in Electronic Health Record. Procedia - Social and Behavioral Sciences, 257 – 262.

Carrascal, E., Guerrero, A., & Llanos, G. (2002). Manual de Registros de Cáncer, *Primer Edi*(Corporacion editorial Médica del Valle).

Clements, P. (1996). "A Survey of Architecture Description Languages." *Proceedings of the International Workshop on Software Specification and Design*.

Correa, P. (2012). The Cali Cancer Registry an example for Latin America. *Colomb Med (Cali)*, 2012;43(4):244–245.

Cross, M. (2002). Decision support systems: using technology for successful management, *75*(CMA Management, Hamilton), 48.

Davenport, T. H., & Harris, J. G. (2007). Competing on Analytics, The New Science of Winning, (Boston : Harvard Business School Press).

Fayyad, U. P., & Shapiro, S. P. (1996). Knowledge discovery and data mining: Towards a unifying framework. Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD)., (Portland).

Florian, B. (2013). Technology-enhaced support for lifelong competence development in higher.

Florian, B., Glahn, C., & Drachsler, H. (2011). Activity-based learner-models for learner monitoring and recommendations in Moodle. *… Ubiquitous Learning*, 1–14. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-23985-4_10

Frawley, W. J., Piatetski-Shapiro, G., & Matheus, C. J. (1991). Knowledge Discovery in Databases, (Menlo Park, California), 1–27.

Gónzalez, J. I. (n.d.). Modelos y Ejemplos de Dashboard - CMI (I), (I).

Han, J., & Kamber, M. (2006). Data Mining, Concepts and Techniques, (Amsterdam: Morgan Kaufmann Publishers.).

Hernández, J., Ramirez, M. ., & Ferri, C. (2004). Introducción a la Minería de Datos, (Editorial Pearson Prentice Hall, pp.680. Madrid, España).

IDC. (n.d.). Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO.

Keim, D. ., Kohlhammer, J., Ellis, G., & Mannsmann, F. (2010). Mastering the Information Age. Solving Problems with Visual Analytics, (Euro- graphics Association, Goslar).

Keim, D., Andrienko, G., Fekete, J., & Carsten, G. (2008). Visual Analytics : Definition , Process , and Challenges, 154–175.

Kriegel, H.-P., Borgwardt, K. M., Kröger, P., Pryakhin, A., & Zimek, A. (2007). Future trends in data mining, (Springer Science + Business Media).

Laudon, K., & Laudon, J. (2004). Sistemas de Información Gerencial, (PRENTICE HALL), 534.

López Viñeglas, A. (1999). El Cuadro de Mando y los Sistemas de Información para la Gestión Empresarial, (Posibilidad de Tratamiento Hipermedia. Madrid, Editora AECA).

MacQueen. (1967). Some methods for classification and analysis of multivariate observations, (Proc. 5th Berkeley Symp. Math. Statisi. Prob., 1:281-297).

Microstrategy. (2006). The 5 styles of Business Intelligence: Industrial- Strength Business Intelligence., (www.microstrategy.com), 1–87.

Nelson, E., & Arias, O. (2013). Registros poblacionales de cáncer: avances en Colombia, Chile y Brasil.

Norton, D., & Kaplan, R. (1992). The Balanced Scorecard. Measures that drive performance, (Harvard Business Review).

Olson, D., & Courtney, J. (1992). Decision Support Models and Expert Systems, (New York, Estados Unidos: Macmillan Publishing Company).

Pete, C. (NCR), Julian, C. (SPSS), Randy, K. (NCR), Thomas, K. (SPSS), Thomas, R. (DaimlerChrysler), Colin, S. (SPSS), & Rudiger, W. (2000). *Crips DM 1.0 Step by Step Data Mining Guide*.

Pyle, D. (2003). Business Modeling and Data Mining. *Morgan Kaufmann Publishers.*

SAS Entreprise Miner. (2012). SAS Institute nc. World Headquartes. Retrieved from http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

Thomas, J., & Cook, K. (2005). Illuminating the Path: Research and Development Agenda for Visual Analytics, (IEEE Press).

Turban, E, Aronson, J. ., Liang, T., & Sharda, R. (2007). Decision Support Business Intelligence Systems, (Upper Saddle River : Pearson Prentice Hall).

Turban, Efraim, & Aronson, J. E. (2001). Decision Support systems and Intelligent systems, *Sexta Edic*, Ed. Prentice–Hall.

Wagar Haque, B. U. (2014). Using Business Intelligence to Analyze and Share Health System Infrastructure Data in a Rural Health Authority. JMIR Medical Informatics.

Weiss, S. . and I. (1998). Predictive Data Mining, (San Francisco: Morgan Kaufmann).

Witten, I. ., & Frank, E. (2005). Data Mining, Practical Machine Learning Tools and Techniques, (New York: Morgan Kaufmann Publishers.).

Wong, M., & Leung, K. (2002). Data Mining Using Grammar Based Genetic Programming And Application, (Editado por Kluwer Academic Publishers, pp. 213. Estados Unidos).

Zimmermann, A., Specht, M., & Lorenz, A. (2005). Personalization and Context Management. User Modeling and User-Adapted Interaction, (doi:10.1007/s11257-005-1092-2), 15(3–4),275–302.

ANEXO 1

patologia.univalle.edu.co/VA_RPCC