Universidad de las Ciencias Informáticas

Facultad 8

Desarrollo de algoritmos de regresión para el Servidor de Análisis Estadísticos: R-SERVER.

Trabajo de Diploma para optar por el Título de Ingeniero en Ciencias Informáticas.

Autoras: Lisandra Cordeiro Rodríguez

Dania Cid Cespedes

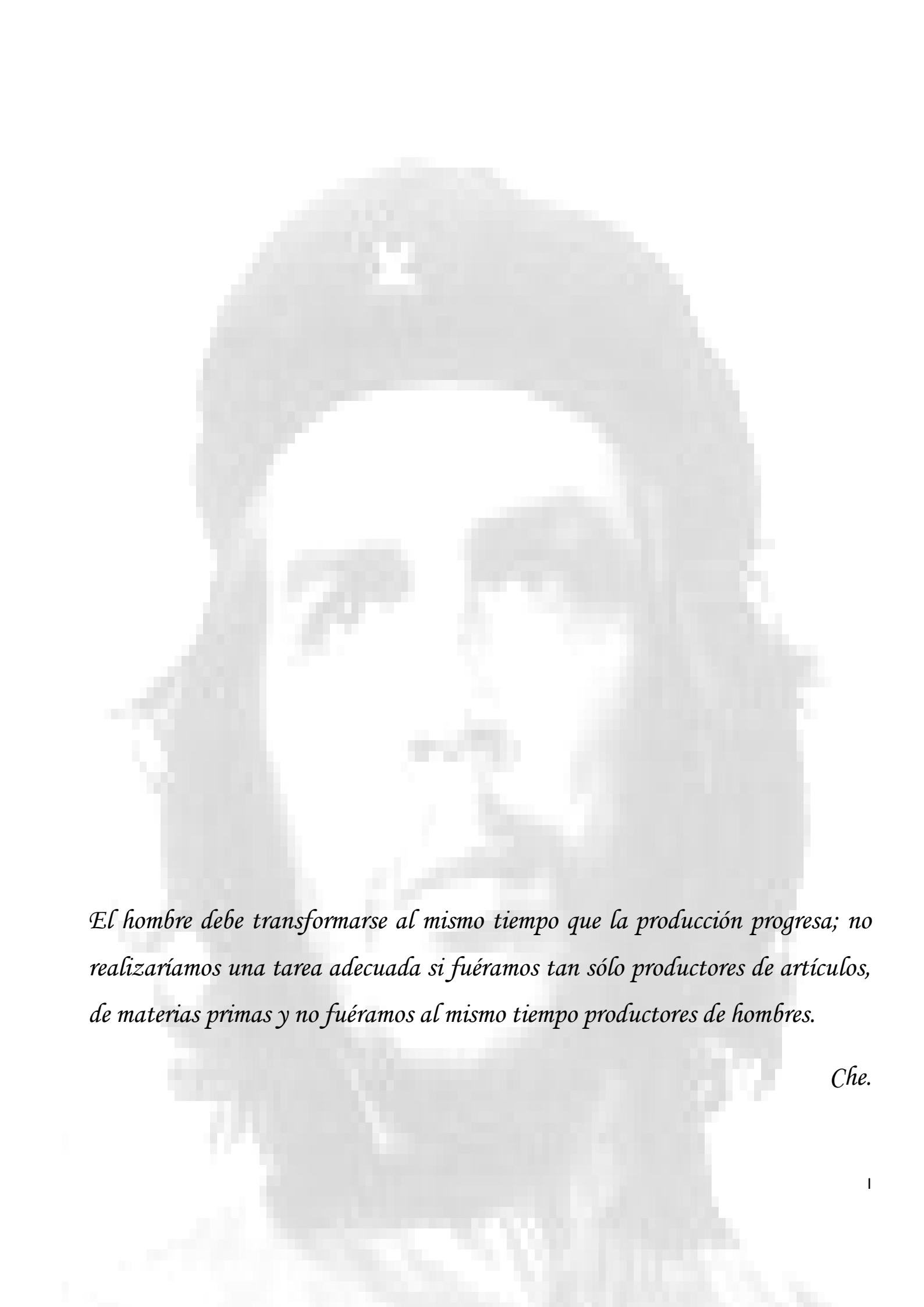
Tutor: Ing. Yunior Miguel Almaguer Bajuelo

Cotutores: Ing. Aldis Joan Abreu Medina

Ing. Juan Carlos Quevedo Lussón

Ciudad de la Habana, Junio 2010

"Año 52 de la Revolución"



El hombre debe transformarse al mismo tiempo que la producción progresa; no realizaríamos una tarea adecuada si fuéramos tan sólo productores de artículos, de materias primas y no fuéramos al mismo tiempo productores de hombres.

Che.

Declaración de Autoría

Declaramos que somos los únicos autores de este trabajo y autorizamos a la infraestructura Productiva de la Universidad de las Ciencias Informáticas (UCI) a hacer uso del mismo para su beneficio y como estime conveniente.

Para que así conste firmamos la presente a los ___ días del mes de _____ de 2009.

_____

Dania Cid Cespedes

Autor

_____

Lisandra Cordeiro Rodríguez

Autor

_____

Ing. Yunior Almaguer Bajuelo

Tutor

_____

Ing. Aldis Joan Abreu Medina

Cotutor

_____

Ing. Juan Carlos Quevedo Lussón

Cotutor

Agradecimientos:

Lisandra:

A mis tutores por su fe y devoción que ha consagrado a nuestro trabajo, a todos mis profesores por su esfuerzo y dedicación en el empeño de enseñarnos; a José Cayetano Bango, por su sabiduría incondicional, a mi compañero Yandris, Luis, Frank y a todos mis compañeros de proyecto por su ayuda sin límites. A mis amistades que siempre han estado presente en cada momento, a los trabajadores del Nodo de redes del docente cuatro, a Fidel y Raúl por darme la oportunidad de ser universitaria.

A todos ellos muchas gracias.

Dania:

Agradezco a:
Mi mamá y a mis abuelos porque han dedicado todo su esfuerzo para que yo pudiera lograr convertirme en una profesional y han seguido cada paso mío por esta universidad viviendo mis alegrías y tristezas junto a mí. Mi papá por brindarme su apoyo incondicional y a mi padrastro que es un padre para mí y ha estado a mi lado siempre y me ha brindado todo su apoyo y confianza. Mis tutores porque realmente fueron de gran ayuda para la realización de este trabajo y porque siempre nos dieron su apoyo. A todos aquellos que me han acompañado durante estos 5 maravillosos años, a los cuales considero mi segunda familia, a esos que con amor, honestidad y mucha humildad se han ganado un lugar en mi corazón. Mis amigas Yanitza, Kenia, Elizabeth e Idalmis que siempre han estado ahí para todo lo que he necesitado y dispuestas a darme todo su apoyo y amor. A Yandris y Michel por su apoyo incondicional, a los trabajadores del nodo de redes del docente 4. En fin todos los que de una forma u otra me ha brindado todo su apoyo y amor para que haya sido posible la realización de este sueño. A nuestro Comandante Fidel, Raúl y a la Revolución Cubana por darnos la oportunidad de convertirnos en profesionales y crear una escuela como esta.

Gracias

Dedicatoria:

Lisandra:

A mis grandes amores:
A mis padres y a Eddy Félix, porque los siento siempre conmigo y porque los amo.

Dania:
A mi madre, mis abuelos, mi papá, mi hermano y mi padrastro, porque ellos son mi espíritu y mi fuerza, pues a ellos les debo todo lo que soy.

Resumen:

En la actualidad, con el creciente auge de la informatización en todas las esferas de la sociedad y el desarrollo de las Tecnologías de la Informática y las Comunicaciones, las mayorías de las empresas e instituciones necesitan automatizar los trabajos que realizan. Debido al creciente y constante volumen de datos y documentación que se transmite en la Oficina Nacional de Estadística (ONE), se hace necesario informatizar los procesos de análisis de datos estadísticos. Por lo tanto, el objetivo principal de este trabajo es desarrollar algoritmos de regresión para el Servidor de Análisis Estadístico: R-SERVER; con vista a contribuir al desarrollo de un sistema eficiente. El documento recoge un estudio de las características de las herramientas y librerías usadas para la propuesta de solución, la descripción de la metodología, arquitectura, los algoritmos a implementar y la validación de la solución mediante diferentes tipos de pruebas.

Palabras claves: Regresión, R-SERVER, Análisis, Estadística.

Índice

Índice de Figuras

Índice de Tablas

Introducción

Actualmente las organizaciones suelen encontrarse dentro de estructuras identificadas con un cambio continúo, es por eso que las empresas públicas así como las privadas deben tener la capacidad de ser adaptativas, de aprender cómo resolver problemas y crear conocimiento. Las organizaciones, buscan obtener los mejores resultados en cuanto a su gestión organizacional, adaptando así la flexibilización como estrategia, con el objetivo de ajustarse a un mercado globalizado, dando origen a un proceso que incide en su sistema estructural. “Así pues, una empresa flexible es la que se orienta hacia los clientes, posee tecnología nueva y presenta acuerdos laterales de organización e innovación”.(MOURITSEN 1999)

Dentro de las aplicaciones para gestionar el flujo de información en las actividades del negocio, existen dos tipos: las aplicaciones que manejan las transacciones y las aplicaciones estadísticas que ayudan a convertir los datos en información útil para la toma de decisiones. En el uso de la Minería de Datos como soporte a la toma de decisiones de las actividades del negocio se quiere mucho más que el uso de aplicaciones con sofisticadas técnicas como las redes neuronales o árboles de decisiones sobre las tablas de datos, ya que es uno de los pasos del proceso de descubrimiento de conocimiento de la base de datos, (KDD, según sus siglas en inglés) [1] y es un proceso que consta de diferentes períodos, en los cuales se utiliza como apoyo, técnicas relacionadas con la estadística, el reconocimiento de patrones y algoritmos de aprendizajes, entre otros.

En las actividades del negocio, “la Minería de Datos es un conjunto de metodologías, aplicaciones y tecnologías que permiten transformar datos de los sistemas e información no estructurada en información estructurada, para la explotación directa o para el análisis y transformación en conocimiento y así dar soporte a la toma de decisiones”.(MARCANO AULAR 2007)

La Minería de Datos es una técnica que está implícita dentro de la estadística inferencial, que es la que se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Dentro de estas técnicas se encuentran: las pruebas de hipótesis, la estimación, correlación y análisis de regresión.

[1] KDD es el proceso completo de extracción de conocimientos, no trivial, previamente desconocidos y potencialmente útil a partir de un conjunto de datos

En Cuba fue creada en 1999 la Unidad Nacional Coordinadora de Farmacovigilancia (UNCFv), con el objetivo de definir, diseñar y desarrollar los sistemas de tratamiento de la información y administrar la base de datos nacional "VigiBaseCuba". Aplicando una serie de transformaciones, validaciones y la adecuación de la metodología CRISP-DM [2]para la elaboración de proyectos de Minería de Datos. (DEBESA F 2007)

El Centro de Tecnología de Gestión de Datos (DATEC) está compuesto por cuatro líneas de desarrollo, en la que se encuentra la línea de integración de soluciones donde se trabaja en un Servidor de Análisis Estadístico, con el objetivo de realizar análisis estadístico a partir de algoritmos de regresión. El principal y actualmente cliente del servidor R-SERVER es el producto PATDSI (Paquete de Herramientas para la Ayuda a la Toma de Decisiones), el cual es concebido como una plataforma web única de inteligencia de negocio que integra en sí las funcionalidades más recurrentes y específicas necesarias para la toma de decisiones en diferentes contextos. PATDSI provee entre sus principales herramientas un Generador Dinámico de Reportes capaz de generar reportes dinámicos desde entornos web, que pueden ser tabulares y análisis estadísticos. Para hacer más completo el funcionamiento de PATDSI, el Servidor de Análisis Estadístico R-SERVER debe contar con el desarrollo de algoritmos de regresión lineal y polinomial (curvilínea), para en un futuro poder integrar el servidor a otros clientes que necesiten realizar análisis estadísticos.

Por lo antes mencionado se define como **problema a resolver**:

¿Cómo desarrollar los algoritmos de regresión para el Servidor de Análisis Estadístico: R-SERVER?

Se define como **Objeto de Estudio** algoritmos estadísticos, **y Campo de Acción** los algoritmos de regresión.

Como **objetivo general** se define:

Desarrollar los algoritmos de regresión para el Servidor de Análisis Estadístico: R-SERVER.

[2] Metodología (CRISP-DM): se distingue entre el modelo de referencia y la guía de usuario. El modelo de referencia presenta una descripción rápida de fases, las tareas, y sus salidas, y describen que hacer en el proyecto de Minería de Datos. La guía de usuario da consejos más detallados e insinuaciones para cada fase y cada tarea dentro de una fase, y representa como realizar un proyecto de Minería de Datos

Los **objetivos específicos** son:

- Realizar el estudio del estado del arte de los algoritmos de regresión, herramientas y metodología para el análisis de la información.
- Desarrollar los algoritmos de regresión para el Servidor de Análisis Estadístico.
- Validar el desarrollo de los algoritmos de regresión para el Servidor de Análisis Estadístico: R-SERVER.

Para dar respuesta a los anteriores objetivos específicos se elaboraron las **tareas** siguientes:

- Identificación de las herramientas y metodología a usar en el análisis y diseño para el Servidor de Análisis Estadístico.
- Documentación de análisis, diseño e implementación de los algoritmos de regresión para el Servidor de Análisis Estadístico.
- Validación de los resultados obtenidos mediante la elaboración de casos de prueba operacionales y del sistema.

Métodos Científicos de Investigación.

No es más que un conjunto de procedimientos que se utilizan para el análisis y estudio de las características del objeto de investigación que no son observables directamente.

Métodos teóricos:

- **Analítico - Sintético:** Se utilizó para el procesamiento de la información que permite analizar los documentos y la extracción de los elementos más importantes acerca del proceso de Minería de Datos para precisar el análisis y diseño, así como para el arribo a las conclusiones de la investigación.
- **Histórico - Lógico:** Mediante este método se pudo determinar de forma teórica las tendencias actuales de desarrollo de la Minería de Datos y algoritmos de regresión así como sus herramientas y su continua evolución.

Estructura del trabajo de diploma.

La tesis se estructura en tres capítulos:

Capítulo 1 Fundamentación teórica:

Se analiza el estado del arte de las diferentes herramientas y librerías que existen en el mundo para el análisis de datos, se identifican sus principales características, y se realiza un análisis crítico de las mismas a través de la comparación teniendo en cuenta los aspectos más relevantes. Se analizan los algoritmos de regresión, centrando la atención en los de regresión lineal (simple) y regresión polinomial (curvilínea) para desarrollar el R-SERVER. Se definen las herramientas, metodología y lenguaje a utilizar, y se argumenta su elección.

Capítulo 2 Análisis y Diseño:

Se identifican los principales conceptos asociados a los algoritmos de regresión, se documentan adicionalmente las funcionalidades a desarrollar en el Servidor de Análisis Estadístico: R-Server. Es definida la estructura del sistema aplicando la técnica de modelado de Casos de Uso propuesta por la metodología de desarrollo. Es documentada la arquitectura del sistema y los patrones de diseño a implementar en el mismo.

Capítulo 3 Implementación y Validación:

Se realiza la implementación de los algoritmos de regresión lineal y regresión polinomial. Es definida la estructura del JSON de entrada y salida, así como la realización de pruebas al módulo, la obtención de los casos de pruebas y sus resultados. Consiste en comprobar que el algoritmo cumpla con la especificación del problema.

Fundamentación Teórica

1.1 Introducción

En este capítulo se abordan temas relacionados con el estado del arte de las diferentes herramientas y librerías que existen en el mundo para el análisis de datos, como sus características, y un análisis crítico de la misma a través de la comparación teniendo en cuenta los aspectos más relevantes. Se analizan los algoritmos de regresión, donde se centra la atención en los de regresión lineal (simple) y regresión polinomial (curvilínea) para desarrollar el R-SERVER. Se establecen las herramientas, metodología y lenguaje a utilizar, y se argumenta su elección.

1.2 Minería de Datos.

La Minería de Datos es una de las herramientas más importantes que utilizan dentro de los programas de gestión del conocimiento como soporte a la toma de decisiones. Con esta se logra extraer la información oculta o análisis de datos mediante técnicas estadísticas usadas en grandes bases de datos. “Las herramientas de Minería de Datos pueden responder a preguntas de negocios empresariales no planteadas o que pueden consumir demasiado tiempo para ser resueltas. La Minería de Datos, como herramienta de búsqueda de información, se utiliza como sistema de apoyo a la toma de decisiones de las altas direcciones de las empresas”. (TOLEDANO 2006)

El algoritmo de Minería de Datos es el mecanismo que crea modelos de Minería de Datos. Estos algoritmos analizan primero un conjunto de datos, buscando patrones y tendencias específicas. Después, el algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de Minería de Datos.

El modelo de Minería de Datos que crea un algoritmo puede tomar diversas formas, incluyendo:

- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción.
- Un árbol de decisión que predice si un cliente determinado comprará un producto.
- Un modelo matemático que predice las ventas.
- Un conjunto de clusters que describe cómo se relacionan los escenarios de un conjunto de datos.

Los algoritmos de Minería de Datos son:

- Algoritmos de clasificación
- Algoritmos de segmentación
- Algoritmos de asociación
- Algoritmos de análisis de secuencia
- Algoritmos de regresión

Según lo antes planteado la Minería de Datos es una de las más importantes herramientas que se utiliza para la extracción de información oculta o en análisis de datos mediante algoritmos estadísticos. Estas herramientas de búsqueda de información se utilizan en sistemas de apoyo a la toma de decisiones de las empresas. Ya que estos algoritmos analizan los datos buscando tendencias específicas y utilizando los resultados del análisis para definir los parámetros del modelo de Minería de Datos.

1.3 Herramientas de Minería de Datos.

1.3.1 DBMINER

Sistema interactivo desarrollado por la Universidad de Simón Fraser de Canadá. Está concebido para la extracción del conocimiento de bases de datos relacionales, almacenes de datos y datos de la web con la facilidad de uso excepcional y de gran versatilidad. Dentro de su arquitectura de diseño es importante destacar el procesamiento analítico en línea (OLAP) y la minería analítica en línea (OLAM). La herramienta de DBMiner posee dos modos de trabajo la Vía interfaz gráfica y de interfaz de script. Esta herramienta es un socio de Microsoft Data Warehousing de la Alianza, y tiene una estrecha colaboración con los líderes mundiales, tales como Microsoft, IBM, HP y Boeing. Posee una licencia GPL y comercial, se pueden instalar en sistemas operativos como: Linux y POSIX. (ORALLO 2004)

1.3.2 WEKA

Esta Herramienta fue desarrollada por la universidad de Waikato (Nueva Zelanda), y se inició en el año 1993. Es una de las primeras aplicaciones Open Source. Es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Su librería de algoritmos es utilizada por la mayoría de herramientas de este tipo, está disponible libremente bajo la licencia pública general de

GNU lo cual ha impulsado que sea una de las herramientas más utilizadas en el área en los últimos años, brinda servicios de documentación, soporte y consultoría. También una herramienta muy portable ya que está completamente implementada en Java y puede correr en casi cualquier plataforma, es de fácil uso gracias a su interfaz gráfica de usuario y brinda funcionalidades de extensibilidad y reusabilidad, además contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado. (ORALLO 2004)

1.3.3 KNIME

Es una plataforma de Minería de Datos que permite el desarrollo de modelos en un entorno visual, la cual está construido bajo la plataforma Eclipse y programado en Java. Está concebido como una herramienta gráfica y dispone de una serie de nodos que encapsulan distintos tipos de algoritmos y flechas que representan flujos de datos que se despliegan y se combinan de manera gráfica e interactiva. Brinda la posibilidad de utilizar la llamada directa y transparentemente a Weka y/ o de incorporar de manera sencilla código desarrollado en R o Python. Es multiplataforma, es decir, que se puede instalar en cualquier sistema operativo, con una licencia Aladdin y actualmente está en producción. (ORALLO 2004)

1.3.4 ORANGE

Es un programa informático que se utiliza para la realización de Minería de Datos y análisis predictivo. Su desarrollo fue llevado a cabo en la facultad de informática de la Universidad de Ljubljana. Consta de una serie de componentes desarrollados en C++ que implementan algoritmos de Minería de Datos, así como operaciones de pre- procesamiento y representación gráfica de datos. Pose componentes que pueden ser manipulados desde programas desarrollados en Python o a través de un entorno gráfico. Se encuentra bajo licencia GPL y es Multiplataforma, es decir se puede instalar en Linux, Windows y Macintosh. (ORALLO 2004)

1.4 Librerías para el análisis de Datos.

Las librerías de análisis de datos son un conjunto de métodos que implementan funcionalidades y utilidades básicas como el acceso a datos, modelos de redes neuronales, métodos bayesianos, exportación de resultados etc. Las librerías se encargan principalmente de facilitar el desarrollo de las tareas de análisis de datos que son más complejas, como el diseño de experimentos, el problema de las librerías, es que es precisa la comprensión de conocimientos de programación.(ORALLO 2004)

1.4.1 XELOPES

Es una librería con licencia pública GNU para el desarrollo de aplicaciones de Minería de Datos. Esta librería está implementada para que sea eficiente para la mayoría de los algoritmos de aprendizaje, por eso, es importante destacar que el usuario puede desarrollar aplicaciones particulares de Minería de Datos. Sus principales características son Acceso a datos, Modelos de redes neuronales, Métodos de agrupamiento, Métodos de reglas de asociación, Árboles lineales, Árboles no lineales, Exportación de datos. Tiene implementaciones para Java, C++, C# y CORBA y las interfaces de servicios web están disponibles actualmente. (ORALLO 2004)

1.4.2 2MLC++

La colección de máquinas de aprendizaje en C++ es un conjunto de librerías que fueron desarrolladas por la Universidad de Standford. La mayoría de las versiones son bajo dominio de investigación, a excepción de la versión 1.3.x, que se distribuye bajo licencia de dominio público. (ORALLO 2004)

Las principales características son:

- Acceso a datos (archivos con formato plano).
- Transformaciones de datos.

1.4.3 SPSS CLEMENTINE

Es uno de los sistemas de Minería de Datos más conocidos. Posee una herramienta visual desarrollada por ISL que tiene una arquitectura cliente/servidor. (ORALLO 2004)

Este sistema se caracteriza por:

- Acceso a datos (fuentes de datos, archivos ASCII).
- Procesamiento de Datos.
- Técnicas de Aprendizaje (redes neuronales, reglas de asociación).
- Técnicas de evaluación de modelos.
- Visualización de resultados (histogramas, diagramas de dispersión, gráficos en 3-D).

- Exportación (informes en HTML o texto).

1.4.4 Librería R

R es un sistema para la realización de cálculos estadísticos. Es un lenguaje que tiene acceso a funciones en el sistema y la habilidad de correr programas guardados en archivos script. Se distribuye gratuitamente bajos los términos de la GNU Licencia Pública General, su desarrollo y distribución son llevados a cabo por varios estadísticos conocidos como el Grupo Nuclear de Desarrollo de R. Este consta de muchas características dentro de las cuales tenemos:

- Que dispone de un amplio almacenamiento y manipulación efectiva de datos.
- Cuenta con operadores para cálculos sobre las variables indexas (*Array*), en particular matrices.
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condiciones, ciclos, funciones recursivas y posibilidad de entradas y salidas.

El código fuente de R está escrito en C y algunas rutinas en Fortran, se encuentra disponible en varias formas fundamentalmente para máquinas Unix y Linux, o como archivos binarios pre-compilados para Windows, Linux (Debian, Mandrake, RedHat, SuSe), Macintosh y Alpha Unix. Posee muchas funciones para el análisis estadístico, cuyos resultados se muestran en la pantalla, y algunos resultados intermedios (como coeficientes de regresión, residuales, etc.) se pueden guardar, exportar a un archivo, o ser utilizados en análisis posteriores.

El lenguaje de R permite al usuario, por ejemplo, programar bucles para analizar conjuntos sucesivos de datos. También es posible combinar un solo programa en diferentes funciones estadísticas para realizar análisis más complejos. (PARADIS 2003)

Una de las características más sobresalientes de R es su enorme flexibilidad. Mientras que programas más clásicos muestran directamente los resultados de un análisis, R guarda estos resultados como un objeto, de tal manera que se puede hacer un análisis sin necesidad de mostrar su resultado inmediatamente. El usuario puede extraer solo aquella parte de los resultados que le interesa. Ejemplo, si corre una serie de 20 regresiones y quiere comparar los coeficientes de regresión, R le puede mostrar únicamente los coeficientes estimados, de esta manera los resultados se muestran en una sola línea, mientras que los programas clásicos le pueden abrir 20 ventanas de resultados. (PARADIS 2003)

R es un lenguaje Orientado a Objetos bajo este término se esconde la simplicidad y la flexibilidad de R. Que es un lenguaje interpretado como Java y no compilado como C++, Fortran, Pascal, etc., lo cual significa que los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir ejecutables. Su sintaxis es muy simple e instructiva, ejemplo una regresión lineal se puede ejecutar con el comando lm (y ˜ x). Para que una función sea ejecutada en R debe estar siempre acompañada de paréntesis, si se escribe el nombre de la función sin los paréntesis, mostrará el contenido (código) mismo de la función.

Todas las acciones en R se realizan con objetos que son guardados en la memoria activa del ordenador, sin usar archivos temporales. (Ver fig. 1). La lectura y escritura de archivo solo se realiza para la entrada y salida de datos, resultados, graficas, etc. (PARADIS 2003)

El usuario ejecuta las funciones con la ayuda de comandos definidos. Los resultados se pueden visualizar directamente en la pantalla, guardar en un objeto o escribir directamente en el disco (particularmente para grafos). Debido a que los mismos resultados son objetos, pueden ser considerados como datos y analizados como tal. Los archivos que contengan datos pueden ser leídos desde el disco duro local o en un servidor remoto a través de la red. (PARADIS 2003)

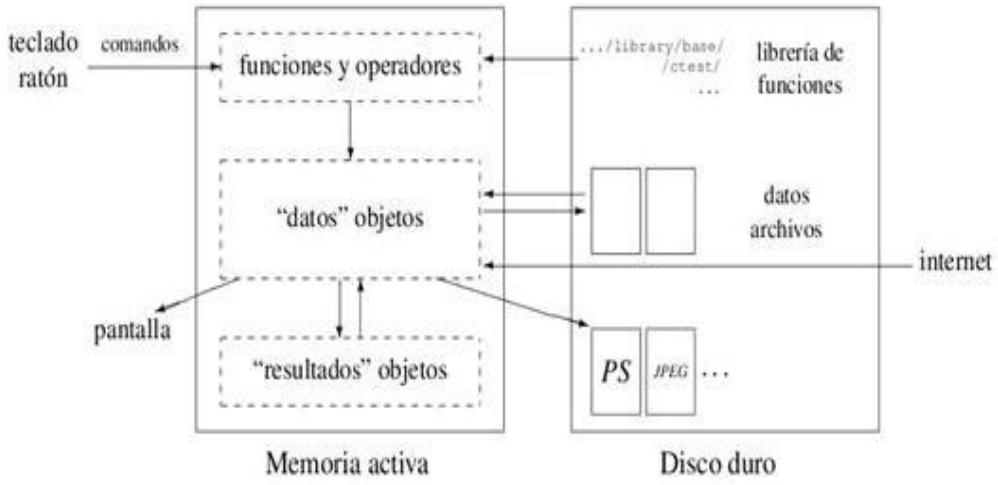


Fig. 1: Visión esquemática del funcionamiento de R.(PARADIS 2003)

Dado que diferentes programas implementan distintos algoritmos para llevar a cabo los mismos tipos de análisis, los usuarios se benefician de una comparación entre los programas más usados. Se han realizado comparaciones de cálculos para los procedimientos (ejemplo: regresión y experimentos factoriales, entre otros) implementados por algunos programas estadísticos. Sin embargo, estos son bastante específicos y se circunscriben a aspectos puntuales. En este contexto, se presenta una

comparación general sobre la base de una serie de aspectos (**Tabla 1**) que brinda toda la información necesaria para escoger la herramienta que se considere más factible a usar.

Programas Estadísticos			
Aspecto	**SPSS**	**SAS**	**R**
Amigabilidad con el usuario	Excelente	Baja-Regular	Baja-Regular
Manipulación de datos	Baja	Buena	Buena
Calidad de gráficos	Regular	Buena-Excelente	Excelente
Control de procesos	Baja	Excelente	Excelente
Costo	U$S 1500	U$S 7200	Gratis
Código fuente disponible	No	No	Si
Variedad de análisis estadístico	Buena	Buena-Excelente	Excelente
Documentación	Excelente	Buena	Buena- Excelente
Soporte técnico	Bueno	Bueno	Bajo
Sistema operativo	Windows	Windows, Macintosh, Linux	Windows, Macintosh, Linux

Tabla 1: Comparación de aspectos generales entre los programas estadísticos SPSS, SAS y R.(SALAS 2008)

1.5 Algoritmos de Regresión.

La regresión tiende a realizar mediciones extremas, por lo que estos algoritmos que no son más que aquellos que predicen una o más variables continuas, como las pérdidas o los beneficios basándose en otros atributos.

El empleo de técnicas de regresión sirve para dos objetivos:

- Estimar la relación entre dos variables teniendo en cuenta la presencia de otros factores.

- Construir un modelo que permita predecir el valor de la variable dependiente para unos valores determinados de un conjunto de variables pronóstico.

Dentro de los algoritmos de regresión que existen se tienen, el lineal y el de polinomial.

1.5.1 Algoritmo de Regresión Lineal Simple

El análisis de regresión se basa principalmente en investigar la relación que existe entre una variable respuesta o dependiente (*Y*) y una variable explicativa o independiente (*X*).El propósito es obtener una función sencilla de la variable explicativa o independiente, que sea capaz de describir lo más ajustado posible la variación de la variable respuesta o dependiente. La función más eficaz es aquella que describe la variable dependiente con la menor diferencia entre los valores observados y predichos. La diferencia entre estos valores observados y predichos se denomina variación residual o residuos. Para evaluar los parámetros en la función se utiliza el ajuste por mínimos cuadrados, en el cual la suma de los cuadrados de las diferencias entre los valores observados y esperados sea menor, lo que es necesario que los residuos o variación residual o errores estén distribuidos normalmente y que varíen de modo similar a lo largo de todo el rango de valores de la variable dependiente. Cuando la variable dependiente es cuantitativa y la relación entre ambas variables sigue una relación recta la función es:

$$\hat{Y}=a+bx$$

Cuando solo existe una variable independiente, esto se reduce a una línea recta. Donde los coeficientes *a* y *b* son parámetros que definen la posición e inclinación de la recta. El parámetro *a*, conocido como la "ordenada en el origen," nos indica cuánto es *Y* cuando *X* = 0. El parámetro *b*, conocido como la "pendiente," nos indica cuánto aumenta *Y* por cada aumento de una unidad en *X*. El problema consiste en obtener estimaciones de estos coeficientes a partir de una muestra de observaciones sobre las variables *Y* y *X.* (COLE 2005)

El procedimiento de regresión lineal matemáticamente.

Anteriormente se ha estudiado detalladamente cómo funciona la librería R y el algoritmo de regresión. En el cual se especifica el análisis de regresión lineal que es una técnica estadística utilizada para estudiar la relación entre variables. En la investigación social, el análisis de regresión se utiliza para predecir un amplio rango de fenómenos, desde medidas económicas hasta diferentes aspecto del comportamiento humano. En el contexto de mercados pueden utilizarse para determinar en cuál de los

diferentes medios de comunicación puede resultar más eficaz invertir o predecir el número de ventas de un determinado producto. Ejemplo:

En una exposición de automóviles, un asistente decidió realizar el conjunto de observaciones relacionado el precio de los vehículos (y) con sus pesos (x), obteniendo los siguientes datos:

X_i= peso en Tm	0.8 1 1.2 1.3 1.5 1.8 2 2.1 2.5
Y_i= precio en millones	1 2 3 5 6 8 9 12 13

Utilizando el método de mínimos cuadrados, calcular el mejor ajuste lineal del tipo y=a+bx. ¿Cuánto podemos esperar que cueste un automóvil de 1.1 Tm? ¿Hasta qué punto es fiable la predicción?

Resultados emitidos:

Xi	Yi	Xi²	Yi²	XiYi	
0.8	1.0	0.64	1	0.8	
1.0	2.0	1	4	2	
1.2	3.0	1.44	9	3.6	
1.3	5.0	1.69	25	6.5	
1.5	6.0	2.25	36	9	
1.8	8.0	3.24	64	14.4	
2.0	9.0	4	81	18	
2.1	12.0	4.91	144	25.2	
2.5	13.0	6.25	169	32.5	
14.2	59	24.92	533	112	Suma

Media

$\bar{X}=1/n\ \Sigma\ Xi$ $Y=1/n\ \bar{\Sigma}\ Yi$

$=14.2/9 =1.58$ $= 59/9 =6.6$

Varianza

$Sxy= 1/n\ \Sigma\ Xi\ Yi - X\bar{Y}$

$= 112/9 – 10.428 = 2.0164$

Covarianza

$S^2x = 1/n\ \Sigma\ Xi^2 - \bar{X^2}$

$= 24.92/9 - 2.4964 = 0.2725$

$S^2y= 1/n\ \Sigma\ Yi^2 - \bar{Y^2}$

$= 533/9 – 43.59 = 15.61$

Recta lineal

$Y = a + bx$

$= -5.306 + 7.3996x$

$a = \bar{Y} - b\bar{X}$ $b = Sxy / S^2x$

$= 6.5 - (7.3996 * 1.57)$ $= 2.0164 / 0.2725$

$= -5.306$ $= 7.3996$

Coeficiente correlación lineal

$r = Sxy / Sx\ Sy$

$= 2.0164/ 0.522 * 3.9509 = 0.98$

Resultados emitidos mediante la herramienta SPSS:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.972	.968	.76109

a. Predictors: (Constant), x

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	142.167	1	142.167	245.432	.000[a]
	Residual	4.055	7	.579		
	Total	146.222	8			

a. Predictors: (Constant), x

b. Dependent Variable: y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5.306	.798		-6.645	.000
	x	7.518	.480	.986	15.666	.000

a. Dependent Variable: y

Fig. 2: Tabla de regresión lineal en SPSS.

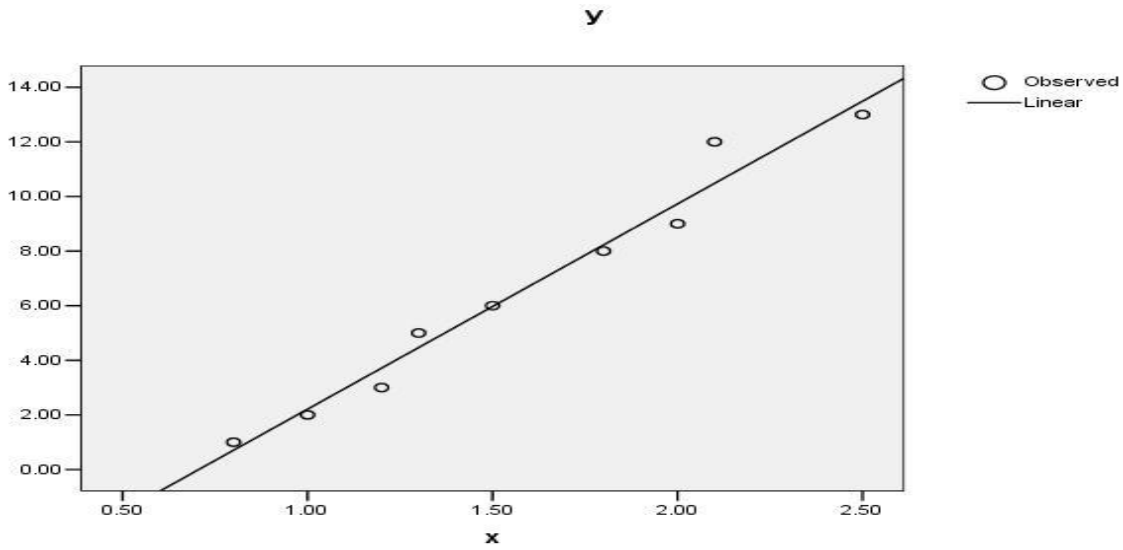


Fig. 3: Gráfica de regresión lineal en SPSS.

Resultados emitidos mediante la herramienta R:

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.7297 -0.4885 -0.2120  0.2915  1.5186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.3057     0.7985  -6.645 0.000292 ***
x             7.5177     0.4799  15.666 1.04e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7611 on 7 degrees of freedom
Multiple R-squared: 0.9723,     Adjusted R-squared: 0.9683
F-statistic: 245.4 on 1 and 7 DF,  p-value: 1.045e-06
```

Fig. 4: Regresión lineal en R.

Como se puede observar mediante el estudio realizado en ambas herramientas si se comparan los resultados que emiten las dos, son los mismos por lo que podemos concluir que el análisis de regresión lineal que se realiza en ambas herramientas es el mismo, aunque la herramienta R proporciona algunos datos adicionales que el SPSS no brinda.

1.5.2 Algoritmo de Regresión Polinomial de Grado 2.

Este algoritmo a diferencia del algoritmo de regresión lineal utiliza una curva para poder expresar la relación que existe entre X e Y. Mediante este modelo se tiene que a medida que X cambia, Y cambia en una cantidad diferente cada vez. La regresión polinomial al igual que un polinomio es una expresión matemática de tipo:

$$Y = \beta o + \beta 1x + \beta 2x^2$$

Donde:

Y: variable dependiente

X: variable independiente

Bo, β1 y β2: son las constantes a estimar.

Este tipo de regresión se utiliza cuando la relación que existe entre las variables dependientes e independientes es no lineal. Algunos fenómenos resultan ser mejor representados por un polinomio y

aunque a veces puede no ser particularmente "natural", o sea aquellos que expresan una relación de causa y efecto entre las variables; suelen ser flexible y tan fácilmente manejable en forma matemática, que resulta de gran utilidad.

El procedimiento de regresión polinomial matemáticamente.

A partir de los datos que muestra la siguiente tabla, ajuste un polinomio de 2do grado utilizando la regresión polinomial.

Xi	Yi
0	2.1
1	7.7
2	13.6
3	27.2
4	40.9
5	61.1

Para este caso específico:

m =2 (el grado del polinomio)

n =6 (la cantidad de datos)

X trazos = 2.5000

Y trazos = 25.4333

- De esta forma el conjunto general de ecuaciones queda instanciado de la siguiente manera:

$$Ao\ n + A1\ \Sigma Xi + A2\ \Sigma Xi^{2} = \Sigma Yi$$

$$Ao\ \Sigma Xi + A1\ \Sigma Xi^{2} + A2\ \Sigma Xi^{3} = \Sigma Xi\ Yi$$

$$Ao\ \Sigma Xi^{2} + A1\ \Sigma Xi^{3} + A2\ \Sigma Xi^{4} = \Sigma Xi^{2}\ Yi$$

Xi	Yi	Xi^2	Xi^3	Xi^4	Xi Yi	Xi^2 Yi	$(Yi- Y trazo)^2$	$(Yi-Ao-A1Xi-A2Xi^2)$
0	2.1	0	0	0	0	0	544.4444	0.14332
1	7.7	1	1	1	7.7	7.7	314.4711	1.00286
2	13.6	4	8	16	27.2	54.4	140.0278	1.08158
3	27.2	9	27	81	81.6	244.8	3.1211	0.80491
4	40.9	16	64	256	163.6	654.4	239.2178	0.61951
5	61.1	25	125	625	305.5	1527.5	1272.1111	0.09439
ΣXi	**ΣYi**	**ΣXi^2**	**ΣXi^3**	**ΣXi^4**	**ΣXi Yi**	**ΣXi^2 Yi**	**St**	**Sr**
15	152.6	55	225	979	585.6	2488.8	2513.3933	3.74657

- Por lo tanto las ecuaciones lineales simultáneas son:

$$6\ Ao + 15\ A1 + 55\ A2 = 152.6$$

$$15\ Ao + 55\ A1 + 225\ A2 = 585.6$$

$$55\ Ao + 225\ A1 + 979\ A2 = 2488.8$$

- Resolviendo este sistema con alguna técnica como la eliminación gausiana se obtiene:

  Ao = 2.47857 A1 = 2.35929 A2 = 1.86071

- El polinomio es:

  $$1.86071\ X^2 + 2.35929\ X + 2.47857$$

- Debemos calcular Sr y St.

Sr: Se utiliza para calcular el error estándar de aproximación basado en la regresión polinomial.

St: Se utiliza para calcular el coeficiente de determinación.

$St = \Sigma (Yi - Y\ trazos)^2$ $Sr = \Sigma (Yi - Ao - A1\ Xi - A2\ Xi^2)$

Error estándar de aproximación

$Sy/x = \sqrt{Sr / n-(m+1)}$ $Sy = \sqrt{St / n-1}$

$= \sqrt{3.74657 / 6-3}$ $= \sqrt{2513.3933 / 6-1}$

$= 1.1175$ $= 22.4205$

Coeficiente de determinación

$r^2 = St - Sr / n - 1$

$= 2513.3933 - 3.74657 / 2513.3933$

$= 0.99851$

- El resultado indica que el 99.851 % de la incertidumbre original se ha explicado mediante el modelo.

Resultados emitidos mediante la herramienta SPSS:

Model Summary and Parameter Estimates

Dependent Variable: y

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	.999	1004.777	2	3	.000	2.479	2.359	1.861

The independent variable is x.

Fig. 5: Gráfica regresión polinomial en SPSS.

Resultados emitidos mediante la herramienta R:

```
Residuals:
      1       2       3       4       5       6
-0.3786  1.0014 -1.0400  0.8971 -0.7871  0.3071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4786     1.0128   2.447  0.09191 .
x             2.3593     0.9527   2.476  0.08955 .
I(x^2)        1.8607     0.1829  10.174  0.00202 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.118 on 3 degrees of freedom
Multiple R-squared: 0.9985,     Adjusted R-squared: 0.9975
F-statistic:  1005 on 2 and 3 DF,  p-value: 5.755e-05
```

Fig. 6: Regresión polinomial en R.

En el estudio realizado con el algoritmo de regresión polinomial se emitieron los mismos resultados en las dos herramientas utilizadas demostrando de esta forma, que el análisis estadístico se realiza de igual manera con ambas herramientas, aunque R proporciona algunos datos adicionales que el SPSS no brinda.

1.6 Herramientas Case.

Las herramientas CASE según sus siglas, es Ingeniería de Software Asistida por Computación, es la aplicación de métodos y técnicas a través de las cuales se hacen útiles a las personas entender las capacidades de las computadoras, por medio de programas, de procedimientos y su documentación. Las herramientas CASE permiten modelar los Procesos de Negocios de las empresas y desarrollar los Sistemas de Información Gerenciales. Permiten organizar y manejar la información de un proyecto informático, permitiéndoles a los participantes de un proyecto que los sistemas, se tornen más comprensibles y además mejorar la comunicación entre los participantes.

Algunos de los componentes de las herramientas CASE permiten:

- Confeccionar la definición de requerimientos de los usuarios.
- Mejorar el diseño de los sistemas.
- Mejorar la eficiencia en la programación (por su generación automática de códigos).
- Otorgar a la administración un mejor soporte en la documentación.

Para ello, y sin importar la arquitectura de la herramienta CASE, en general tales herramientas deben abarcar las siguientes propiedades:

- Tener una interfaz gráfica y textual, que le permita al usuario manejar los objetos de diseño.
- Contar con un Diccionario de Datos, a fin de rastrear y controlar los objetos diseñados.
- Disponer de un conjunto de herramientas que permitan: chequear las reglas del diseño y analizar la lógica del diseño.

Las herramientas CASE son un elemento muy importante, que le permitirá al administrador llevar adelante un proyecto informático de forma eficaz y eficiente. También estas mismas herramientas, como toda Tecnología de la Información se encuentra en continua evolución y existe además una gran variedad de proveedores y productos y cada uno de ellos con sus diferentes aplicaciones y especificaciones.

Otro elemento importante es que las herramientas CASE, son herramientas que permiten aumentar la productividad en el desarrollo de un proyecto y estas deben ser aplicadas a una metodología determinada.

Las herramientas CASE en sí mismas son una metodología; su uso está restringido a la metodología elegida para llevar adelante el análisis y diseño del proyecto. Después del estudio realizado sobre estas herramientas se decide que como herramienta case a usar:

1.6.1 Visual Paradigm

Visual Paradigm para es una herramienta CASE que admite el diseño de software utilizando notación UML, soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una más vertiginosa construcción de aplicaciones de calidad y a un menor costo. Admite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. La herramienta UML CASE también facilita cuantiosos tutoriales de UML, demostraciones interactivas de UML y proyectos UML. Y además de estas características posee muchas más, igual de significativas, algunas de estas son las que se muestran a continuación:

- Soporte de UML versión 2.1.

- Diagramas de Procesos de Negocio - Proceso, Decisión, Actor de negocio, Documento Modelado colaborativo con CVS y Subversión.
- Ingeniería de ida y vuelta.
- Ingeniería inversa - Código a modelo, código a diagrama.
- Diagramas de flujo de datos.
- Generación de bases de datos - Transformación de diagramas de Entidad-Relación en tablas de base de datos.
- Ingeniería inversa de bases de datos - Desde Sistemas Gestores de Bases de Datos (DBMS) existentes a diagramas de Entidad-Relación.
- Distribución automática de diagramas - Reorganización de las figuras y conectores de los diagramas UML.
- Importación y exportación de ficheros XMI.
- Integración con Visio - Dibujo de diagramas UML con plantillas de MS Visio.

1.7 Lenguajes y Tecnologías

El lenguaje de programación es un idioma diseñado para que las máquinas como la computadora la entienda de manera que puedan crear programas que controlen el comportamiento físico y lógico de la máquina, para expresar algoritmos de precisión y como modo de comunicación humana. El lenguaje de programación está compuesto por un conjunto de símbolos y reglas semánticas y sintácticas que definen las estructuras y significados de sus expresiones.

1.7.1 PHP

PHP (Hipertext Preprocessor) es un lenguaje de programación que está diseñado básicamente para la creación de páginas web, fue creado originalmente por Rasmus Lerdorf en 1994. Es un lenguaje interpretado con un uso amplio y puede ser incrustado dentro de código HTML, es desplegado en la mayoría de los servidores web y en casi todos los sistemas operativos y plataformas sin costo alguno. PHP se encuentra instalado en más de 20 millones de sitios web y en un millón de servidores, su versión más reciente es la 5.3.2 del 04 de marzo del 2010. Posee una capacidad de conexión con la

mayoría de los motores de base de datos que se utilizan en la actualidad, destacando su conectividad con MySQL y PostgreSQL, al ser libre se presenta como una alternativa de fácil acceso para todos , permite aplicar técnicas de programación orientada a objetos y tiene una biblioteca natural de funciones intensamente amplia e incluida.(GRACIA 2004)

Ventajas:

- Fácil de aprender.
- Se caracteriza por ser un lenguaje rápido.
- Soporta en cierta medida la orientación a objeto. Clases y herencia.
- Es un lenguaje multiplataforma: Linux, Windows, entre otros.
- Capacidad de conexión con la mayoría de los manejadores de bases de datos: MySQL, PostgreSQL, Oracle, MS SQL Server, entre otras.
- Capacidad de expandir su potencial utilizando módulos.
- Posee documentación en su página oficial la cual incluye descripción y ejemplos de cada una de sus funciones, además de una gran comunidad de desarrolladores.
- Es libre, por lo que se presenta como una alternativa de fácil acceso para todos.
- Incluye gran cantidad de funciones.
- No requiere definición de tipos de variables.

1.8 IDE de desarrollo

Los IDE de desarrollo, según sus siglas (Entorno de Desarrollo Integrado) es un programa informático compuesto por un conjunto de herramientas de programación. Un IDE es un medio de programación que ha sido empaquetado como un programa de aplicación, como un compilador, editor de código y constructor de interfaces gráficas. También suministran un marco de trabajo amigable para la mayoría de los lenguajes de programación. Existen varios tipos de IDE que utilizan las empresas para su desenvolvimiento.

1.8.1 NetBeans

NetBeans comenzó como un proyecto estudiantil en República Checa (originalmente llamado Xelfi).

El NetBeans IDE es un ambiente libre de desarrollo integrado con Open Source para desarrolladores de software. El mismo ofrece todas las herramientas necesarias para crear escritorios profesionales, Enterprise, Web y aplicaciones móviles con el lenguaje Java, JavaFX, C / C + + y lenguajes dinámicos como PHP, Java Script, Groovy y Ruby. El NetBeans IDE es de fácil instalación y uso directamente desde la caja y se ejecuta en Windows, Linux, Mac OS X y Solaris (TM).

Su nueva versión NetBeans IDE 6.8 brinda soporte completo para Java EE 6 y Sun GlassFish, Enterprise Server v3 y ofrece PHP mejorado, soporte para JavaFX y C/C ++ además ofrece otras nuevas características y mejoras que incluyen:

Una integración más ajustada con Project Kenai.

Mejora de C / C + + Profiling: Perfila y sintoniza aplicaciones C / C + + con el nuevo indicador Microstate Accounting, supervisor de uso I/O.

JavaFX: Código de finalización mejorado, sugerencias y navegación para JavaFX en el editor NetBeans.

1.9 Metodología para realizar el análisis en la Línea de Desarrollo de Herramientas de Análisis de Datos.

La metodología de trabajo que propone el grupo de análisis de la línea de desarrollo de herramienta para el análisis de datos, soporta el modelo de desarrollo basado en líneas de producto de software. Para lograr el desarrollo de esta metodología se tuvo en cuenta las oportunidades de mejoras para la línea, pues anteriormente existían algunos problemas con las metodologías aplicadas como Open Up ó RUP. Además se alinean los artefactos que genera dicho grupo con los previstos por la Universidad de Ciencias Informáticas. Esta metodología está siendo aplicada en la Línea Integral de Soluciones del Centro de Tecnología de Gestión de Datos (DATEC), ha organizado la producción mediante el concepto de líneas de productos, estableciéndose para ello tres líneas principales:

En la Línea Integral de Soluciones se han definido 3 grupos para el desarrollo de todos sus proyectos, el grupo de análisis, el grupo de arquitectura y el grupo de desarrollo. La metodología aplicada en esta línea está basada en Open Up, Scrum y PMBok, esto ha dado muy bueno resultados al centro pues en

poco tiempo y con pocos recursos humanos se ha logrado tener un gran avance y reutilización del código y la documentación.(ALMAGUER 2009)

La puesta en marcha del modelo de desarrollo basado en líneas de producto de software traía como problema que la aplicación de algunas metodologías clásicas como RUP, Open Up entre otras, además la universidad tiene un expediente de proyecto que de alguna forma también guía el desarrollo de los proyectos informáticos y el volumen de información que se repetía en los proyectos de la misma línea era grande y difícil de gestionar lo que hacía el proceso de desarrollo y liberación de un producto un poco engorroso. Por lo que se decidió realizar una metodología de trabajo que estuviera alineada con el modelo de desarrollo basado en líneas de producto de software y con el modelo de expediente de proyecto exigido por la universidad para el desarrollo de los proyectos informáticos.(ALMAGUER 2009)

EL grupo de análisis debe ser capaz de realizar el análisis de diferentes proyectos en un mismo tiempo además debe tener la flexibilidad de poder cambiar los recursos humanos de proyecto en el momento que sea necesario por lo que es importante definir una estructura flexible y bien organizada, es importante recordar que un importante volumen de los recursos humanos con que cuenta el grupo son estudiantes de la universidad esto puede traer algunos problemas como la falta de experiencia y preparación desarrollando actividades de este tipo, sin contar el tiempo en que estos recursos le pueden dedicar a la producción por lo que se decide que el grupo tenga la siguiente estructura:

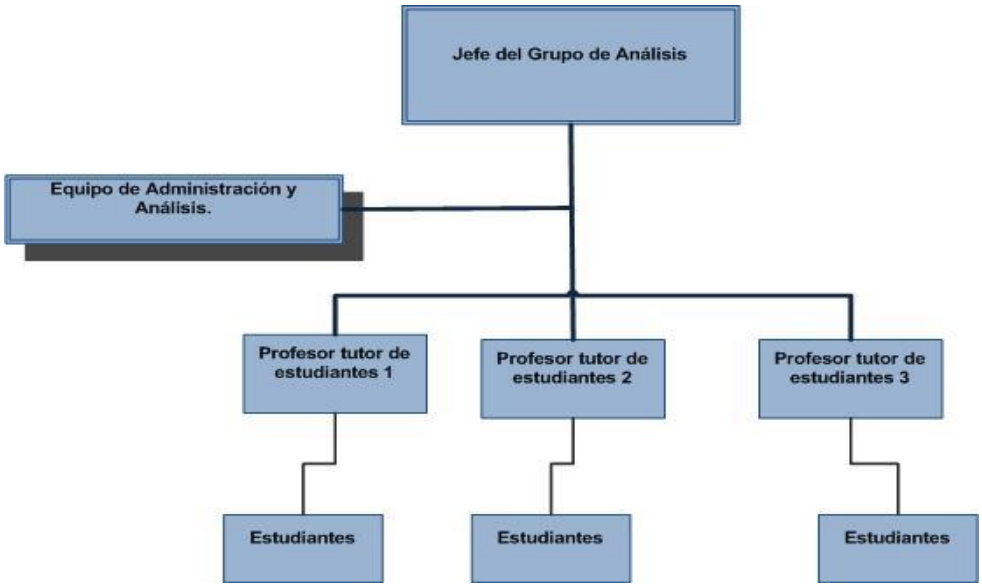


Fig. 7: Estructura del grupo de análisis.(ALMAGUER 2009)

Los artefactos que el grupo de análisis debe generar para cada proyecto debe estár lineado a los artefactos del expediente de proyecto que define la universidad,en la figura 10 se muestran los artefactos que se realizan en la diferentes faces de los proyectos .(ALMAGUER 2009)

:

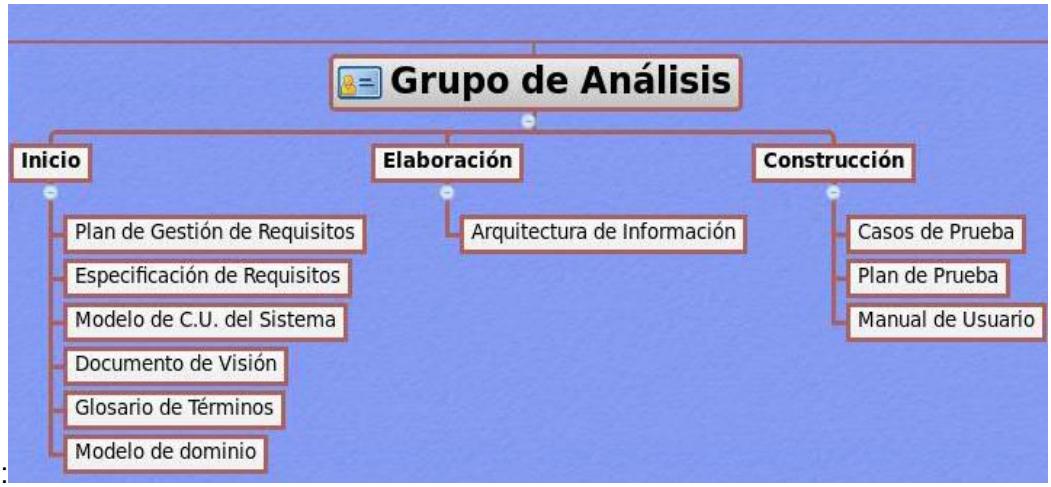


Fig. 8: Artefactos generados por el grupo de análisis.(ALMAGUER 2009)

1.10 Conclusiones parciales

Después del estudio realizado de los diferentes aspectos tratados en este capítulo se describieron las herramientas y lenguaje a utilizar así como la metodología del desarrollo del software para lograr una propuesta de solución.

Se escoge la librería R debido a que no es necesario utilizar un menú de funciones para su ejecución, cuenta con un excelente control de procesos y logra así, que sea más flexible por ser de código abierto y posee una muy buena documentación. Es gratis, tiene un código fuente que se encuentra disponible y es multiplataforma ya que corre en diferentes sistemas operativos. Se desarrollan algoritmos de regresión como el lineal simple y polinomial (Curvilíneo), como herramienta Case el Visual Paradigm para la modelación visual ya que está definida como política del centro. Y como metodología a utilizar, es la definida por el grupo de análisis en la Línea Integral de Soluciones, logrando tener un gran avance en el centro con muy buenos resultados.

Análisis y diseño de la solución propuesta

2.1 Introducción

En este capítulo se comienza el estudio del dominio actual para entenderlo y lograr definir las mejoras que se pueden realizar. Se elabora una lista con los requerimientos funcionales y no funcionales, se identifican los casos de usos y se elabora un diagrama para representar la interacción de los actores con los casos de usos con el objetivo de proporcionarle un mayor entendimiento al cliente. Además se explica la arquitectura que se va a implementar y los patrones de diseño aplicados.

2.2 Análisis

2.2.1 Modelo de dominio

Un Modelo de Dominio es un artefacto de la disciplina de análisis, construido con las reglas de UML durante la fase de concepción, en la tarea construcción del modelo de dominio, presentado como uno o más diagramas de clases y que contiene, no conceptos propios de un sistema de software sino de la propia realidad física.

Los modelos de dominio pueden utilizarse para capturar y expresar el entendimiento ganado en un área bajo análisis como paso previo al diseño de un sistema, ya sea de software o de otro tipo. Similares a los mapas mentales utilizados en el aprendizaje, el modelo de dominio es utilizado por el analista como un medio para comprender el sector industrial o de negocios al cual el sistema va a servir.(JACOBSON 2000; PRESSMAN 2002)

Diagrama de modelo de dominio

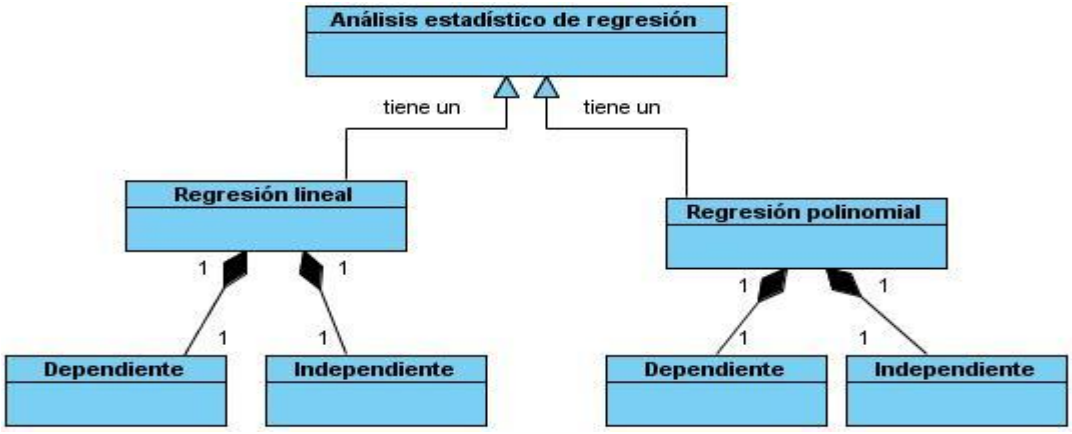


Fig. 9: Diagrama del Modelo del dominio.

Definición de clases del dominio

Modelo estadístico: Representa un módulo que brinda la opción de realizar análisis estadísticos a través de diferentes conceptos.

Regresión lineal: Es un método matemático que brinda un modelo de la relación entre una variable dependiente y una variable independiente.

Regresión polinomial: Es un método matemático que nos permite estimar los parámetros de cualquiera de los modelos estadísticos estándar, guardar los valores pronosticados (valores teóricos) y dibujar la curva de todos los modelos.

Variables Independientes: Es dato de entrada obligatorio, pues es una de las variables necesarias para la realización del análisis estadístico de regresión lineal

Variables Dependientes: Es dato de entrada obligatorio, pues es una de las variables necesarias para la realización del análisis estadístico de regresión lineal

2.2.2 Especificación de los requisitos

La Especificación de Requisitos Software (ERS) es una descripción completa del comportamiento del sistema que se va a desarrollar. Incluye un conjunto de casos de uso que describe todas las interacciones que tendrán los usuarios con el software. Los casos de uso también son conocidos como requisitos funcionales. Además de los casos de uso, también contiene requisitos no funcionales (o complementarios), los cuales imponen restricciones en el diseño o la implementación. (JACOBSON 2000; PRESSMAN 2002)

2.2.3 Requisitos Funcionales

Definen el comportamiento del sistema describiendo las tareas que este debe realizar. Al definir un requisito funcional es importante mantener el equilibrio entre la excesiva generalidad, insuficiencia de detalle o ambigüedad, y el exceso de detalle con precisiones o descripciones innecesarias o redundantes.

Requisitos del componente propuesto

Para realizar el análisis estadístico, a partir del algoritmo de Regresión Lineal definido en el componente propuesto como solución, se definen los siguientes requisitos:

RF1: Permitir insertar los datos de la variable dependiente (los datos son obligatorios y con la misma cantidad de la variable independiente).

RF2: Permitir insertar los datos de la variable independiente (los datos son obligatorios y con la misma cantidad de la variable dependiente).

RF3: Mostrar Salida de regresión lineal.

- Mostrar los valores de la variable independiente de la tabla anova.
- Mostrar los valores residuales de la tabla anova.
- Mostrar los valores de las variables intersect e independiente de la tabla de coeficientes.

RF4: Mostrar Salida regresión polinomial.

- Mostrar los valores coeficientes de la variable intersect, independiente e independiente de potencia dos.
- Mostrar los grados de libertad, el residuo cuadrado corregido múltiple y el valor probabilístico.

2.3.4 Requisitos no funcionales

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener. Estas propiedades son las características que conciben que un producto sea atractivo, usable, rápido o confiable. Usualmente están afines a los requerimientos funcionales, una vez que se conozca lo que el sistema debe hacer podemos determinar cómo ha de comportarse, qué cualidades debe tener o cuán rápido o grande debe ser.

RNF1: De hardware

RNF1.1 Para el servidor se debe tener como mínimo una PC de RAM: 2GB,con un procesador Pentium IV y como mínimo 1GB espacio libre en el disco duro.

RNF3: Restricciones de diseño

RNF3.1 El lenguaje de programación a utilizar es PHP.

RNF3.2 Se define la notación JSON para propiciar las peticiones hacia el servidor de análisis (R server).

RNF3.3 La respuesta del sistema será mediante arreglo asociativo, se implementará para retornar los resultados de los análisis estadísticos.

RNF3.4 Se utilizará la notación JSON para proporcionar las respuestas al usuario desde el servidor de análisis.

RNF7: De fiabilidad.

RNF7.1 El sistema debe permanecer online en los horarios de trabajo establecidos por las entidades, excepto cuando sea necesario reiniciarlo o detenerlo por mantenimiento.

RNF7.2 Disponibilidad de Servicio.

RNF7.2.1 El sistema permanecerá fuera de servicios solo si está en configuración o mantenimiento.

2.2.5 Modelo de casos de uso del sistema

Los casos de uso son artefactos narrativos que describen, bajo la forma de acciones y reacciones, el comportamiento del sistema desde el punto de vista del usuario. Por lo tanto, establece un acuerdo entre clientes y desarrolladores sobre las condiciones y posibilidades (requisitos) que debe cumplir el sistema. Los casos de uso candidatos también se encuentran entre las actividades a automatizar. Esto no significa que una actividad se convierta en un caso de uso, porque un caso de uso es un proceso que da un resultado de valor para un actor determinado y una secuencia de actividades a automatizar puede implicar pasos dentro de un caso de uso. (JACOBSON 2000; PRESSMAN 2002)

Actores del sistema

Actor	Descripción
Analizador	Responsable de construir la URL con la escritura del JSON contenedor del tipo de análisis estadístico. El actor es de preferencia un sistema externo que se comunica mediante un protocolo HTTP.
R	Es el responsable de brindarle una respuesta al R server mediante un JSON. El actor de preferencia un sistema externo, simplemente recibe el fichero con las instrucciones del tipo de análisis estadístico, ejecuta las acciones y genera el resultado mediante un fichero.

Tabla 2: Actores del sistema.

Diagrama de casos de uso del sistema

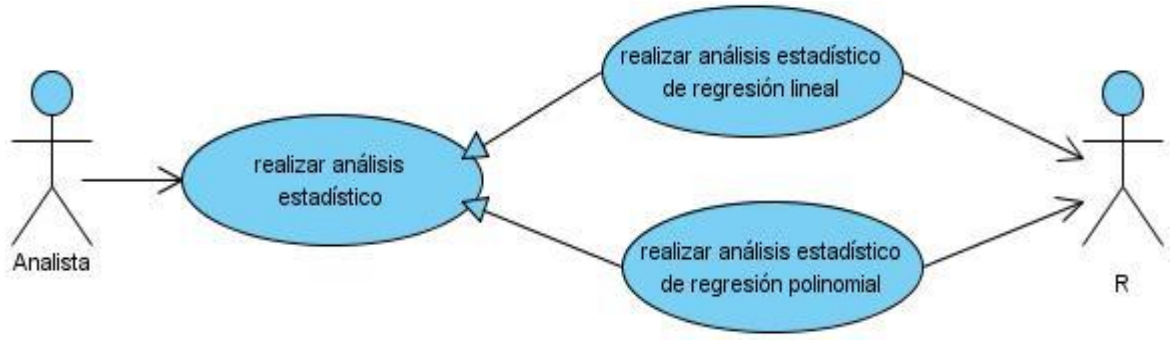


Fig. 10: Diagrama de casos de uso del sistema.

Especificación de casos de uso

Descripción de Caso de uso generalizado: “Realizar análisis estadístico”.

<table>
<tr><td>Caso de Uso:</td><td colspan="2">Realizar análisis estadístico</td></tr>
<tr><td>Actores:</td><td colspan="2">Analizador</td></tr>
<tr><td>Resumen:</td><td colspan="2">Representación del concepto implícito de la realización de análisis de estadístico.</td></tr>
<tr><td>Precondiciones:</td><td colspan="2">1. Es obligatorio el envío del tipo de análisis estadístico solicitado

2. Es obligatorio el envío de la escritura que contiene las configuraciones del análisis estadístico.</td></tr>
<tr><td>Referencias</td><td colspan="2">RF1</td></tr>
<tr><td>Prioridad</td><td colspan="2"></td></tr>
<tr><td colspan="3">Flujo Normal de Eventos</td></tr>
<tr><td>Acción del Actor</td><td colspan="2">Respuesta del Sistema</td></tr>
<tr><td>1. El analizador solicita la realización de un análisis estadístico entrando todos los datos obligatorios y acciones para realizar la misma.</td><td colspan="2">2. El sistema recibe la petición.

3. El sistema identifica el tipo de análisis estadístico solicitado, en caso de solicitar una no implementada, ver Flujos Alternos 3.1.

4. Según el tipo análisis estadístico identificado se invoca la realización del caso de uso especializado correspondiente para el caso de: <<regresión lineal>> ver caso de uso especializado “Realizar análisis estadístico del algoritmo de regresión lineal”, en caso que sea <<regresión polinomial>> ver caso de uso especializado “Realizar análisis</td></tr>
</table>

	estadístico del algoritmo de regresión polinomial". 5. Envía al actor los datos de la respuesta del análisis estadístico. 6. Termina el caso de uso
Prototipo de Interfaz	
Flujos Alternos	
Acción del Actor	**Respuesta del Sistema**
	3.1. Se notifica al actor que la petición realizada no es correcta.
Prototipo de Interfaz	
Poscondiciones	Se envió al actor la respuesta del análisis estadístico solicitado.

Tabla 3: Descripción de Caso de uso generalizado.

Descripción de Caso de uso especializado "Realizar análisis de Regresión Lineal

Caso de Uso:	Realizar análisis estadístico del algoritmo de regresión lineal
Actores:	R
Resumen:	Responsable de realizar el análisis estadístico de regresión lineal «regression_linear». La entrada para la ejecución del análisis es suministrada en una escritura estandarizada que contiene los datos, las cuales serán ejecutadas en la interfaz de línea de comando de la aplicación R. La aplicación R es la responsable de realizar el análisis estadístico y retornará como respuesta un archivo físico en el nodo servidor con el resultado. El resultado será interpretado y se construirá una respuesta para el actor en el mismo formato que la entrada.
Precondiciones:	1. La escritura que contiene la solicitud enviada para la realización de análisis estadístico no debe obviar los datos obligatorios: variable dependiente y variable independiente.

<table>
<tr><td>Referencias</td><td colspan="2">RF1, RF2</td></tr>
<tr><td>Prioridad</td><td colspan="2"></td></tr>
<tr><td colspan="3">Flujo Normal de Eventos</td></tr>
<tr><td colspan="2">Acción del Actor</td><td>Respuesta del Sistema</td></tr>
<tr><td colspan="2">5. La aplicación R realiza el análisis estadístico de regresión lineal y se devuelve como resultado un fichero con la respuesta en una dirección física de la máquina servidora.</td><td>1. El sistema valida que se hayan enviado los datos obligatorios: variable dependiente y variable independiente; en caso de no haberse enviado los datos, ver Flujo Alterno 1.1.
2. Se capturan los datos obligatorios para construir el objeto “regression_linear” que los contendrá.
3. Se procede a construir el archivo con las instrucciones en lenguaje R para ejecutarlo en la aplicación R.
4. Se ejecutan las instrucciones conformadas en la aplicación R.

6. Se interpreta el fichero de salida de la aplicación R y se construye la respuesta en una escritura estandarizada lista para notificar al actor.
7. Continúa el Flujo Básico del Caso de uso generalizado: “Realizar análisis estadístico” en el paso 5.</td></tr>
<tr><td colspan="3">Prototipo de Interfaz</td></tr>
<tr><td colspan="3">Flujos Alternos</td></tr>
<tr><td colspan="2">Acción del Actor</td><td>Respuesta del Sistema</td></tr>
</table>

	1.1 En caso de no haberse enviado los datos obligatorios, se notifica un error y pasa directamente al paso 6 del Flujo Básico.
Prototipo de Interfaz	
Poscondiciones	Se construyó la respuesta estandarizada del análisis de estadístico de regresión lineal.

Tabla 4: Descripción de Caso de uso especializado.

Descripción de Caso de uso especializado "Realizar análisis de Regresión Polinomial.

Caso de Uso:	Realizar análisis estadístico del algoritmo de regresión polinomial.
Actores:	R
Resumen:	Responsable de realizar el análisis estadístico de regresión polinomial «polinomial». La entrada para la ejecución del análisis es suministrada en una escritura estandarizada que contiene los datos, las cuales serán ejecutadas en la interfaz de línea de comando de la aplicación R. La aplicación R es la responsable de realizar el análisis estadístico y retornará como respuesta un archivo físico en el nodo servidor con el resultado. El resultado será interpretado y se construirá una respuesta para el actor en el mismo formato que la entrada.
Precondiciones:	1. La escritura que contiene la solicitud enviada para la realización de análisis estadístico no debe obviar los datos obligatorios: variable dependiente y variables independientes.
Referencias	RF1, RF2
Prioridad	
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
	1. El sistema valida que se hayan enviado los datos obligatorios: variable dependiente

	y variables independientes; en caso de no haberse enviado los datos, ver Flujo Alterno 1.1. 2. Se capturan los datos obligatorios para construir el objeto “polinomial” que los contendrá. 3. Se procede a construir el archivo con las instrucciones en lenguaje R para ejecutarlo en la aplicación R. 4. Se ejecutan las instrucciones conformadas en la aplicación R.
5. La aplicación R realiza el análisis estadístico de regresión polinomial y se devuelve como resultado un fichero con la respuesta en una dirección física de la máquina servidora.	6. Se interpreta el fichero de salida de la aplicación R y se construye la respuesta en una escritura estandarizada lista para notificar al actor. 7. Continúa el Flujo Básico del Caso de uso generalizado: “Realizar análisis estadístico” en el paso 5.
Prototipo de Interfaz	
Flujos Alternos	
Acción del Actor	**Respuesta del Sistema**
	1.1 En caso de no haberse enviado los datos obligatorios, se notifica un error y pasa directamente al paso 6 del Flujo Básico.
Prototipo de Interfaz	
Poscondiciones	Se construyó la respuesta estandarizada del análisis de estadístico Regresión polinomial.

Tabla 5: Descripción de Caso de uso especializado.

2.3 Diseño

2.3.1 Descripción de la arquitectura del sistema

La arquitectura del software es el diseño de más alto nivel de la estructura de un sistema. También es conocido como arquitectura lógica, que no es más que un conjunto de patrones y abstracciones coherentes que proporcionan la referencia necesaria para guiar la construcción de un software. Esta establece los fundamentos para que analistas, desarrolladores y programadores trabajen en una línea común que permita alcanzar los objetivos esperados cubriendo de esta forma todas las necesidades. (JACOBSON 2000; PRESSMAN 2002)

PATDSI es una aplicación que corre bajo tecnología web y se conecta a un gestor de base de datos. Esta aplicación cuenta con un módulo de generador de reportes, que es el encargado de generar reportes, que pueden ser tabuladores o análisis estadísticos.

El análisis estadístico tiene la opción de elegir el tipo de análisis estadístico que se quiere realizar, el cual es el encargado de confeccionar una escritura JSON que es enviada desde PATDSI al R-Server por el método POST (ver fig.8).

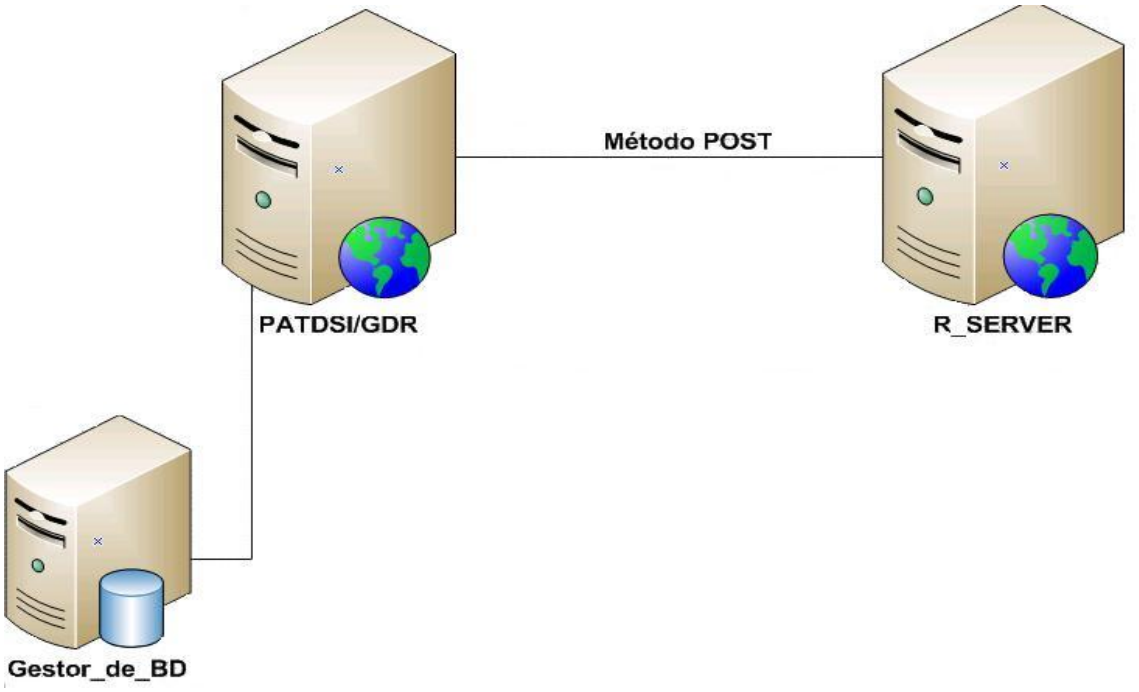


Fig. 11: Arquitectura del sistema (Integración de PTDSI con el R_Server).

El Servidor de Análisis Estadístico está compuesto por cuatro capas, las cuales son:

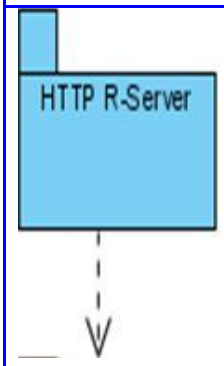
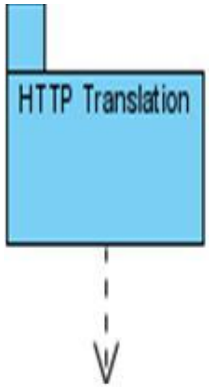
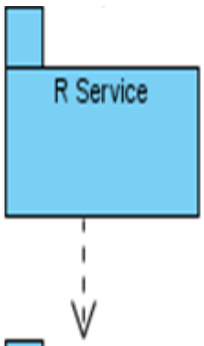
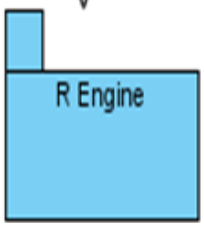
Capa	Responsabilidad
	**HTTP R-Server** Es la capa más externa, se encarga del servicio completo de análisis. Comprende el establecimiento de las conexiones HTTP y el lanzamiento de hilos y procesos para ejecutar las peticiones de forma concurrente.
	**HTTP Translation** Recibe un JSON con los datos asociados al tipo de análisis a realizar, interpreta el mismo y genera los objetos del negocio correspondientes para conformar en la capa inferior la escritura R. Una vez obtenido el fichero con el resultado generado por R, es interpretado el mismo y generado un JSON que será enviado como respuesta a la petición inicial de ejecución de análisis.
	**R Service** Esta es una capa de servicio de alto nivel. Es la responsable de conformar el fichero con la escritura R a ejecutar en R y retornará el fichero de respuesta generado por el propio R.
	**R Engine** La más primitiva de las capas. Simplemente recibe el fichero con las instrucciones del tipo de análisis estadístico, ejecuta las acciones y genera el resultado mediante un fichero.

Tabla 6: Descripción de la arquitectura del sistema.

2.3.2 Patrones

Patrones arquitectónicos

Según Bruschmann se define como patrón arquitectónico, sobre los aspectos fundamentales de la estructura de un sistema de software. Especifican un conjunto predefinido de subsistemas con sus responsabilidades y una serie de recomendaciones para organizar los distintos componentes.

Patrón en Capa:

El patrón en capa según lo planteado por Rivera y Robacio, descompone una aplicación en un conjunto de capas independientes y ordenadas jerárquicamente según el nivel de abstracción que tenga cada una de ellas. Cada nivel o capa usa los servicios del nivel inmediatamente inferior y ofrece servicios a la capa superior.

Ventajas

- El estilo resiste un diseño basado en niveles de abstracción progresivo.
- El estilo permite optimizaciones y refinamientos.
- Facilita una amplia reutilización.

Desventajas

- Muchos problemas no aceptan un buen mapeo en una estructura jerárquica.
- Los cambios en las capas de bajo nivel tienden a filtrarse hacia las de alto nivel.
- A veces es difícil encontrar el nivel de abstracción correcto.

Siendo este patrón, el seleccionado para guiar la arquitectura del sistema del Servidor de Análisis Estadístico porque se enfoca en la distribución de roles y responsabilidades de forma jerárquica, suministrando de una forma muy efectiva la separación de responsabilidades.

Patrones de diseño

Los patrones de diseño son la principal base para la búsqueda de soluciones a problemas comunes en el desarrollo de software y otros ámbitos referentes al diseño de interacción o interfaces. Son una solución a un problema de diseño, y para que esa solución sea considerada un patrón debe poseer características como la efectividad y debe ser reusable lo que significa que es aplicable a diferentes problemas de diseño en distintas circunstancias. (JACOBSON 2000; PRESSMAN 2002)

Estos patrones también pueden considerarse como un documento que define una estructura de clases que aborda una situación particular, se dividen en tres grupos principales dentro de los cuales tenemos los patrones de creación, los funcionales y los estructurales donde se encuentra el patrón Facade que es el se escoge como patrón a utilizar.

Patrones GoF

Los patrones GoF Gang of Four o también llamado como la pandilla de los cuatros se adaptan a los problemas generales del diseño al describir como se comunican las clases y objetos entre sí. Este facilita el aprendizaje y la comunicación entre los programadores y los diseñadores. (JACOBSON 2000; PRESSMAN 2002)

Se clasifica en dependencia de la solución como son:

- Los creacionales se encargan de abstraer el proceso de instanciación y ocultar los detalles de cómo los objetos son creados o inicializados, entre los que se encuentran Singleton, Abstract Factory y Factory Method.
- Los de comportamiento nos ayudan a definir la comunicación e interacción entre los objetos de un sistema, entre los que se encuentran Observer y Strategy.
- Los estructurales describen como las clases y objetos pueden ser combinados para formar grandes estructuras y proporcionar nuevas funcionalidades, dentro de este se encuentran Proxy, Decorator, Composite y Facade.

Facade (Fachada): Le proporciona al R-SERVER una interfaz de alto nivel que hace que el subsistema sea más fácil de usar. También hace más fácil el uso y el entendimiento de una biblioteca

del software ya que implementa métodos ventajosos para tareas frecuentes. Este patrón se aplica en las capas de rservices, para abstraer la complejidad de esta capa a través de la clase r_services.

2.3.3 Diagrama de clases del diseño

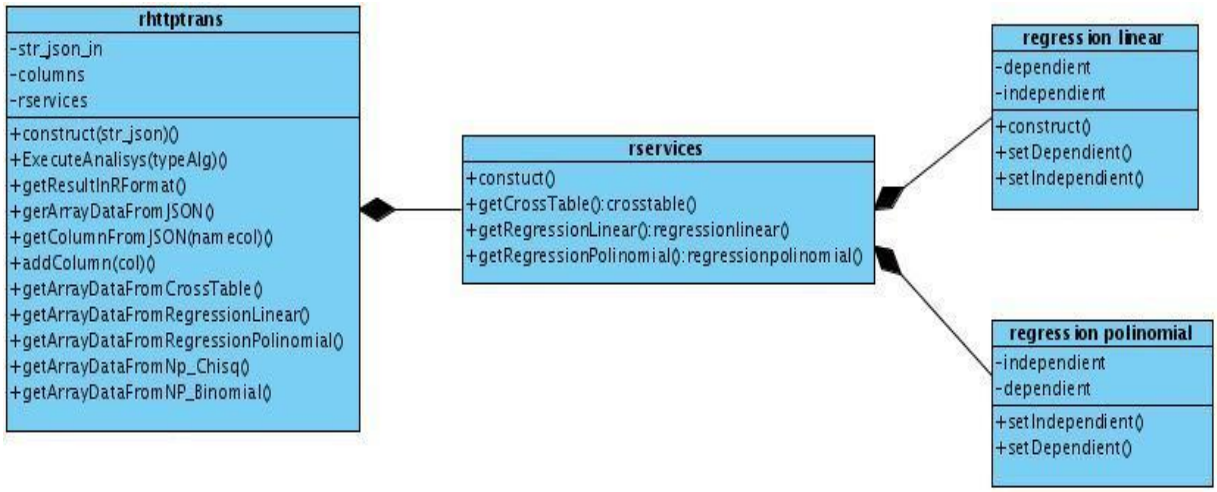


Fig. 12: Diagrama de clases del diseño.

Descripción de las clases

Clase	Descripción
rhttptrans	Implementa las conversiones de notación JSON a objetos en ambas direcciones
rservices	Implementa la fachada del trabajo con las librerías R a través de un lenguaje de alto nivel.
regresión linear	Define la estructura de regresión lineal que implementa los métodos necesarios para esta clase.
regresión polinomial	Define la estructura de regresión polinomial que implementa los métodos necesarios para esta clase.

Tabla 7: Descripción de las clases.

2.4 Conclusiones Parciales

En este capítulo se modelaron los casos de usos que dan respuestas a las necesidades planteadas. Se obtuvo el análisis y diseño de los algoritmos de regresión para el Servidor de Análisis Estadístico: R-SERVER. Se propuso un diseño basado en un modelo de arquitectura apoyado en el patrón arquitectónico en capas y de diseño la fachada, necesario para lograr un sistema con calidad.

Implementación y Validación

3.1 Introducción

En este capítulo se comienza el estudio de los principales conceptos de elementos usados para la elaboración de la implementación de estos algoritmos estadísticos, con presentación del código necesario para esta implementación. Además, se comienza la validación de la solución propuesta a través de las diferentes pruebas de unidad y es descrito un sistema de pruebas capaz de lograr la ausencia de errores en el producto final propuesto.

3.2 Implementación

3.2.1 Diagrama de componente

Los diagramas de componente describen los elementos físicos del sistema y sus relaciones. Muestran las opciones de realización incluyendo código fuente, binario y ejecutable. (JACOBSON 2000; PRESSMAN 2002)

En el diagrama de componente (fig.15), se muestran todas las capas que componen el servidor de análisis datos estadísticos (R-Server) donde se reflejan los componentes que componen cada capa y las relaciones que existen en cada una de ellas, permitiendo una validación del desarrollo.

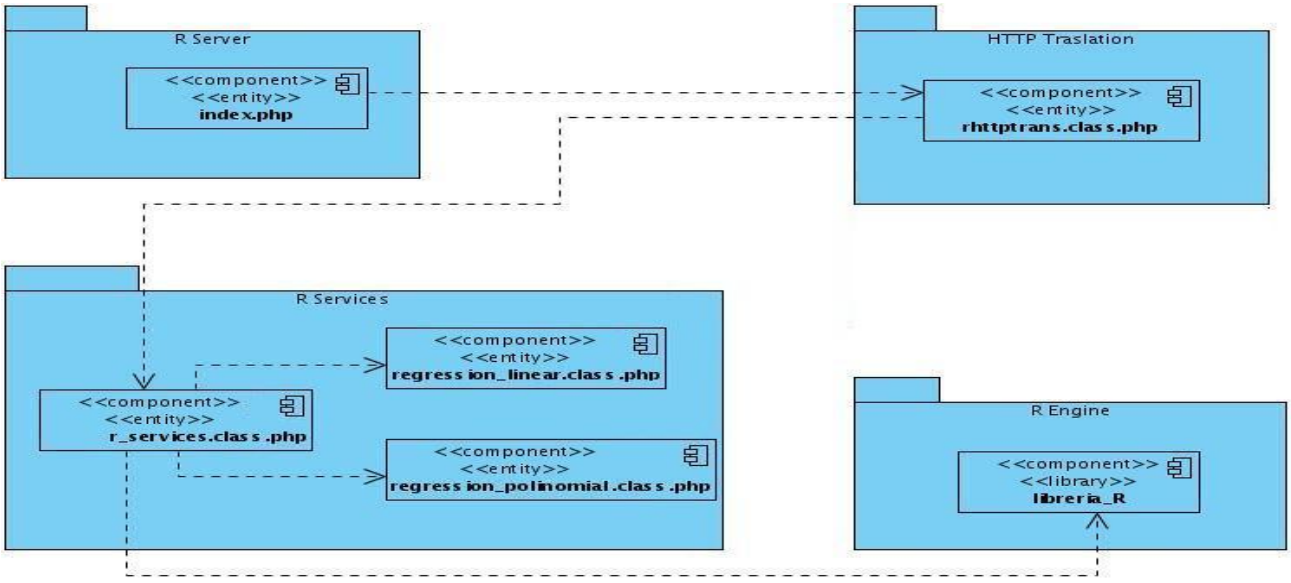


Fig. 13: Diagrama de componentes.

El diagrama de componente está compuesto por cuatro paquetes los cuales son: R Server, HTTP Translation, R Services y el R Engine. En el paquete R Server se agrupa el componente index.php que tiene la función de recibir la solicitud del algoritmo y del JSON por el método post de la URL y muestra el resultado generado por R, que está relacionado con el paquete HTTP Translation que contiene el componente http_trans.class.php que es una clase que tiene la funcionalidad de recoger en un arreglo el resultado generado por R.

Este componente está relacionado con el componente r_services.class.php del paquete de R Services, que Implementa la fachada del trabajo con las librerías R que está relacionada con los componentes de regression_linear.class.php y regression_polinomial.class.php que son las que contiene las instrucciones con el código R de los algoritmos a ejecutar. El componente r_services.class.php a su vez está relacionado con el componente librería_R del paquete R Engine que contiene las librerías a utilizar por los algoritmos y por el JSON.

3.2.2 JSON

Un JSON según sus siglas en ingles (*Java Script Object Notation*) no es más un formato ligero utilizado para el intercambio de datos, es un subconjunto de la notación literal de objetos de Java Script pero que no requiere el uso de XML. Gracias a su simplicidad el JSON ha dado lugar a la difusión de su uso, principalmente usándolo como una alternativa XML en AJAX (*Asincronous Javascript and XML* según siglas en inglés), puesto que una de las ventajas que presenta JSON sobre XML es que es mucho más sencillo escribir un analizador semántico de JSON. (JAVAHISPANO 2007)

Habitualmente en entornos donde el tamaño del flujo de datos entre el cliente y servidor es de vital importancia se emplea el JSON. Se utiliza para representar estructuras de datos simples llamados objetos y arreglos asociativos, existe código para analizar y generar datos JSON para una larga variedad de lenguajes de programación , dentro de algunas ventajas que presenta el JSON se encuentran que es más simple que el XML, es un mejor formato para intercambiar datos, su procesamiento por parte de los ordenadores es rápido, se necesitan librerías muy pequeñas para trabajar con él y en algunos de los casos es posible procesarlo sin el uso de librerías, dado su naturaleza es ideal para entornos AJAX. Pero a pesar de presentar estas ventajas tiene dentro de sus desventajas que es un formato mas reciente y con menos soporte a nivel de herramientas, es complicado de entender para los seres humanos, no puede ser utilizado para transportar imágenes y sonido, no permite describir interfaces gráficas y no es extensible.

Para la realización del análisis estadístico de regresión lineal y regresión polinomial se utiliza el código JSON para la entrada y salida de datos al Servidor de Análisis Estadístico R Server.

JSON de entrada para Regresión Lineal y Polinomial es el siguiente:

‘{

“dependiente”: {“nombre”:” valor”, “data”: [ ] },

“independiente”: {“nombre”: “valor”, “data”: [ ] }

}’

Los datos en el JSON no se validan, sino, se validan en la aplicación que va a interactuar con el Servidor de Análisis Estadístico. En los que se especifica bien en los requisitos funcionales la entrada de datos de las variables en el JSON.

Ejemplo de JSON de salida:

Regresión Lineal

{“dependiente”: “valor”, “independiente”:” valor”, “anova”:{ "independiente":{“df”:” valor”, “Sum_Sq”:”valor”, “Mean_Sq”:”valor”, “F_value”:”valor”, “Pr_f”:”valor”}, “residual”:{ "df":"valor","Sum_Sq":"valor","Mean_Sq":"valor"}},"coeficiente":{"intercep":{"Estimate":"valor", "Std_Error":"valor","t_value":"valor","Pr_t":"valor"},"independiente":{"Estimate":"valor","Std_Error":"valor ","t_value":" valor ","Pr_t":" valor "}}}

Regresión Polinomial

{"dependiente":" valor ","independiente":" valor ","coeficiente":{"Intersect":" valor ","Independiente":"valor ","Independiente2":" valor "},"summary":{"1":{"R_Square":" valor,"}," valor ":{"F":" valor ","DF1":" valor ","DF2":" valor ","P_Value":" valor "}}}

3.3 Pruebas

Las pruebas del software no son más que un elemento crítico para lograr la calidad del software, representa una revisión final de las especificaciones del diseño y de la codificación. Las pruebas son también actividades en la cual se somete un sistema o uno de sus componentes a una evaluación de

los resultados que proyectan en una base a la ejecución de éste en condiciones especificadas. Existen una gran variedad de pruebas del software, dentro de todas estas utilizamos la de caja negra.

Las pruebas de caja negra, también denominadas prueba de comportamiento, se centran en los requisitos funcionales del software, estas pruebas intentan encontrar errores de funciones incorrectas o ausentes, errores de interfaz, en estructuras de datos, errores de rendimiento y de inicialización y terminación. A diferencia de las pruebas de caja blanca estas pruebas se aplican durante las fases posteriores a la prueba, ya que ignora de forma intencional la estructura de control, y coloca su atención en el campo de la información. El sistema de pruebas de caja negra no considera la codificación dentro de sus parámetros a evaluar, es decir, no están basadas en el conocimiento del diseño interno del programa.(PRESSMAN 2002)

3.3.1 Casos de Prueba

Los casos de pruebas son un conjunto de variables o condiciones que lanzan un conjunto de resultados esperados desarrollados con un objetivo en particular. La primera prueba se toma como línea base para los siguientes siglos de pruebas y lanzamientos del producto. Los casos de pruebas descritos, incluyen una descripción de la funcionalidad que se probara, la cual es tomada ya sea de los requisitos o de los casos de usos.

Definición de prueba del CU “Realizar análisis estadístico de regresión lineal”

Descripción del caso de prueba

El caso de uso es una representación del concepto implícito de la realización de análisis de estadístico de regresión lineal, se inicia cuando el actor realiza la petición de análisis estadístico de algoritmos de regresión. (PRESSMAN 2002)

El caso de uso termina una vez que se realiza el análisis estadístico del algoritmo de regresión lineal y se devuelve el resultado al actor

Secciones a probar en el Caso de Uso.

Nombre de la sección	Escenarios de la sección	Descripción de la funcionalidad
Realizar análisis estadístico de regresión lineal.	EC 1.1 Realizar análisis estadístico de regresión lineal.	Permite realizar el análisis estadístico de regresión lineal.
	EC 1.2 Datos incorrectos	Termina el caso de uso.

Descripción de variables.

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Variable independiente	Arreglo JSON	No	Debe ser introducido el label de la independiente y sus datos asociados. El JSON debe ajustarse al siguiente formato: Independiente: name: label, data[<valores>]
2	Variable dependiente	Arreglo JSON	No	Debe ser introducido el label de la dependiente y sus datos asociados. El JSON debe ajustarse al siguiente formato: Dependiente: name: label, data[<valores>]

Matriz de datos.

Escenario	**Variable 1** ***Variable*** **Dependiente**	**Variable 2** ***Variable*** **Independiente**	**Respuesta del Sistema**	**Resultado de la Prueba**	**Flujo Central**
1.1: Realizar análisis estadístico de regresión lineal.	V	V	El sistema recibe todos los datos y realiza el análisis estadístico escogido.	Satisfactorio.	1- Se valida que se hayan enviado los datos obligatorios. 2- Se capturan los datos obligatorios para realizar el análisis estadístico. 3- Se realiza el análisis estadístico del algoritmo de regresión lineal. 4- Se devuelve un resultado.
1.2: Realizar análisis estadístico de regresión lineal con datos inválidos.	I	V	El sistema muestra un error y no puede ejecutar el análisis estadístico, porque existen datos inválidos.	Satisfactorio	1.1El sistema muestra un error.
	V	I			

3.3.2 Definición de prueba del CU “Realizar análisis estadístico de regresión polinomial”

Descripción del caso de prueba

El caso de uso es una representación del concepto implícito de la realización de análisis de estadístico de regresión polinomial, se inicia cuando el actor realiza la petición de análisis estadístico de algoritmos de regresión polinomial.

El caso de uso termina una vez que se realiza el análisis estadístico del algoritmo de regresión polinomial y se devuelve el resultado al actor.

Secciones a probar en el Caso de Uso.

Nombre de la sección	Escenarios de la sección	Descripción de la funcionalidad
Realizar análisis estadístico de regresión polinomial.	EC 1.1 Realizar análisis estadístico de regresión polinomial.	Permite realizar el análisis estadístico de regresión polinomial
	EC 1.2 Datos incorrectos	Termina el caso de uso.

Descripción de variables.

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Variable dependiente	Arreglo JSON	No	Debe ser introducido el nombre (label) de la variable independiente y sus datos asociados. El JSON debe ajustarse al siguiente formato: Dependiente: name : label, data[<valores>]
2	Variable independiente	Arreglo JSON	No	Debe ser introducido el nombre (label) de la independiente y sus datos asociados. El JSON debe ajustarse al siguiente formato: Independiente: name : label, data[<valores>]

Matriz de datos.

Escenario	**Variable 1** ***Variable*** **Dependiente**	**Variable 2** ***Variable*** **Independiente**	**Respuesta del Sistema**	**Resultado de la Prueba**	**Flujo Central**
1.1: Realizar análisis estadístico de regresión Polinomial.	V	V	El sistema recibe todos los datos y realiza el análisis estadístico escogido.	Satisfactorio.	1- Se valida que se hayan enviado los datos obligatorios. 2- Se capturan los datos obligatorios para realizar el análisis estadístico. 3- Se realiza el análisis estadístico del algoritmo de regresión polinomial. 4- Se devuelve un resultado.
1.2: Realizar análisis estadístico de regresión polinomial con datos inválidos.	I	V	El sistema muestra un error y no puede ejecutar el análisis estadístico, porque existen datos inválidos.	Satisfactorio	1.1El sistema muestra un error.
	V	I			

3.3.3 Validación con las pruebas de tiempo

Esta prueba consiste en entrar una cierta cantidad de datos en el JSON de entrada de las variables dependiente e independiente, logrando mostrar el resultado del algoritmo y a su vez el tiempo que demora en dar respuesta este análisis.

Para el algoritmo de regresión lineal simple:

Para 100 datos.

Array ([dependiente] => peso [independiente] => precio [anova] => Array ([independiente] => Array ([df] => 1 [Sum_S
=> 169 [Mean_Sq] => 168.67 [F_value] => 0.1988 [Pr_f] => 0.6567) [residual] => Array ([df] => 98 [Sum_Sq] => 8315
[Mean_Sq] => 848.53)) [coeficiente] => Array ([intercep] => Array ([Estimate] => 5.072e+01 [Std_Error] => 2.955e+(
[t_value] => 17.163 [Pr_t] => <2e-16) [independiente] => Array ([Estimate] => -9.276e-07 [Std_Error] => 2.081e-06
[t_value] => -0.446 [Pr_t] => 0.657))) ,``0,``0

Tiempo de respuesta:
1.4601380825043 segundos

Fig. 14: Prueba de tiempo.

Para 1000 datos.

Array ([dependiente] => peso [independiente] => precio [anova] => Array ([independiente] => Array ([df] => 1 [Sum_S
=> 1.6016e+28 [Mean_Sq] => 1.6016e+28 [F_value] => 1.1552e+39 [Pr_f] => <) [residual] => Array ([df] => 9998
[Sum_Sq] => 0.0000e+00 [Mean_Sq] => 0.0000e+00)) [coeficiente] => Array ([intercep] => Array ([Estimate] =>
0.000e+00 [Std_Error] => 3.725e-08 [t_value] => 0.000e+00 [Pr_t] => 1) [independiente] => Array ([Estimate] =>
1.000e+00 [Std_Error] => 2.942e-20 [t_value] => 3.399e+19 [Pr_t] => <2e-16))) ,``0,``0

Tiempo de respuesta:
12.511286973953 segundos

Fig. 15: Prueba de tiempo.

Para el algoritmo de regresión Polinomial:

Para 100 datos.

Array ([dependiente] => peso [independiente] => precio [coeficiente] => Array ([Intercet] => 0.000000e+00
[Independiente] => 1.000000e+00 [Independiente2] => -7.408448e-31) [summary] => Array ([1] => Array ([R_Square]
=> 1,) [2] => Array ([F] => 4.547e+35 [DF1] => 2 [DF2] => 997 [P_Value] => 2.2e-16)))

Tiempo de respuesta:
1.6010069847107 segundos

Fig. 16: Prueba de tiempo.

Para 1000 datos.

Array ([dependiente] => peso [independiente] => precio [coeficiente] => Array ([Intercet] => 0.000000e+00
[Independiente] => 1.000000e+00 [Independiente2] => -1.239322e-30) [summary] => Array ([1] => Array ([R_Square]
=> 1,) [2] => Array ([F] => 5.777e+38 [DF1] => 2 [DF2] => 9997 [P_Value] => 2.2e-16)))

Tiempo de respuesta:
12.511190891266 segundos

Fig. 17: Prueba de tiempo.

Como se puede observar a medida que aumenta la cantidad de de datos de entradas de las variables dependientes e independientes es mayor el tiempo de ejecución del algoritmo.

3.3.4 Prueba de Ancho de Banda

El ancho de banda en conexiones a internet no es más que la cantidad de información o de datos que se pueden enviar a través de una conexión de red en un período de tiempo. Es la longitud medida en HZ, de rango de frecuencias en el que se concentra la mayor parte de la potencia de la señal. En las redes de ordenadores, el ancho de banda a menudo se utiliza como sinónimo para la tasa de transferencia de datos, la cantidad de datos que se pueden llevar de un punto a otro en un período dado (generalmente en un segundo). Este tipo de ancho de banda generalmente se expresa en bits por segundos (bps), aunque en ocasiones se expresa como bytes por segundos (Bps). En general, una conexión con ancho de banda alto es aquella que puede llevar la suficiente información como para sostener la sucesión de imágenes en una presentación de video. Este puede referirse a la capacidad de ancho de banda disponible en bits, lo cual significa el rango neto de bits o la máxima salida de una huella de comunicación lógico o físico en un sistema de comunicación digital.

MRTG (Multi Router Traffic Grapher) es una herramienta para monitorizar la carga de tráfico sobre determinados nodos de una red. Genera páginas HML que incluyen representaciones gráficas, en formato GIF, del tráfico registrado en un determinado nodo de red. Hace un chequeo de comportamiento de ancho de banda en el cliente y el servidor, cuando el cliente consulta a este último, generando gráficas de consumo vs tiempo.

Recolecta la información del tráfico del dispositivo (habitualmente *routers*) la herramienta utiliza el protocolo SNMP (*Simple Network Management Protocol*). La Herramienta MRTG se pueden instalar en sistemas operativos como: GNU/Linux, Windows, AIX. Posee una licencia pública general.

SNMP es un protocolo de gestión de red o sea un conjunto de estructuras que permiten tener datos concretos del tráfico que se produce en la red, así como quien lo produce. Opera en el nivel de aplicación utilizando el protocolo de transporte TCP/IP, lo que ignora los aspectos específicos del hardware sobre el que funciona. Este protocolo está compuesto por dos elementos el agente (Agent), y el gestor (Manager). Es una arquitectura cliente servidor en la cual el agente desempeña el papel del servidor y el gestor de cliente. Este protocolo proporciona la información en crudo de la cantidad de bytes que han pasado por ellas distinguiendo la entrada de la salida. Esta cantidad bruta deberá ser tratada adecuadamente para la generación de informe.

Una vez que estuvo listo el servidor y la aplicación fue publicada, se corrieron 594 peticiones desde las PC clientes al servidor, utilizando 7 PC clientes y un servidor donde fue monitoreada la prueba con un ancho de banda de 10-100 Mbps y una velocidad máxima de 12.5Mbytes\s. Como se puede observar (fig. 20), el color azul indica el consumo de salida y el color verde indica el consumo de entrada. Esta prueba fue realizada en un período de tiempo de una hora, con un consumo de entrada aproximadamente de 1bytes\s y de salida aproximadamente 9.1bytes\s.

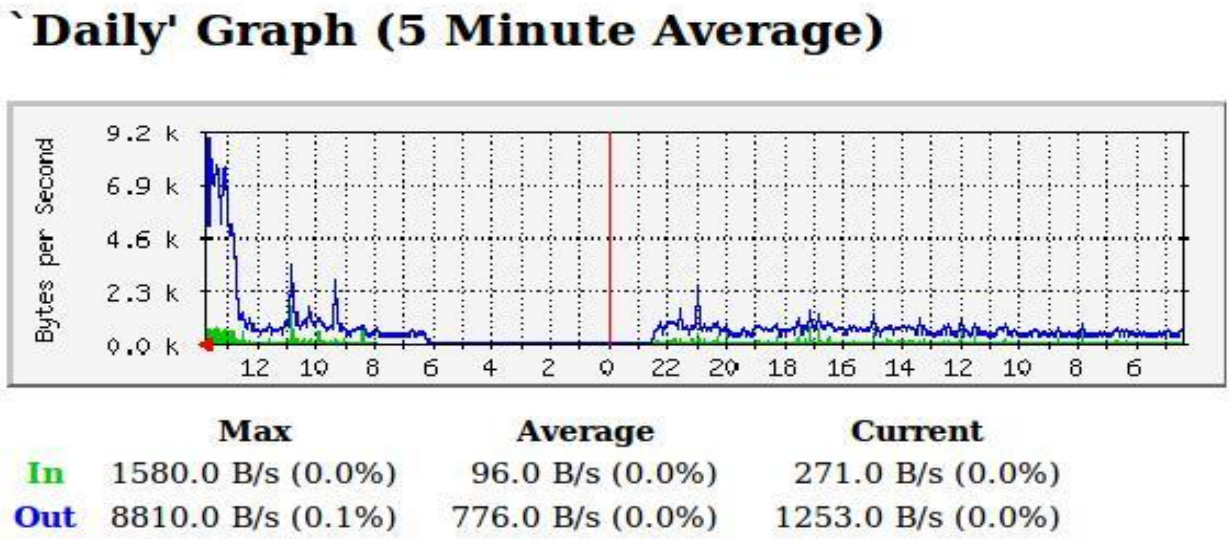


Fig. 18: Prueba de Ancho de Banda.

También se utilizó VMStat para recoger los datos de consumo del CPU, IN, OUT, SWAP (espacio de intercambio de un disco). El cual muestra información relativa al sistema de memoria. Se puede especificar el intervalo entre la muestra y el número de muestra. También realiza un chequeo del comportamiento del servidor y el cliente cuando este se conecta al servidor.

Muestra los valores siguientes:

r---- procesos esperando a ser ejecutado.

b----proceso durmiendo ininterrumpidamente.

swpd-----memoria virtual en uso (kb).

free-----memoria física libre (kb).

buff-----memoria usada como buffer.

cache---memoria usada como cache.

si---memoria intercambiada desde disco (KB\s).

so----memoria intercambiada hacia disco (KB\s).

bi----bloque de memoria por segundo enviados al disco.

bo--- bloque de memoria por segundos recibidos desde disco.

in---interrupciones por segundo.

cs----cambio de contexto por segundo.

us---uso del procesador ejecutando código de usuario.

sy----uso del procesador ejecutando código del sistema operativo.

Id—porcentaje de tiempo con el procesador ocioso.

Se observa con esta prueba que a medida que aumentan las peticiones del servidor el consumo del CPU y las interrupciones por segundo aumenta y disminuye en función de las peticiones realizadas.

```
procs -----------memory---------- ---swap-- -----io----   -system-- ----cpu----
 r  b   swpd   free   buff  cache   si   so    bi    bo     in   cs us sy id wa
 3  0      0  77052 108908 518300    0    0    23    12     70  226  2  1 97  0
 0  0      0 118508 108920 518440    0    0     0    24    260  217 34  0 65  0
 0  0      0 118880 108920 518288    0    0     0     0    117  231  1  0 99  0
 0  0      0 114532 108928 521360    0    0     0    22    246 2203  5  2 93  0
 0  0      0 115408 108936 520464    0    0     0     8     79  169  0  0 100  0
 2  0      0 111068 108936 524872    0    0     0     2    305  379 15  1 84  0
 0  0      0 109704 108948 526772    0    0     0    18    193 1919  3  1 96  0
 0  0      0 115284 108956 520668    0    0     0    77    180  410  1  0 99  0
 0  0      0 115532 108964 520528    0    0     0     9    205  758  1  0 99  0
 0  0      0 111440 108972 524856    0    0     0    10    108  259  2  0 97  0
 0  0      0 114168 108980 522108    0    0     0    14     75  171  0  0 100  0
 1  0      0 115284 108980 520488    0    0     0     6    251  804  2  1 97  0
 0  0      0 116772 108988 518952    0    0     0     8    142  268  1  0 99  0
 0  0      0 106480 108996 529716    0    0     0    24    227 1919  5  1 94  0
 0  0      0  98916 109004 536664    0    0     0    10    394 1348  4  2 94  0
 0  0      0 114044 109004 521288    0    0     0     6    345 1701  1  2 97  0
 0  0      0 115904 109012 519524    0    0     0    34    366 2290  2  2 96  0
 0  0      0 116028 109028 519328    0    0     0    12    117  424  0  0 99  0
 0  0      0 112556 109028 523208    0    0     0     0    147  649  1  1 99  0
```

Fig. 19: Prueba de Ancho de Banda con VMStat.

3.4 Conclusiones parciales

En este capítulo se desarrolló toda la implementación y validación de los algoritmos de regresión lineal simple y polinomial, donde se realizó pruebas que juegan un papel primordial ya que proporciona al cliente y al desarrollador confiabilidad en el producto final. Fueron descritas las principales pruebas de casos de usos con el objetivo de asegurar que el sistema cumpla con los requerimientos funcionales planteados, las pruebas de tiempo el cual se puede observar claramente que a medida que aumenta la cantidad de datos de entrada en el JSON aumenta también el tiempo de respuesta del sistema. Las pruebas de Ancho de Banda donde se analizó la conexión de red en un período de tiempo en que se en vio los datos.

Conclusiones generales

Una vez culminado el presente trabajo y después de implementar los algoritmos de regresión del Servidor de Análisis Estadístico, se puede decir que se cumplió con el objetivo general propuesto. Y se considera además que se cumplieron los siguientes objetivos específicos.

- Se realizó un estudio del arte de las distintas herramientas, lenguajes y librerías que existen en el mundo para el análisis de datos, teniendo en cuenta las características más relevantes donde se obtuvo que la librería R es la más eficiente para el desarrollo de los algoritmos de análisis de datos estadísticos.

- Se lograron los artefactos necesarios del análisis y la implementación cumpliendo con los requisitos funcionales y no funcionales, logrando un correcto desarrollo del software.

- Se alcanzó incorporar al R-Server el desarrollo de los métodos estadísticos de regresión lineal y polinomial, aumentando las funcionalidades del mismo.

- Se realizó la validación a partir de la realización de diferentes tipos de pruebas para probar la fiabilidad del sistema y validar el trabajo realizado.

Recomendaciones

Luego de concluir este trabajo se recomienda lo siguiente:

- Continuar con el desarrollo de los algoritmos de regresión logística binaria y multinominal, regresión no lineal, mínimos ponderados, ordinal, prolsit y escalamiento óptimo para el Servidor de Análisis Estadísticos: R-SERVER.

ALMAGUER, Y. M. B. J. L. M. S. J. C. Q. L. A. R. P. Propuesta de metodología de un grupo de análisis que utiliza un modelo de desarrollo basado en líneas de productos de software, 2009

COLE, J. H. Enciclopedia Multimedia Virtual Interactiva. "Nociones de Regresión Lineal", 2005.

DEBESA F, J. G., PÉREZ J, ÁVILA J. La estrategia de Farmacoepidemiología en Cuba. Características y papel de la Unidad Coordinadora de Farmacovigilancia en Cuba, 2007. 41: 3.

GRACIA, J. Manual de PHP, 2004. [Disponible en: http://www.desarrolloweb.com

JACOBSON, G. B. J. R. I. El Proceso Unificado del Desarrollo de Software. 2000. 464 p.

JAVAHISPANO, C. javaHispano, 2007. [Disponible en: http://www.javahispano.org/contenidos/es/JSON_vs_xml

-MARCANO AULAR, Y. J. Y. T. P., ROSALBA. Minería e Datos como soporte a la toma de decisiones empresariales, 2007. 23: 104-118.

MOURITSEN, A. H. Y. J. "La tecnología de redes de gestión y anotó: Competitividad en Acción - El trabajo de traducir el rendimiento en una empresa de alta tecnología, la Organización", 1999. 6: 3.

ORALLO, C. F. R. M. J. R. Q. J. H. Introducción a la Minería de Datos. Prentice Hall. 2004. p. 8420540919

PARADIS, E. Libro R para principiantes. 2003. p. F-34095

PRESSMAN, R. S. Ingeniería del Software. Un enfoque práctico., 2002. p.

SALAS, C. ¿Por qué comprar un programa estadístico si existe R? , 2008. 223-231 p. 1667-782X

TOLEDANO, M. J. Minería de Datos, 2006. [2009]. Disponible en: http://datamining.iespana.es

Bibliografía

1-ALMAGUER, Y. M. B. J. L. M. S. J. C. Q. L. A. R. P. Propuesta de metodología de un grupo de análisis que utiliza un modelo de desarrollo basado en líneas de productos de software, 2009

2-BUSCHAMANN, R. M. H. R. P. S. M. S. F. Pattern Oriented Software Architecture. 1996. p.

3-Toledo, A. NetBeans 6.8. NetBeans IDE, 2010. [Disponible en: http://netbeans-ide.uptodown.com/

4-COLE, J. H. Enciclopedia Multimedia Virtual Interactiva. "Nociones de Regresión Lineal", 2002. [Disponible en: http://fce.ufm.edu/catedraticos/jhcole/regresion.htm

5-Damodar, N. Gujarati. Libro Econometría. [2010].

6-DEBESA F, J. G., PÉREZ J, ÁVILA J. La estrategia de Farmacoepidemiología en Cuba. Características y papel de la Unidad Coordinadora de Farmacovigilancia en Cuba, 2007. 41: 3. [Disponible en: http://scielo.sld.cu/scielo.php?pid=S0034-75152007000300003&script=sci_arttext

7-Fernández, J. S.P. Gestión SNMP con Linux, 2001. [Disponible en: http://beta.redes-linux.com/manuales/Monitorizacion_redes/snmp.pdf

8-Freund, John E. M. I. R. J.R. Libro Probabilidad y Estadística para ingenieros. 2006.

9-Free Download Manager. Paradigma Visual, 2007. [2010]. Disponible en: http://www.freedownloadmanager.org/es/downloads/Visual_Paradigm_for_UML_%28LE%29_%5BWindows%5D_14728_p/

10-Garzón, J.A. Monitorización gráfica del tráfico de red y otros parámetros del sistema. [2010]. [Disponible en: http://beta.redes-linux.com/manuales/Monitorizacion_redes/mrtg.pdf

11-GRACIA, J. Manual de PHP, 2004. [Disponible en: http://www.desarrolloweb.com/articulos/436.php

12-JACOBSON, G. B. J. R. I. El Proceso Unificado del Desarrollo de Software. 2000. 464 p.

13-JAVAHISPANO, C. javaHispano, 2007. [Disponible en: http://www.javahispano.org/contenidos/es/JSON_vs_xml/

14-MARCANO AULAR, Y. J. Y. T. P., ROSALBA. Minería e Datos como soporte a la toma de decisiones empresariales, 2007. 23: 104-118.

15-MARCIAL, F. TAWS Grupo de investigación FIEC-ESPOL, 20009. [Disponible en: http://blog.espol.edu.ec/taws/2009/01/20/

16-MOURITSEN, A. H. Y. J. "La tecnología de redes de gestión y anotó: Competitividad en Acción - El trabajo de traducir el rendimiento en una empresa de alta tecnología, la Organización", 1999. 6: 3.

17-MUÑOZ, J. L. S. Descripción de la línea de desarrollo de soluciones integrales, 2009.

18-ORALLO, C. F. R. M. J. R. Q. J. H. Introducción a la Minería de Datos. Prentice Hall. 2004. p. 8420540919

19-PARADIS, E. Libro R para principiantes. 2003. p. F-34095

20-PRESSMAN, R. S. Ingeniería del Software. Un enfoque práctico., 2002. p.

21-PROYECTOSENA. Minería de Datos, proyectosena, 2008. [2009]. Disponible en: http://proyectosena.soy.es/2008/08/29/algoritmos-de-mineria-de-datos-2/

22-SALAS, C. ¿Por qué comprar un programa estadístico si existe R? , 2008. 223-231 p. 1667-782X

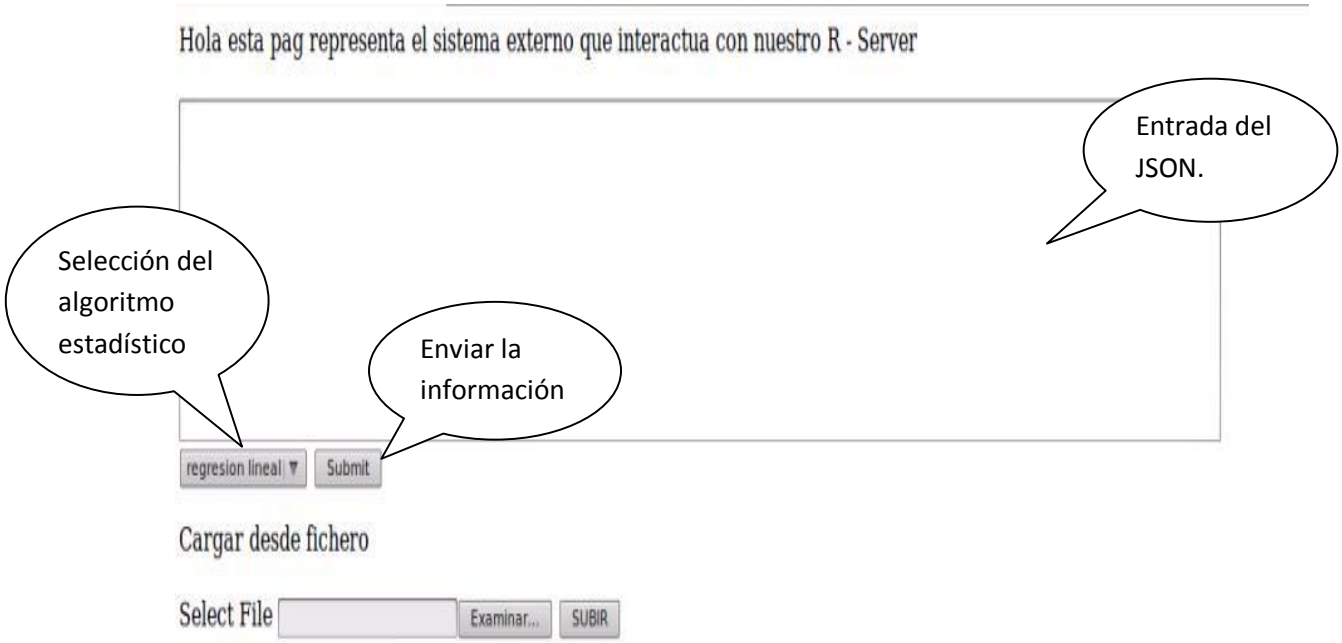
23-Sitio Web de la E.U.de Ingeniería Técnica Informática de Oviedo. IDE, 2001. [2010]. Disponible en: http://petra.euitio.uniovi.es/~i1667065/HD/documentos/Entornos%20de%20Desarrollo%20Integrado.pdf

24-TOLEDANO, M. J. Minería de Datos, 2006. [2009]. Disponible en: http://datamining.iespana.es

25-VALLS, I. P. Hola Mundo Java con NetBeans 6, 2008. [Disponible en: http://www.javadabbadoo.org

Anexos

Anexo: 1 Sistema externo que interactúa con el R Server.



Anexo: 2 Clase de Regresión Lineal en PHP

```
class RegressionLinear extends Algoritmo{
    public function CreateRFile($str){
        $path = tempnam('/var/www/rserver/r_services/', 'in');
        $_SESSION["inputFile"] = $path;
        $file = fopen($path, 'w');
        fputs($file, "library(rjson)");
        fputs($file, "\n");
        fputs($file, "library(rjson)");
        fputs($file, "\n");
        fputs($file, "str_json_in<-");
        fputs($file, $str);
        fputs($file, "\n");
        fputs($file, "obj<- fromJSON(str_json_in)");
        fputs($file, "\n");
        fputs($file, "RegModel<- lm(dependiente\$data ~ independiente\$data, data=obj)");
        fputs($file, "\n");
        fputs($file, "anova(RegModel)");
        fputs($file, "\n");
        fputs($file, "summary(RegModel)");
        fclose($file);
    }
    //put your code here
}
?>
```

Anexo: 3 Clase de Regresión Lineal en PHP

```php
class RegressionPolinomial extends Algoritmo{
    public function CreateRFile($str){
        $path = tempnam('/var/www/rserver/r_services/', 'in');
        $_SESSION["inputFile"] = $path;
        $file = fopen($path, 'w');
        fputs($file, "library(rjson)");
        fputs($file, "\n");
        fputs($file, "library(rjson)");
        fputs($file, "\n");
        fputs($file, "str_json_in<-");
        fputs($file, $str);
        fputs($file, "\n");
        fputs($file, "ob<- fromJSON(str_json_in)");
        fputs($file, "\n");
        fputs($file, "RegModel<- lm(dependiente\$data ~ indeperdiente\$data +I(independiente\$data^2), data=ob)");
        fputs($file, "\n");
        fputs($file, "RegModel\$coef");
        fputs($file, "\n");
        fputs($file, "sunmary(RegModel)");
        fclose($file);
    }
    //put your code here {
}
?>
```

Anexo5: Pruebas de tiempo de regresión lineal

Para 2500 datos.

Array ([dependiente] => peso [independiente] => precio [anova] => Array ([independiente] => Array ([df] => 1 [Sum_S
=> 4.8043e+27 [Mean_Sq] => 4.8043e+27 [F_value] => 9.4897e+34 [Pr_f] => <) [residual] => Array ([df] => 2698
[Sum_Sq] => 0.0000e+00 [Mean_Sq] => 0.0000e+00)) [coeficiente] => Array ([intercep] => Array ([Estimate] =>
4.946e-06 [Std_Error] => 4.333e-06 [t_value] => 1.142e+00 [Pr_t] => 0.254) [independiente] => Array ([Estimate] =>
1.000e+00 [Std_Error] => 3.246e-18 [t_value] => 3.081e+17 [Pr_t] => <2e-16))) ,``0,``0

Tiempo de respuesta:
3.457967042923 segundos

Para 7500 datos.

Array ([dependiente] => peso [independiente] => precio [anova] => Array ([independiente] => Array ([df] => 1 [Sum_S
=> 1.2812e+28 [Mean_Sq] => 1.2812e+28 [F_value] => 3.3162e+35 [Pr_f] => <) [residual] => Array ([df] => 7498
[Sum_Sq] => 0.0000e+00 [Mean_Sq] => 0.0000e+00)) [coeficiente] => Array ([intercep] => Array ([Estimate] =>
0.000e+00 [Std_Error] => 2.271e-06 [t_value] => 0.000e+00 [Pr_t] => 1) [independiente] => Array ([Estimate] =>
1.000e+00 [Std_Error] => 1.737e-18 [t_value] => 5.759e+17 [Pr_t] => <2e-16))) ,``0,``0

Tiempo de respuesta:
9.3536961078644 segundos

Anexo 6: Pruebas de tiempo de regresión polinomial

Para 2500 datos.

Array ([dependiente] => peso [independiente] => precio [coeficiente] => Array ([Intercet] => 4.909553e-07 [Independiente] => 1.000000e+00 [Independiente2] => 3.171150e-30) [summary] => Array ([1] => Array ([R_Square] => 1,) [2] => Array ([F] => 4.743e+34 [DF1] => 2 [DF2] => 2697 [P_Value] => 2.2e-16)))

Tiempo de respuesta:
3.4975349903107 segundos

Para 7500 datos.

Array ([dependiente] => peso [independiente] => precio [anova] => Array ([independiente] => Array ([df] => 1 [Sum_Sq] => 1.2812e+28 [Mean_Sq] => 1.2812e+28 [F_value] => 3.3162e+35 [Pr_f] => <) [residual] => Array ([df] => 7498 [Sum_Sq] => 0.0000e+00 [Mean_Sq] => 0.0000e+00)) [coeficiente] => Array ([intercep] => Array ([Estimate] => 0.000e+00 [Std_Error] => 2.271e-06 [t_value] => 0.000e+00 [Pr_t] => 1) [independiente] => Array ([Estimate] => 1.000e+00 [Std_Error] => 1.737e-18 [t_value] => 5.759e+17 [Pr_t] => <2e-16))) ,``0,``0

Tiempo de respuesta:
9.3536961078644 segundos

Juego de Datos de Regresión Lineal Simple				
No	**Variables**			**Tiempo de Respuesta (segundos)**
1	**Dependiente**			1.46
	Nombre	**Tipo**	**Cantidad**	
	Precio	Numérico	100	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Peso	Numérico	100	
2	**Dependiente**			3.45

	Nombre	**Tipo**	**Cantidad**	
	Precio	Numérico	2500	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Peso	Numérico	2500	
3	**Dependiente**			9.35
	Nombre	**Tipo**	**Cantidad**	
	Precio	Numérico	7500	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Peso	Numérico	7500	
4	**Dependiente**			12.51
	Nombre	**Tipo**	**Cantidad**	
	Precio	Numérico	1000	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Peso	Numérico	100	

Juego de Datos de regresión polinomial.				
No	**Variables**			**Tiempo de Respuesta (segundos)**
1	**Dependiente**			1.60
	Nombre	**Tipo**	**Cantidad**	
	Pulgadas		100	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Cargas		100	
2	**Dependiente**			3.49
	Nombre	**Tipo**	**Cantidad**	
	Pulgadas		2500	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Cargas		2500	
3	**Dependiente**			9.24
	Nombre	**Tipo**	**Cantidad**	
	Pulgadas		7500	
	Independiente			
	Nombre	**Tipo**	**Cantidad**	
	Cargas		7500	

<table>
<tr><td>4</td><td colspan="3">Dependiente</td><td>12.51</td></tr>
<tr><td></td><td>Nombre</td><td>Tipo</td><td>Cantidad</td><td></td></tr>
<tr><td></td><td>Pulgadas</td><td></td><td>1000</td><td></td></tr>
<tr><td></td><td colspan="3">Independiente</td><td></td></tr>
<tr><td></td><td>Nombre</td><td>Tipo</td><td>Cantidad</td><td></td></tr>
<tr><td></td><td>Cargas</td><td></td><td>1000</td><td></td></tr>
</table>