UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 6

Obtención de preferencias de usuario a partir de métodos de regresión lineal de Minería de datos.

Trabajo de diploma para optar por el título de Ingeniero en Ciencias Informáticas.

Autor:
Nixys Leydis Báez Polanco.

Tutor:
Ing. Yunier Albrecht Delgado.

Co-Tutor:
Msc. Yunier Emilio Tejeda Rodríguez.

Ciudad de la Habana, Junio de 2011

Año 53 de la Revolución

"El conocimiento es experiencia,
todo lo demás es información"

Albert Einstein

Declaración de Autoría

Yo: Nixys Leydis Báez Polanco declaro que soy la única autora del trabajo titulado “Obtención de preferencias de usuario a partir de métodos de regresión lineal de Minería de Datos”, y autorizo a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmo la presente a los 30 días del mes de junio del año 2011.

Nixys Leydis Báez Polanco — Firma del Autor

Yunier Albrecht Delgado — Firma del Tutor

Yunier Emilio Tejeda.

Firma del Co-Tutor

Opinión de Tutor

SOBRE EL TRABAJO DE DIPLOMA PRESENTADO PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Título: Obtención de preferencias de usuario a partir métodos de regresión lineal de Minería de datos.
Autor(es): Nixys Leydis Báez Polanco

El tutor del presente Trabajo de Diploma considera lo siguiente:

El documento presentado se encuentra correctamente estructurado en correspondencia con lo establecido para una tesis de este tipo. Existe correspondencia entre los objetivos planteados y los cumplidos, los contenidos están bien referenciados y la bibliografía utilizada es actual.
El resultado obtenido es de una gran importancia para el proyecto Plataforma VideoWeb porque le brinda al mismo, métodos y conocimientos para mejorar los servicios ofrecidos en el producto desarrollado, ofrece un gran valor agregado y esta es una cuestión que es muy valorada en el mercado de software donde pretende insertarse el producto y que es de vital importancia para la economía del país.
El cumplimiento de las metas trazadas no habría sido posible sin la gran responsabilidad mostrada por la estudiante ante las tareas a realizar, su independencia y creatividad le ayudaron en la búsqueda de soluciones ante los problemas encontrados durante el camino del desarrollo de la investigación, demostrando además la capacidad para apropiarse de los conocimientos necesarios para llevar a término completo el objetivo propuesto.
Teniendo en cuenta lo antes expresado, la exposición realizada y el cumplimiento de los objetivos propuestos, se considera que la estudiante se encuentra apta para ejercer como Ingeniera en Ciencias Informáticas y se propone al Tribunal la calificación de 5 puntos.

Ing. Yunier Albrecht Delgado — **30/06/11**

Firma — Fecha

Datos de Contacto

Tutor: Ing. Yunier Albrecht Delgado.

- ✓ Ingeniero en Ciencias Informáticas, Universidad de las Ciencias Informáticas, 2011.
- ✓ Jefe del Proyecto de Video y Sonido Digital, Facultad 6, Universidad de las Ciencias Informáticas.
- ✓ Profesor del Departamento de Técnicas de Programación, Facultad 6, Universidad de las Ciencias Informáticas.

Correo electrónico: yalbrecht@uci.cu

Co Tutor: Msc. Yunier Emilio Tejeda.

- ✓ Máster en Ciencias Matemáticas.
- ✓ Profesor del Departamento de Ciencias Básicas, Facultad 6, Universidad de las Ciencias Informáticas.

Correo electrónico: yuniere@uci.cu

Opiniones y Avales

Dedicatoria

A mi mamá, por la educación que me ha dado, por la confianza que siempre ha depositado en mi y por ser mi guía e inspiración.

A todas las personas que me ayudaron, me apoyaron y estuvieron siempre a mi lado aún cuando la cima parecía estar más lejos, a aquellos que siempre creyeron en mí y no me abandonaron ni un momento, a todos ustedes dedico mi triunfo.

Agradecimientos

A la Universidad de las Ciencias Informáticas por los conocimientos adquiridos.

A los Yunier, gracias por su dedicación, orientación y consejos.

A los profesores Frank y Angel, gracias por la ayuda y el apoyo brindado, sin ustedes este sueño no se hubiera realizado.

A mi mamá y mi segundo padre Carlos, gracias por creer en mi en todo momento y estar siempre a mi lado, esta meta es de ustedes.

A mis dos abuelitas, gracias por el apoyo y cariño que me han brindado.

A mi amiga y hermana Diana, por su apoyo incondicional durante estos cinco años, hoy vemos realizado uno de tantos sueños.

A mi prima Martha que mucho ha ayudado a que todo saliera como hasta hoy.

A mis amigas doradas, Any, Arlén, Glenda y Daylis, gracias por su eterno amor.

A mi familia, especialmente a mis hermanos Bia, Carla, Kleidis y Kevin, a mi papá Carlos, a tías Nixy, Leydis y China, y mis primos Gaby, Kamila y Barbara.

A la madrasta que más quiero Katy, gracias por tu apoyo y cariño.

A mis mejores amigos de la universidad, los que se dicen de toda la vida Yanet, Michelena, Adaylis, Yalili, la Flaca y el Paca, por ser tan especiales y saber siempre sobrellevarme incluso en esos malos días.

A Alejandro por ayudarme tanto en este mi último año.

A mi mejor amigo y guía Rolando, gracias por brindarme tu tiempo y amor.

A las chicas locas del apartamento Lusmey, Yara, Elizabetha, Lisandra, Maylén, Grechyn, Leydis y Yanet Ramos, gracias a ustedes por todos los momentos buenos que pasamos en este nuestro último año de universidad.

A mis amistades de la UCI, los que están y los que ya no están, que me ayudaron mucho no sólo en el desarrollo de mi tesis, sino a lo largo de estos cinco años de estudio.

Resumen

La capacidad para almacenar información ha crecido en los últimos años a velocidades exponenciales. Contrario a esto, la velocidad de procesar y utilizar estos datos no ha ido a la par. Por tal motivo la Minería de Datos, se presenta como una tecnología de apoyo para extraer información y conocimiento útil contenido en grandes volúmenes de datos. La información obtenida será de gran utilidad para descubrir nuevos caminos y tendencias ocultos en los datos que ayuden en la toma de decisiones para lograr predecir el comportamiento de un futuro cliente. Este documento ofrece una perspectiva general de todo el proceso que trae consigo la Minería de Datos, basando su uso en el empleo de la técnica estadística de Regresión Lineal Múltiple. Seguidamente se describen los cambios propuestos en la estructura de almacenamiento de información, para poder almacenar el conocimiento que permita llevar a cabo el proceso de personalización de la interfaz de la Plataforma VideoWeb. Igualmente se describen además, algunas de las herramientas estadísticas existentes en el mundo, que permitan validar los resultados obtenidos.

Palabras clave: Minería de Datos, personalización, Regresión Lineal Múltiple.

Abstract

The capacity of storage of data has grown in the recent years to exponential speeds. Contrary to this, the speed of processing and using these data has not kept pace. For this reason data mining is presented as a supporting technology to extract useful information and knowledge contained in large volumes of data. The information obtained will be useful to discover new ways and trends hidden in data that help in making decisions that can make us to predict the future behavior of a client. This document provides a general overview of the process that brings data mining, basing its use on the use of the statistical technique of multivariate linear regression. Also it describes the proposed changes in the structure of information storage, to store knowledge that allows doing the process of customizing the interface VideoWeb Platform. Also describes some statistical tools in the world, to validate the results.

Keywords: data mining, personalization, Multiple Linear Regression.

Índice

Índice de Figuras

Índice de Tablas

Introducción

Internet ha supuesto una revolución sin precedentes en el mundo actual, las nuevas generaciones viven en un mundo de colaboración básica y accesible, donde la web se ha convertido en una herramienta primaria de comunicación global.

La web abre un mercado con grandes oportunidades, dando paso al surgimiento de empresas virtuales, comercio electrónico y aplicaciones que brindan la posibilidad de comerciar productos y servicios, mediante la interacción con todo tipo de contenidos, sean estos: videos, imágenes, textos u otros contenidos que permitan la interrelación de información. Una de las tendencias en la búsqueda de más ganancias y mejor tratamiento a los clientes, fue la personalización de la información orientada a los usuarios. En este sentido, la información necesita ser procesada, donde los contenidos exhibidos estén centrados a las preferencias de los usuarios, con la finalidad de encontrar relaciones dentro de los datos, que reflejen la satisfacción del cliente y la posibilidad de adecuarse a cada usuario en particular.

Teniendo en cuenta todo lo planteado anteriormente, se puede concretar, que personalizar es mostrar a cada usuario los contenidos más apropiados de acuerdo con alguna preferencia, hábito o gusto que queda recogido en variables tales como: el destino preferido, comportamiento manifestado durante una sesión de navegación o el perfil de usuario.

El manejo de la información recogida de la interacción del cliente con estas aplicaciones puede ser difícil, debido a que el volumen de datos que se genera se multiplica día a día. Actualmente se presenta la paradoja de que, cuantos más datos están disponibles, menor información se tiene.

Para superar este problema, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento de los datos y permiten realizar un análisis a fondo de los mismos. La idea clave es que los datos contienen más información de la que se ve a simple vista. El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (*Data Mining*), que entre otras sofisticadas técnicas aplica la Estadística para encontrar tendencias o patrones dentro de los datos.

En Cuba, en la Universidad de las Ciencias Informáticas (UCI), el proyecto Plataforma VideoWeb, dentro del Centro de Geoinformática y Señales Digitales (Geysed) de la Facultad 6, tiene como objetivo desarrollar una solución web para la gestión y transmisión de contenidos multimedia a través de la red de datos. Este sistema está diseñado para el consumo bajo demanda de materiales audiovisuales, pero además ofrece servicios de tv y radio en vivo. La información publicada para los usuarios obedece a una determinada organización previamente diseñada y se mantiene de manera estática para todos los usuarios por igual. En la primera versión del producto, la interacción de los usuarios con el sistema en el consumo de los materiales publicados, las preferencias o necesidades de información manifestadas por estos, no son registradas de ninguna forma. La aplicación no cuenta con una infraestructura de almacenamiento de datos en la cual guardar todo el historial de navegación que se registra a partir de la interacción de los usuarios con la Plataforma VideoWeb. De tener esta infraestructura de almacenamiento, se hace necesario el análisis de los datos guardados para convertirlos en información útil, que posibilite la personalización de la información orientada al usuario en aplicaciones como la que se desarrolla.

A raíz de las situaciones expuestas, surge el siguiente **problema** a resolver:

¿Cómo obtener una base de conocimientos sobre preferencias de usuario a partir de su interacción con la Plataforma VideoWeb?

Para dar respuesta al problema antes mencionado, se plantea como **objetivo general**:

Utilizar un algoritmo basado en la técnica de Regresión Lineal para el análisis del comportamiento de los usuarios en la Plataforma VideoWeb.

El **objeto de estudio** lo constituyen los algoritmos dentro de la técnica de Regresión Lineal en la Minería de Datos.

El **campo de acción** que abarca este trabajo está enmarcado en la obtención de preferencias de usuario durante su interacción con la Plataforma VideoWeb.

Esta investigación se guiará **defendiendo la idea** de que aplicando un algoritmo basado en la técnica de Regresión Lineal de la Minería de Datos se garantizará la obtención de preferencias de usuario en la Plataforma VideoWeb.

Para alcanzar dichos objetivos se planteó desarrollar las siguientes **tareas de investigación**:

1. Analizar la técnica de Regresión Lineal en la Minería de Datos.
2. Definir las variables para formar la base de conocimientos.
3. Analizar soluciones existentes.
4. Modificar la base de datos de la Plataforma VideoWeb, teniendo en cuenta las variables definidas para la base de conocimiento.
5. Plantear el algoritmo a utilizar.
6. Aplicar el algoritmo en la base de datos de la Plataforma VideoWeb.
7. Valorar los resultados de la aplicación del algoritmo planteado.

Con el correcto cumplimiento de las tareas se espera obtener los siguientes **resultados**:

1. El algoritmo a aplicar para la obtención de preferencias de usuario.
2. Definición de la base de conocimientos a obtener con el algoritmo a aplicar.
3. Informe de aplicación del algoritmo en la Plataforma VideoWeb.

Para obtener los conocimientos necesarios que hagan posible el cumplimiento del objetivo trazado en el trabajo y cumplir las tareas propuestas, se lleva a cabo una investigación en la que se utilizan los siguientes métodos científicos:

El **Analítico-Sintético**: se utiliza para la sistematización de la información contenida, de todo el material acopiado durante la indagación, para arribar a los criterios y las conclusiones expuestos en la tesis.

El **Análisis Histórico-Lógico**: se utiliza para revelar el origen de las tecnologías de Minería de Datos, así como de los métodos de Regresión Lineal aplicados a esta, lo que posibilitó obtener referentes teóricos e históricos para determinar cuál es el más indicado en la obtención de las preferencias de usuarios, y revelar de esta manera la esencia de la investigación.

La **Modelación**: posibilita llevar a cabo, la elaboración de un modelo para la obtención de preferencias de usuario, mediante un algoritmo de la Regresión Lineal.

Capítulo 1. “Fundamentación Teórica”

1.1 Introducción

En el presente capítulo se brinda una visión general de las diferentes técnicas para el análisis de datos; enmarcándose en un contexto donde la Inteligencia Artificial y la Estadística van de la mano, a través del análisis de Regresión Lineal enfocado en la Minería de Datos. Este tema se fundamenta en el análisis de regresión tradicional de la Estadística e intenta extender su aplicación al manejo de los datos, determinando su comportamiento y centrando su atención en la Regresión Lineal Múltiple.

1.2 Conceptos asociados al dominio del problema

Para el correcto entendimiento del presente documento es necesario hacer referencia a conceptos propios de Estadística, enfocados a las Técnicas de Minería de Datos y algoritmos de Regresión Lineal.

Para determinar problemas de predicción, tanto la Inteligencia Artificial como la Estadística, tienen un modelo a desarrollar, cada una con sus propias terminologías. Esto puede llevar a una incomunicación entre ellas, siendo diversas las opiniones de conveniencia en cuanto a cual utilizar para resolver un determinado problema. Sin embargo, no es difícil encontrar áreas donde convergen estas dos disciplinas. Por ejemplo, la Minería de Datos orientada a extraer patrones predictivos ocultos en grandes bases de datos es una de estas áreas.

Dentro del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), la Minería de Datos es una de sus etapas. Esta se puede definir como un “proceso analítico diseñado para explorar grandes cantidades de datos (generalmente datos de negocio y mercado) con el objetivo de detectar patrones de comportamiento o relaciones entre las diferentes variables”. (Jaramillo, 2009) La realización de un proyecto de Minería de Datos atraviesa por diferentes fases para la obtención de conocimiento.

Fases de la Minería de Datos:

Filtrado de datos

El formato de los datos contenidos en la fuente de datos, la mayoría de las veces no es el correcto, siendo imposible utilizar algún algoritmo de minería sobre los datos iniciales sin que requieran alguna transformación.

En este paso se filtran los datos con el objetivo de eliminar valores incorrectos, no válidos o desconocidos; según las necesidades y el algoritmo a utilizar. Además, se obtienen muestras de los datos en busca de mayor velocidad y eficiencia de los algoritmos, o se reducen el número de valores posibles para los atributos de análisis.

Selección de Variables

Después de realizar la limpieza de los datos, en la mayoría de los casos se tiene una gran cantidad de variables o atributos. La selección de características reduce el tamaño de los datos sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería; seleccionando las variables más influyentes en el problema.

Algoritmos de Extracción de Conocimiento

La extracción del conocimiento es la esencia de la Minería de Datos donde mediante una técnica, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.

En general las técnicas de Minería de Datos se pueden dividir en aquellas que hacen uso de ecuaciones tales como la estadística o las redes neuronales o las que se basan en la lógica como los árboles de decisión y las reglas.

En esta fase es donde se concentra la investigación pues es donde se aplicarán técnicas para extraer el conocimiento de grandes volúmenes de información que sirvan para la personalización de la interfaz de usuario de la plataforma.

Interpretación y Evaluación

Una vez obtenido el modelo, se procede a su validación; donde se comprueba que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos para buscar el que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

1.3 Regresión Lineal en la Minería de Datos.

1.3.1 Descripción General

En la Minería de Datos confluyen varias disciplinas, en especial la Estadística, que puede constituirse en un aliado muy productivo y eficaz para los gestores de bases de datos. El campo de la Estadística tiene que ver con la recopilación, presentación, análisis y uso de datos para tomar decisiones y resolver problemas (Montgomery, y otros, 2003). Las técnicas estadísticas para el análisis, también se consideran parte de la Minería de Datos, dado que su propósito es el mismo. Uno de los enfoques en la Minería de Datos dentro de su amplia gama de usos es la predicción. En la literatura, este tipo de problema es comúnmente llamado regresión.

El concepto de regresión es uno de los pilares de la estadística, y data al menos desde principios de 1800. Es posible que el término regresión sea debido a Francis Galton, quien acuñó el término "regresión hacia la media" para describir la observación de que los hijos de padres muy altos tienden a ser algo más bajos que sus progenitores, y por el contrario los hijos de padres muy bajos suelen ser algo más altos, y por lo tanto acercarse en ambos casos más a la media de la población.

En muchas situaciones de la vida real, se presentan problemas en los cuales existe una relación entre dos o más variables y se hace necesario encontrar la naturaleza de esta relación. Es común que existan relaciones entre variables, pero cuando estas relaciones no han sido identificadas o no están completamente determinadas, toma valor el análisis de regresión. El análisis de regresión es una técnica estadística para el modelado y la investigación de la relación entre dos o más variables (Montgomery, y otros, 2003). Según Neter, el análisis de regresión sirve principalmente para tres propósitos: descripción, control y predicción (Neter, y otros, 1996).

Tipos de regresión

- ✓ **Regresión Lineal Simple**: cuando las variables X e Y se relacionan según un modelo de línea recta.
- ✓ **Regresión no Lineal o Curvilínea**: cuando las variables X e Y se relacionan según una línea curva.
- ✓ **Regresión Múltiple**: cuando se tiene más de una variable independiente (x_1, x_2, ...,x_p), y una sola variable dependiente Y.

La mayoría de los estudios conllevan la obtención de datos en un número más o menos extenso de variables. En algunos casos el análisis de dicha información se lleva a cabo centrando la atención en pequeños subconjuntos de las variables recogidas utilizando para ello análisis sencillos que involucran únicamente técnicas bivariadas. Un análisis apropiado, sin embargo, debe tener en consideración toda la información recogida o de interés, siendo necesario requerir de técnicas estadísticas múltiples más complejas. En particular, el modelo de Regresión Lineal Simple es un método sencillo para analizar la relación lineal entre dos variables cuantitativas. Sin embargo, en la mayoría de los casos lo que se pretende es predecir una respuesta en función de un conjunto más amplio de variables, siendo necesario considerar el modelo de Regresión Lineal Múltiple como una extensión de la recta de regresión que permite la inclusión de un número mayor de variables.

Regresión Lineal Múltiple

La Regresión Lineal Multivariante o Múltiple es un modelo de regresión que considera p −1 regresores o variables independientes para explicar una variable de respuesta o variable dependiente Y. La forma general del modelo de Regresión Lineal Múltiple, es (Weisberg, 2005):

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad (1)$$

Donde:

- ✓ Sean $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ los parámetros del modelo que representa los coeficientes de regresión.
- ✓ Sean $x_{i,1}, x_{i,2}, \ldots, x_{i,p-1}$ la i-ésima observación de las variables independientes explicativas.
- ✓ Sea y_i la i-ésima observación de la variable respuesta.
- ✓ Sea ε_i el i-ésimo error aleatorio.
- ✓ Sea i =1,2,..., n el índice que corresponde al conjunto de tuplas o registros en la base de datos, que sirven como muestra de aprendizaje.

El modelo de regresión se puede escribir con notación matricial, así:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \vdots & x_{n,p-1} \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (2)$$

Es común la presentación del modelo en forma resumida:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (3)$$

Donde **Y** es el vector de respuestas, **X** la matriz de constantes o variables independientes, $\boldsymbol{\beta}$ el vector de coeficientes y $\boldsymbol{\varepsilon}$ es el vector de error aleatorio.

En el modelo de Regresión Lineal Múltiple el parámetro β_0 es el intercepto (de la línea, plano, curva, hiperplano o hipercurva, dependiendo del número de variables independientes y de las posibles transformaciones); si el alcance del modelo incluye al intercepto, el parámetro β_0 es la media de la distribución de la variable de respuesta Y, cuando las demás variables explicativas son cero.

Los coeficientes de regresión $\beta_1, \beta_2, \dots, \beta_{p-1}$ miden el cambio esperado en la variable de respuesta Y, por unidad de cambio en la correspondiente variable explicativa cuando las demás se mantienen constantes. En el caso de no existir linealidad en las variables explicativas, la interpretación de los coeficientes de regresión puede ser mucho más compleja y dependerá de la forma final del modelo.

El problema central del análisis de regresión consiste en encontrar los estimadores más apropiados de los parámetros β_i, utilizando los datos observados. El científico alemán Gauss, propuso estimar los parámetros β_i minimizando la suma de los cuadrados de las desviaciones o las diferencias entre valores observados y ajustados. Este criterio para estimar los coeficientes de regresión se conoce como método de mínimos cuadrados (Montgomery, y otros, 2003).

La función de mínimos cuadrados es:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} \right) \right)^2 \quad (4)$$

Al realizar las derivadas correspondientes e igualar a cero, resulta el sistema de ecuaciones a resolver:

$$\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'Y} \quad (5)$$

La expresión (5) se conoce como las ecuaciones normales en forma matricial de las cuales se obtiene, el estimador de mínimos cuadrados de $\hat{\boldsymbol{\beta}}$ como (Montgomery, y otros, 2003):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \quad (6)$$

Coeficiente de Determinación R-Cuadrado

Se denomina coeficiente de determinación R^2 como el coeficiente que indica el porcentaje del ajuste que se ha conseguido con el modelo lineal, es decir el porcentaje de la variación de Y que se explica a través del modelo lineal estimado, por medio del comportamiento de X. A mayor porcentaje mejor es el modelo para predecir el comportamiento de la variable Y. También se puede entender este coeficiente de determinación como el porcentaje de varianza explicada por la recta de regresión y su valor siempre

estará entre 0 y 1. Es una medida de la proximidad o de ajuste de la recta de regresión a la nube de puntos. También se le denomina bondad del ajuste.

Anova, Contraste de Hipótesis

En estadística, el análisis de la varianza (*ANOVA*, según terminología inglesa) es una colección de modelos estadísticos y sus procedimientos asociados, en el cual la varianza está particionada en ciertos componentes debido a diferentes variables explicativas. Las técnicas iniciales del análisis de varianza fueron desarrolladas por el estadístico y genetista R. A. Fisher en los años 1920 y 1930, siendo algunas veces conocido como Anova de Fisher o análisis de varianza de Fisher, debido al uso de la distribución F de Fisher como parte del contraste de hipótesis.

Tabla ANOVA del modelo de Regresión Lineal Múltiple

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Varianzas
Por la recta	$scE=\sum_i (\hat{y}_i - \bar{y})_2$	k	$\hat{s}_{e^2} = \frac{scE}{K}$
Residual	$scR=\sum_i (y_i - \bar{y})_2$	n – (k+1)	$\hat{s}_{R^2} = \frac{scR}{n-(k+1)}$
Global	$scG=\sum_i (y_i - \bar{y})_2$	n - 1	$\hat{s}_{y^2} = \frac{scG}{n-1}$

Tabla 1: Anova del modelo de Regresión Lineal Múltiple.

De esta tabla de ANOVA se deduce el siguiente contraste acerca de la influencia “conjunta” del modelo de regresión en la variable respuesta.

Contraste de regresión múltiple de la F.

El contraste que se desea resolver es el siguiente:

$$C_M \equiv \left\{ \begin{array}{l} H_0 \equiv \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \\ H_0 \equiv \text{algún } \alpha_i \neq 0 \text{ para algún i} \end{array} \right\}$$ Contraste Conjunto de la F

Si H_0 es cierto ninguna de las variables regresoras influye en la variable respuesta (el modelo no influye). En este supuesto se verifica que $\hat{y}_1 \approx \bar{y} \Rightarrow scE \approx 0$, por ser ésta una medida absoluta se compara con la varianza residual, lo que lleva a utilizar como estadístico del contraste el siguiente: $\hat{F}_M = \frac{\hat{s}^2_e}{\hat{s}^2_R}$

Bajo la hipótesis nula y por la hipótesis de independencia se sigue que $\hat{F}_M$ sigue una distribución F (Contraste de la F) con k y n – (k + 1) grados de libertad, $\hat{F}_M | H_0 = \frac{\hat{s}^2_e}{\hat{s}^2_R} \sim F_{k,n-(K+1)}$. Deduciéndose que p – valor de contraste es donde $F_{k,n-(K+1)}$ denota una variable aleatoria que sigue una distribución F con k y n - (k+1) grados de libertad. El contraste de la F es unilateral (de una cola) y generaliza el contraste de regresión expuesto para el modelo de Regresión Lineal Simple.

Sí el valor crítico (p-valor) del contraste es grande (mayor que el nivel de significación $\propto$) se acepta H_0, el modelo de regresión no es influyente y debe buscarse un modelo alternativo.

Supuestos del modelo

El modelo de Regresión Lineal Múltiple requiere que se satisfagan básicamente los mismos supuestos que el modelo de Regresión Lineal Simple.

1. La relación entre las variables es lineal.
2. Los errores en la medición de las variables explicativas son independientes entre sí.
3. Los errores tienen varianza constante.
4. Los errores tienen una esperanza matemática igual a cero (los errores de una misma magnitud y distinto signo son equiprobables).
5. El error total es la suma de todos los errores.
6. Los errores siguen una distribución normal con media cero y varianza uno.

1.3.2 Descripción actual del dominio del problema.

La Plataforma VideoWeb ubicada en la Universidad de la Ciencias Informáticas (UCI) es un producto que brinda a la comunidad universitaria la posibilidad de interactuar con todo tipo de contenidos, a través de una interfaz intuitiva y amigable. Los materiales audiovisuales que en ella se presentan, están definidos de acuerdo al género al que pertenezcan y son publicados conforme al criterio del administrador del sistema. Estos pueden ser videos, películas o documentales. Además ofrece servicios de transmisión de tv y radio

en vivo. Al acceder al sitio, la aplicación no tiene funcionalidades para registrar las preferencias de los usuarios. Esto determina que la interfaz se comporte de manera estática para todos los clientes por igual.
La aplicación está implementada a partir del CMS Drupal, el cual cuenta con una serie de funcionalidades que viabilizan el almacenamiento de variables, tales como: el objeto al que se obtuvo acceso, el url que lo solicitó, la fecha y el usuario que realizó dicha acción. Sin embargo, la información que se adquiere, no se utiliza con ningún fin en específico.

1.3.3 Situación Problemática.

Para comercializar productos y servicios de software es necesario que los mismos tengan calidad y profesionalismo, pero además lo que asegura que los clientes mantengan la fidelidad, son los valores agregados que puedan tener estos productos y servicios que se comercializan.
La Plataforma VideoWeb es un producto que de manera general se dedica a la gestión y transmisión de contenidos audiovisuales a través de la red de datos. En la versión actual del producto no se personaliza la interfaz que se muestra a los usuarios y esto hace que sea estática para todos por igual. Los contenidos están organizados atendiendo a las políticas definidas por los administradores del sistema y para incorporar un valor agregado como la personalización de la información de acuerdo a las preferencias de los usuarios, es necesario analizar el comportamiento de los mismos al interactuar con la plataforma.
Para el análisis del comportamiento de los usuarios con el fin de saber sus preferencias, el registro de la información asociada a la interacción de los usuarios con la aplicación es un paso importante y esto actualmente no es soportado por la aplicación. El proceso de gestión de la información que se acumula, puede ser difícil, debido al manejo poco eficiente de tan alto volumen de información. El sistema no cuenta con un mecanismo que ayude al procesamiento de dichos datos, con el fin de darle valor a la información, de acuerdo a los gustos manifestados por el usuario en su navegación.

1.4 Análisis de soluciones existente.

Bajo el nombre de Minería de Datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligada al proceso de personalización de un sitio web, ya que resulta muy útil para aprovechar la información disponible en las bases de datos, analizando el comportamiento de los usuarios y de esta forma aumentar su fidelidad. Las técnicas de la Minería de Datos provienen de la Inteligencia Artificial y de la Estadística. Dichas técnicas,

no son más que algoritmos, que se aplican sobre un conjunto de datos para obtener resultados y dar solución a problemas como la predicción.

Las técnicas más representativas son:

Redes neuronales: Son modelos de predicción no lineales que aprenden como detectar un patrón para emparejar un perfil particular a través de un proceso de entrenamiento. Las redes neuronales son conocidas en la estructura del aprendizaje automático cómo “aproximaciones universales” con un gran carácter paralelo de cálculo y buenas capacidades de generalización, pero también como cajas negras debido a la dificultad para penetrar dentro de las relaciones aprendidas. (Arévalo, y otros)

Regresión Lineal Simple: es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables. (Montgomery, y otros, 2003)

Árboles de decisión: un árbol de decisión es un modelo de predicción utilizado en el ámbito de la Inteligencia Artificial; es una estructura en forma de árbol que visualmente describe una serie de reglas (condiciones) que causan que una decisión sea tomada. (Arévalo, y otros)

Teniendo en cuenta el principal objetivo de esta investigación que no es más que predecir el comportamiento de los usuarios, de acuerdo con sus preferencias, se determinó el uso del método estadístico Regresión Lineal Múltiple, ante cualquiera de las técnicas dentro de la Inteligencia Artificial antes mostradas. El por qué de esta selección se basa en que ninguna de estas técnicas supera la capacidad predictiva del método estadístico estudiado, mostrando ser significativamente mejor que las redes neuronales, además de que gracias al fácil acceso por parte de los usuarios a diversos software estadísticos, hoy en día resulta más económico de aplicar un modelo estadístico que las redes neuronales, en cuanto a los recursos computacionales involucrados para este tipo de análisis.
Un aspecto importante que se deberá tener en cuenta es que así como un método estadístico da una solución única ante unos mismos datos (siendo además la mejor solución posible, definida como el mejor ajuste a un modelo lineal), una red neuronal no garantiza que la solución dada sea la óptima. (Pitarque, y otros)

Otra desventaja que presentan las técnicas dentro de la Inteligencia Artificial es que en aplicaciones reales los árboles de decisión tienden a ser demasiado grandes y, por tanto, difíciles de interpretar.

Otro punto significativo que demuestra la superioridad del uso del método estadístico Regresión Lineal Múltiple, es que este permite obtener relaciones entre datos a través de un conjunto más amplio de variables, teniendo en cuenta que la Regresión Lineal Simple es insuficiente en espacios multidimensionales.

1.5 Conclusiones

En este capítulo se abordó un estudio referente a la Minería de Datos, así como los pasos a seguir para desarrollar un proyecto de este tipo, se analizaron técnicas de Inteligencia Artificial como las redes neuronales, y arboles de decisión frente a técnicas estadísticas como la Regresión Lineal Simple y Regresión Lineal Múltiple; concluyendo que la última es la más apropiada para esta investigación.

Capítulo 2. “Descripción y análisis de la solución propuesta”

2.1. Introducción

En el presente capítulo se selecciona un conjunto de variables para formar la base de conocimientos, así como los cambios necesarios para el diseño de la base de datos de la Plataforma VideoWeb. Se selecciona el algoritmo a utilizar y se abordarán brevemente algunos de los múltiples paquetes de software estadístico existentes en el mundo, que pueden satisfacer las necesidades del proyecto.

2.2. Estructura de almacenamiento de la información.

El estudio del comportamiento de los usuarios requiere el análisis de información relacionada a los mismos, y en el caso de la Plataforma VideoWeb por ser una solución basada en la web, esta información no es más que el historial de acciones dejadas durante su interacción con el sistema. Un conjunto de tablas relacionadas entre sí conforman la base de datos de la Plataforma VideoWeb. Algunas de estas tablas están relacionadas con el propio funcionamiento del CMS que soporta el sistema, y otras almacenan la información que se maneja dentro del mismo (información audiovisual, gestión de usuarios y tipologías). Para la presente investigación solo interesan el último grupo de tablas mencionadas, las cuales se listan y describen a continuación.

Listado de tablas a analizar:

- ✓ streaming.
- ✓ publicacion_am.
- ✓ filehtml.
- ✓ almacenamiento.
- ✓ archivo_multimedia.
- ✓ internos.
- ✓ tipologias_am.
- ✓ peliculas.
- ✓ videos.
- ✓ docentes.
- ✓ documentales.
- ✓ series.

- ✓ temporada.
- ✓ capitulo.
- ✓ users.

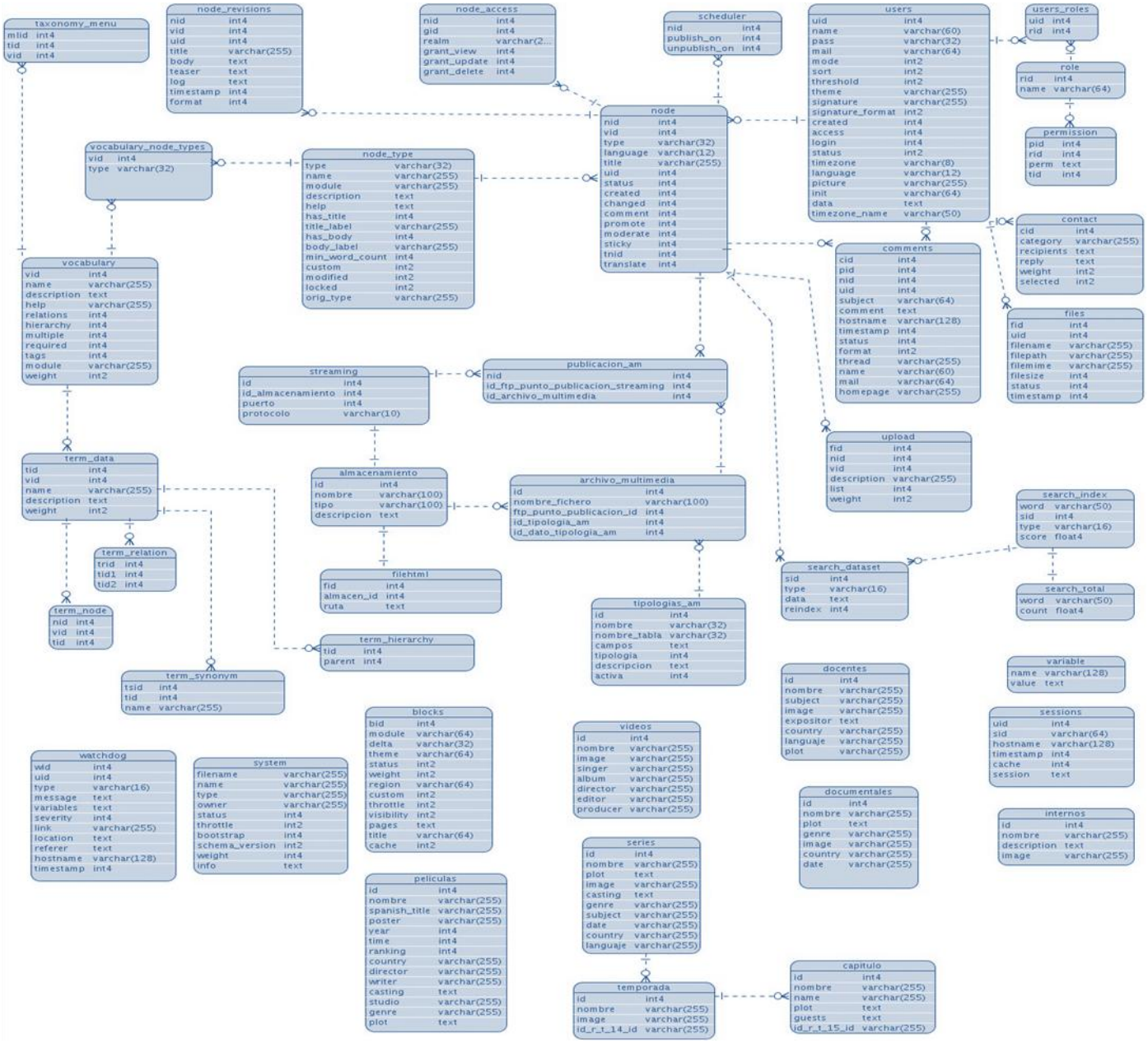

Figura 1: Estructura de almacenamiento de información del sistema.

Descripción de las tablas y atributos.

Nombre: streaming		
Descripción: En esta tabla se almacenan los datos de los puntos de publicación streaming que posee el sistema.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los puntos de publicación de streaming.
id_almacenamiento	int4	Identifica del almacenamiento que se va a utilizar como streaming.
puerto	int4	Puerto que se utiliza.
protocolo	varchar(10)	Protocolo que se utiliza.

Tabla 2: Descripción de la clase streaming.

Nombre: publicacion_am		
Descripción: En esta tabla se almacenan los datos de donde está publicado cada archivo multimedia.		
Atributo	**Tipo**	**Descripción**
nid	int4	Valor numérico único y auto incremental que identifica a cada nodo al que está asociado el archivo multimedia.
id_ftp_punto_publicacion_streaming	int4	Id de almacenamiento donde se va a publicar el archivo multimedia.
id_archivo_multimedia	int4	Id donde se va a publicar el archivo multimedia.

Tabla 3: Descripción de la clase publicacion_am.

Nombre: filehtml		
Descripción: En esta tabla se almacenan los datos del servidor de almacenamiento filehtml.		
Atributo	**Tipo**	**Descripción**
fid	int4	Identificador del servidor de almacenamiento.
almacenamiento_id	int4	Id del almacenamiento al cual él pertenece.
ruta	Int4	Ruta donde se van a almacenar los archivos multimedia.

Tabla 4: Descripción de la clase filehtml.

Nombre: almacenamiento		
Descripción: En esta tabla se almacenan los datos de los servidores de almacenamiento que utiliza la Plataforma VideoWeb.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los almacenamientos de acuerdo al nombre y tipo.
nombre	varchar(100)	Nombre que identifica cada uno de los almacenamientos
tipo	varchar(100)	Identifica qué tipo de servidor de almacenamiento se va a utilizar (ftphtml, filehtml).
descripcion	text	Descripción del almacenamiento.

Tabla 5: Descripción de la clase almacenamiento.

Nombre: archivo_multimedia		
Descripción: En esta tabla se almacenan los datos de las publicaciones de archivos multimedia que posee el sistema.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los archivos multimedia que posee el sistema.
nombre_fichero	varchar(100)	Nombre que posee la publicación para ser mostrada.
ftp_punto_publicacion_id	int4	Id del servicio de almacenamiento donde se va a guardar el archivo multimedia.
id_tipología_am	int4	Id de la tipología a la que pertenece el archivo multimedia.
id_dato_tipología_am	int4	Id del dato del archivo según la tipología.

Tabla 6: Descripción de la clase archivo_multimedia.

Nombre: internos		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como interno.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los materiales internos.
nombre	varchar(255)	Nombre del archivo multimedia interno.
description	text	Descripción del archivo multimedia interno.
image	varchar(255)	Imagen del archivo multimedia interno.

Tabla 7: Descripción de la clase internos.

Nombre: tipologias_am		
Descripción: En esta tabla se almacenan los datos de las tipologías de archivo multimedia que posee el sistema.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada una de las tipologías de archivo multimedia que posee el sistema.
nombre	varchar(32)	Nombre de la tipología.
nombre_tabla	varchar(32)	Nombre de la tabla en la base de datos.
campos	text	Arreglo con la estructura de cada uno de los campos que conforman la tipología.
tipología	int4	Identifica si es una tipología que tiene asociado video o no.
descripcion	text	Breve descripción de la tipología.
activa	int4	Identifica si fue creada o no la tipología.

Tabla 8: Descripción de la clase tipologias_am.

Nombre: peliculas		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como película.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada una de las películas.
nombre	varchar(255)	Título original de la película.
spanish_title	varchar(255)	Título en español de la película.
poster	varchar(255)	Imagen de la película.
year	int4	Año en que fue producida la película.
time	int4	Duración de la película.
ranking	int4	Valor de votación de los usuarios a la película.
country	varchar(255)	Nombre del país al que pertenece la película.
director	varchar(255)	Nombre del director de la película.
writer	varchar(255)	Nombre del escritor de la película.
casting	text	Nombre de los actores de la película.
studio	varchar(255)	Estudio donde se produjo la película.
genre	varchar(255)	Género de la película.
plot	text	Resumen de la película.

Tabla 9: Descripción de la clase películas.

Nombre: videos		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como video.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los videos.
nombre	varchar(255)	Nombre que identifica el video.
imagen	varchar(255)	Imagen del video.
singer	varchar(255)	Nombre del cantante o agrupación que interpreta la canción del video.
album	varchar(255)	Nombre del álbum del video.
director	varchar(255)	Nombre del director del video.
editor	varchar(255)	Nombre del editor del video.
producer	varchar(255)	Nombre del productor del video.

Tabla 10: Descripción de la clase videos.

Nombre: docentes		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como docente.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los materiales docentes.
nombre	varchar(255)	Nombre del archivo multimedia docente.
subject	varchar(255)	Tema del archivo multimedia docente.
image	varchar(255)	Imagen del archivo multimedia docente.
expositor	text	Ponente del material docente.
country	varchar(255)	País del archivo multimedia docente.
language	varchar(255)	Idioma.
plot	varchar(255)	Resumen del material docente.

Tabla 11: Descripción de la clase docentes.

Nombre: documentales		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como documental.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los documentales.
nombre	varchar(255)	Nombre del documental.
plot	text	Resumen del documental.
genre	varchar(255)	Género del documental.
image	varchar(255)	Imagen del documental.
country	varchar(255)	País del documental.
date	varchar(255)	Fecha.

Tabla 12: Descripción de la clase documentales.

Nombre: series		
Descripción: En esta tabla se almacenan los datos de un archivo multimedia definido como serie.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada una de las series.
nombre	varchar(255)	Nombre que identifica a la serie.
plot	text	Resumen de la serie.
imagen	varchar(255)	Imagen de la serie.
casting	text	Nombre de los actores que interpretan la serie.
genre	varchar(255)	Género de la serie.
subject	varchar(255)	Tema de la serie.
date	varchar(255)	Fecha.
country	varchar(255)	Nombre del país de la serie.
language	varchar(255)	Idioma.

Tabla 13: Descripción de la clase series.

Nombre: temporada		
Descripción: En esta tabla se almacenan los datos de la temporada de una serie.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada de las temporadas de la serie.
nombre	varchar(255)	Nombre que identifica a la temporada de la serie.
imagen	varchar(255)	Imagen de la temporada de la serie.
id_r_t_14_id	varchar(255)	Id de la serie a la que pertenece la temporada.

Tabla 14: Descripción de la clase temporada.

Nombre: capitulo		
Descripción: En esta tabla se almacenan los datos de un capítulo de una serie.		
Atributo	**Tipo**	**Descripción**
id	int4	Valor numérico único y auto incremental que identifica a cada uno de los capítulos de la temporada de una serie.
nombre	varchar(255)	Nombre que identifica al capítulo.
imagen	varchar(255)	Imagen que identifica al capítulo.
id_r_t_14_id	varchar(255)	Id de la temporada a la que pertenece el capítulo.

Tabla 15: Descripción de la clase capítulo.

Nombre: users		
Descripción: En esta tabla se los datos de los usuarios registrados.		
Atributo	**Tipo**	**Descripción**
uid	int4	Valor numérico único y auto incremental que identifica a cada uno de los usuarios.
name	varchar(60)	Nombre único del usuario.
pass	varchar(32)	Contraseña del usuario.
mail	varchar(64)	Dirección de correo del usuario.
mode	int2	Modo de visualización de los comentarios para cada usuario.
sort	int2	Forma de ordenar los comentarios por fecha.
threshold	int2	Anteriormente usado para las preferencias de usuario, ya no se usa.
theme	varchar(255)	Tema por defecto.
signature	varchar(255)	Firma de usuario.
signature_format	int2	Formato de la firma del usuario.
created	int4	Marca de tiempo de la creación del usuario.
acess	int4	Marca de tiempo para el acceso anterior

		del usuario al sitio.
login	int4	Marca de tiempo para el último acceso del usuario.
status	int2	Valor que muestra si el usuario está activo (1) o bloqueado (0).
timezone	varchar(8)	Zona horaria del usuario.
language	varchar(12)	Lenguaje por defecto del usuario.
picture	varchar(255)	Dirección de la imagen cargada por el usuario.
init	varchar(64)	Dirección de correo usado para la creación inicial de una cuenta.
data	text	Arreglo de nombre y valores relacionados con el usuario. No se recomienda el uso de este campo.
timezone_name	varchar(50)	Nombre de la zona horaria.

Tabla 16: Descripción de la clase users.

Luego de analizar las tablas que conforman actualmente la base de datos del sistema, se concluyó que cada una de ellas tiene un objetivo específico dentro del funcionamiento de la aplicación y los datos que guardan tienen un fin determinado. Ninguna de estas tablas está preparada para almacenar el historial de navegación que deja un usuario cuando visita el sitio, tampoco a través de relaciones entre ellas es posible obtener información relevante para la obtención de preferencias.

Guardar el historial completo de navegación de un usuario no sería aún la solución, es preciso filtrar más esa información de forma tal que solo se guarden los datos que sean relevantes y que luego son analizados para personalizar. El filtrado de esa información es referente a las variables que luego se convertirán en preferencias de un usuario, estas se definirán más adelante en la investigación.

2.3. Definición de las variables para formar la base de conocimientos.

Una aplicación Web debe estar centrada en los usuarios, reflejando la satisfacción del cliente. Un enfoque de estas características, por medio de una base de conocimiento, suministrará servicios eficaces a diferentes tipos de usuarios en una variedad de dominios de aplicaciones, cambiando la experiencia de los clientes y la accesibilidad de los mismos. Estos cambios permitirán que las aplicaciones web sean

utilizadas por cualquier usuario, donde se brinden servicios basados en las preferencias de los mismos. La relación entre el usuario y la información presentada es vital para el éxito de cualquier aplicación.
Una base de conocimiento provee una alternativa excelente para aplicaciones como la que se desarrolla, donde la atención al cliente, la adaptabilidad y la posibilidad de ajustarse a distintas clases de usuario son la prioridad. El más importante aspecto de una base de conocimiento es la calidad de la información que esta contiene, de ahí la necesidad de definir una serie de variables, donde queden recogidos los hábitos o gustos de cada usuario. Para poder personalizar, las variables antes mencionadas, se clasifican en tres tipos: variables demográficas, de sesión y de operaciones, logrando así una visión completa del cliente.

1. **Las variables demográficas:** representan información relacionada con características propias del cliente como persona física, tales como su nombre, apellidos, teléfono, fecha de nacimiento, sexo, entre otras.
2. **Las variables de sesión:** representan información sobre cómo interactúa el usuario con la web. Con ellas será posible conocer hábitos y preferencias de los clientes en base a la utilización que hace de los contenidos, los servicios y la interfaz de usuario.
3. **Las variables de operaciones:** representan información sobre el conjunto de operaciones a nivel de transacción que el cliente ha efectuado durante su navegación. (IWorld.com, 2003)

La información suministrada por este conjunto de variables es fundamental a la hora de diseñar esquemas de personalización y programas basados en los hábitos o preferencias del usuario. Según las clasificaciones antes nombradas las variables seleccionadas están divididas de acuerdo a la tipología a la que pertenecen y las cuales se listan a continuación:

Tipologías				
películas	**docentes**	**videos**	**documentales**	**series**
género	exponente	cantante	género	reparto
director	país	álbum	país	género
reparto	idioma	editor		país
escritor		productor		idioma
país				

Tabla 17: Variables para la base de conocimiento.

2.4. Cambios necesarios para el diseño en la base de datos.

Después de examinar la estructura actual que presenta la base de datos del sistema, se concluyó que no era suficiente para almacenar los datos necesarios para hacer el tipo de análisis que se pretende en la investigación. Es necesario proponer una serie de cambios con vistas a solucionar la insuficiencia identificada. A continuación se muestra y describe la propuesta de solución.

Listado de tablas a incluir:

- ✓ campo
- ✓ campo_valor
- ✓ usuario_campo
- ✓ historial

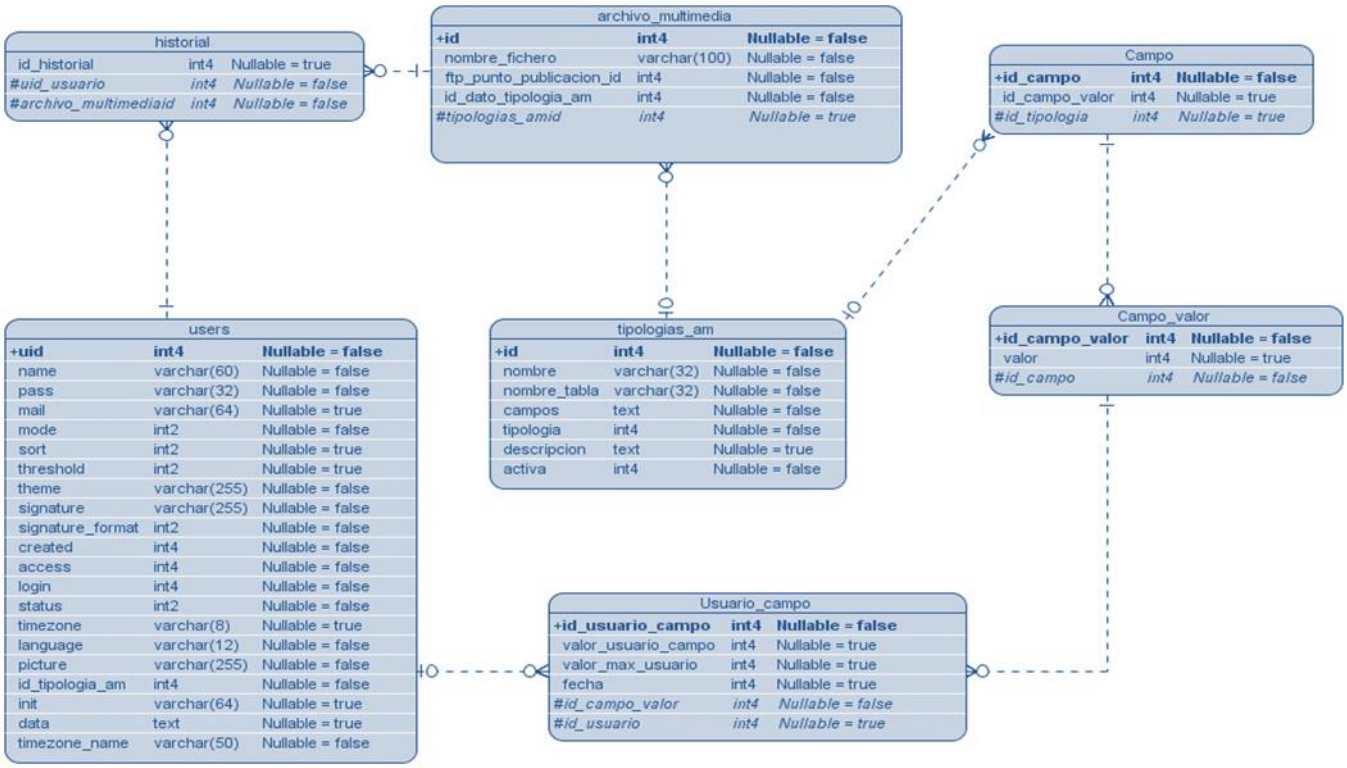

Figura 2: Estructura de almacenamiento de información del sistema.

Descripción de las tablas y atributos.

Nombre: campo		
Descripción: En esta tabla se almacenan los campos relevantes en la obtención de preferencias de usuario.		
Atributo	**Tipo**	**Descripción**
id_campo	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
id_tipologia	int4	Identificador de la tipología de archivo multimedia.
id_campo_tipologia	int4	Identificador del campo dentro de la tipología de archivo multimedia.

Tabla 18: Descripción de la clase campo.

Nombre: campo_valor		
Descripción: En esta tabla se almacenan los valores que toman los campos relevantes en la obtención de preferencias de usuario.		
Atributo	**Tipo**	**Descripción**
id_campo_valor	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
id_campo	int4	Identificador de los campos relevantes para la obtención de preferencias.
valor	int4	Valor a tomar por el campo para obtener las preferencias.

Tabla 19: Descripción de la clase campo_valor.

Nombre: usuario_campo		
Descripción: En esta tabla se almacenan la relación del usuario con los valores que toman los campos relevantes en la obtención de preferencias.		
Atributo	**Tipo**	**Descripción**
id_usuario_campo	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
id_usuario	int4	Identificador de los usuarios.
id_campo_valor	int4	Identificador del valor que toman los campos relevantes para la obtención de preferencias de usuarios.
valor_usuario_campo	int4	Valor de preferencia del usuario por el campo.
valor_max_usuario	int4	Valor máximo histórico del usuario por el campo.
fecha	int4	Fecha en que se registra.

Tabla 20: Descripción de la clase usuario_campo.

Nombre: historial		
Descripción: En esta tabla se almacenan los historiales de acceso a los archivos multimedia.		
Atributo	**Tipo**	**Descripción**
id_historial	int4	Valor numérico único y auto incremental que identifica a cada uno de los campos de esta tabla.
uid_usuario	int4	Identificador del usuario.
archivo_multimediaid	int4	Identificador del archivo multimedia.

Tabla 21: Descripción de la clase historial.

Se han agregado una serie de tablas a la base de datos del sistema como propuesta de solución a las deficiencias encontradas en el análisis realizado del modelo de datos inicial de la Plataforma VideoWeb. A partir de este momento se cuenta con las condiciones necesarias para guardar la información de los usuarios para la obtención de sus preferencias una vez aplicado un algoritmo que analice dicha información. A continuación se plantea qué algoritmo utilizar.

2.5. Selección del algoritmo.

La velocidad con la que se almacenan los datos es muy superior a la velocidad con la que estos son analizados. Para obviar este problema se necesitan soluciones de investigación tendientes a minar grandes y masivas bases de datos, desarrollar algoritmos y sistemas para profundizar nuevos tipos de datos y mejorar la utilización de los mismos en pos del cliente. Un análisis general de las diferentes técnicas para el análisis de datos permitió definir como propuesta a utilizar el método de Regresión Lineal Múltiple.

Los métodos multivariados son extraordinariamente útiles para ayudar a los investigadores a hacer que tengan sentido conjuntos grandes de datos, que constan de una gran cantidad de variables a analizar. La propuesta de Regresión Lineal Múltiple tiene como objetivo primario resumir grandes cantidades de datos por medio de relativamente pocos parámetros.

Un estudio de la regresión permite averiguar hasta que punto una variable puede ser prevista conociendo otra. Un paso importante en la construcción de dicho modelo de regresión, es el de la elección de variables a incluir y cuáles no. Para elegirlas se pueden utilizar algunos de los siguientes algoritmos:

Selección hacia adelante

Comienza con un modelo vacío (o variables preseleccionadas) y añade aquella variable al modelo que optimice un criterio, y continua añadiendo variables hasta que una regla de parada se active.

Eliminación hacia atrás

Comienza con el modelo completo (o con variables preseleccionadas) y elimina aquella variable del modelo que optimice un criterio, continua eliminando variables hasta que una regla de parada se active.

Regresión por pasos

Combina las estrategias anteriores, la decisión sobre si una variable será añadida o eliminada (y que variable) se basa sobre un criterio. (Varmuza, y otros, 2009)

Luego del correspondiente análisis de cada una de las técnicas antes mencionadas para la selección de variables y la construcción de un mejor modelo a partir de su selección, se determinó el uso del algoritmo *Stepwise* o como también es nombrado Regresión por pasos. La decisión de aplicar dicho algoritmo se basa principalmente en la característica propia de su nombre, donde en cada paso se introducen o eliminan variables de acuerdo al criterio tomado, teniendo en cuenta que dicho algoritmo es una combinación de las técnicas de Selección hacia adelante y de Eliminación hacia atrás. Permite además la posibilidad de arrepentirse de decisiones tomadas en pasos anteriores. Esta selección es la que más se ajusta a las condiciones del proyecto y al objetivo de la investigación.

2.6. Herramientas para el análisis de datos.

El análisis de regresión es uno de los métodos estadísticos más empleados en varias disciplinas, siendo necesario el uso de diversas herramientas estadísticas para su estudio. El uso de un programa de computación estadístico es importante tanto en la ciencia básica como en la aplicada y tiene como objetivo el análisis estadístico de los datos. En la práctica, tanto investigadores como profesionales emplean algún programa estadístico para realizar pruebas de hipótesis, ajustes de modelos y análisis de diseños experimentales complejos. En consecuencia, este epígrafe centrará su atención a una breve descripción de algunos de los múltiples paquetes de software estadísticos existentes en el mundo.

SPSS

SPSS (*SPSS Inc. 2007*) es un software lanzado al mercado en 1968. Originalmente se desarrolló para las ciencias sociales, por lo que ofrece un uso sencillo de las opciones, acceso rápido a datos y procedimientos, generación de salidas y gráficos. SPPS es un programa con una interfaz gráfica de usuario (término denominado en computación, “GUI”) amigable, y sólo a través de esta se accede a sus opciones (abrir los datos y ejecutar cálculos), mediante el uso de los botones de la interfaz gráfica. (Salas, 2008) SPSS es el más fácil de aprender para los investigadores principiantes y tiene un manual que explica la filosofía y los mecanismos de las técnicas estadísticas.

SAS

SAS (*SAS Institute Inc. 2007*) ha sido por largos años el software más utilizado en la comunidad estadística y, por lo tanto, también se ha propagado su uso entre investigadores de diferentes disciplinas. SAS, a diferencia de SPSS, es un programa que requiere el ingreso de comandos (sintaxis) para ejecutar

gran parte de sus rutinas y opciones. Por lo tanto, necesita del conocimiento de la sintaxis antes de su uso. SAS ha llegado a ser el programa estándar empleado en ensayos clínicos y por la industria farmacéutica en los Estados Unidos. (Salas, 2008) Ofrece la mayor flexibilidad para personalizar el manejo y análisis de datos, sin embargo su principal inconveniente es que no resulta fácil aprender a usarlo.

R

R es un lenguaje de programación y un entorno para análisis estadístico. Fue inicialmente escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda. Actualmente es el resultado de un esfuerzo de colaboración de personas del todo el mundo. El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL, y puede ser instalado tanto en sistemas operativos tipo Windows como en Linux o MacOS X. (Salas, 2008)
Existen varias interfaces gráficas muy amigables para el usuario, entre las que cabe destacar R-Commander. Posee tratamiento gráfico bastante complejo y operaciones con matrices, por lo que puede ser sustituto de matlab en este aspecto. Es capaz de importar bases de datos en distintos formatos: xls, stata (dta), csv, text, entre otros.

Es sabido que existen en el mercado distintos paquetes de software estadístico propietarios, como SPSS, Statgraphics, Minitab, Statistica, SAS, S-Plus, que cubren todas las necesidades de cualquier usuario de técnicas estadísticas, básicas o avanzadas. Frente a estas opciones, la decisión de utilizar R, es debido a su constante desarrollo, enorme flexibilidad y distribución gratuita, siendo actualmente el resultado del esfuerzo de colaboración de personas de todo el mundo. El paquete estadístico R es uno de los más potentes y profesionales que existen actualmente para realizar tareas estadísticas de todo tipo, desde las más elementales, hasta las más avanzadas. Además de adaptarse a las nuevas exigencias estadísticas y econométricas del mundo académico, considerándose mucho más rápido que los software de pagos. Cuenta, además, con la ventaja de ser gratuito y de descarga e instalación sencillas. Estas razones garantizan un resultado de importancia para el proyecto porque además de que no es necesario pagar por la utilización del software, la curva de aprendizaje de este ayudante es muy baja con lo cual se ahorra en tiempo.

2.7. Conclusiones

En este capítulo se describió la estructura actual de almacenamiento de información del sistema, concluyendo que la misma no está preparada, para almacenar todo el historial de navegación que se genera a partir de la interacción de los usuarios con la plataforma. Se realizó una propuesta de cambios, con el fin de soportar las preferencias obtenidas de los usuarios en su navegación, y de esta forma eliminar la insuficiencia identificada.

Luego del análisis de los algoritmos existentes dentro de la técnica de Regresión Lineal Múltiple, se determinó el uso del algoritmo Regresión por pasos, para la selección de las variables y construcción de un mejor modelo de predicción. Se propuso el uso del software estadístico R, para un mejor manejo de los datos e implementar el algoritmo seleccionado.

Capítulo 3. “Análisis de resultado”

3.1 Introducción

En el presente capítulo se muestran los resultados obtenidos en consecuencia de la aplicación del modelo matemático seleccionado, a un conjunto de datos de muestra, con el objetivo de determinar con cierto grado de precisión las preferencias de los usuarios en su interacción con la Plataforma VideoWeb. Se realizaron pruebas estadísticas que arrojaron resultados satisfactorios poniendo de manifiesto la bondad de los modelos.

3.2 Caso de estudio.

Para el análisis de los datos y la valoración de los resultados se propuso la utilización de un caso de estudio, a continuación se presenta el mismo.

Diseño del modelo de predicción

Teniendo en cuenta un posible consumo por parte de los usuarios de materiales audiovisuales publicados en la Plataforma VideoWeb, se diseñó el presente caso de estudio con el objetivo de ofrecer una herramienta matemática que haga posible representar la probabilidad de preferencia de un género determinado, dentro de la tipología películas, para cada usuario en particular. Los datos utilizados para este caso de estudio fueron generados por la herramienta *Data Generator for Postgre SQL*, ya que el proyecto hasta el momento no cuenta con los datos reales necesarios para probar este tipo de análisis. La tabla obtenida almacena una muestra de 15 usuarios, valor obtenido de la muestra general de integrantes pertenecientes al proyecto, que suman un total de 27 usuarios, se guarda además el número de películas visitadas por género, durante 1 mes. Se consideró como variable dependiente aquel género que mantuviera una correlación positiva con el resto de los géneros en estudio. Este análisis se realizó a partir de la matriz de correlación, las variables a tener en cuenta para dicho análisis se enumeran a continuación:

- ✓ ID: id del usuario
- ✓ A: Acción
- ✓ T: Terror
- ✓ Av: Aventura
- ✓ D: Drama
- ✓ C: Comedia
- ✓ S: Suspenso

Análisis estadístico

Para llevar a cabo un análisis de regresión es necesario que los datos tomados como muestra estén correlacionados, es decir que la variable respuesta dependa linealmente de las variables explicativas, y de esta forma a través de la matriz de correlación poder determinar la relación que existe entre cada uno de los géneros existentes en el modelo.

La matriz de correlaciones ayuda a identificar correlaciones lineales entre pares de variables. Cuanto más extremo sea el coeficiente, mejor asociación lineal existe entre el par de variables, cuando es cercano a cero no. A continuación se comprueba la dependencia lineal de las variables.

	drama x_1	**acción** x_2	**comedia** x_3	**terror** x_4	**suspenso** x_5	**aventura** x_6
x_1	1.0000000	-0.6170457	0.90492774	-0.42403562	-0.3864323	0.8968542
x_2	-0.6170457	1.0000000	-0.33327590	0.95023513	0.8700936	-0.4630765
x_3	0.9049277	-0.3332759	1.00000000	-0.09061804	-0.1699561	0.8581194
x_4	-0.4240356	0.9502351	-0.09061804	1.00000000	0.8476790	-0.3024968
x_5	-0.3864323	0.8700936	-0.16995612	0.84767899	1.0000000	-0.1285660
x_6	0.8968542	-0.4630765	0.85811943	-0.30249678	-0.1285660	1.0000000

Tabla 22: Matriz de correlaciones.

De acuerdo con lo anteriormente planteado y a partir de la relación que existe entre cada una de las variables mostradas, con respecto a la matriz de correlación, los modelos de regresión obtenidos son los siguientes:

1. $D = \beta_{0D} + \beta_{2D}A + \beta_{3D}C + \beta_{6D}Av + \varepsilon_D$
2. $A = \beta_{0A} + \beta_{1A}D + \beta_{4A}T + \beta_{5A}S + \varepsilon_A$
3. $C = \beta_{0C} + \beta_{1C}D + \beta_{6C}A + \varepsilon_C$
4. $T = \beta_{0T} + \beta_{2T}A + \beta_{5T} + \varepsilon_T$
5. $S = \beta_{0S} + \beta_{2S}A + \beta_{4S} + \varepsilon_T$
6. $Av = \beta_{0Av} + \beta_{1Av}D + \beta_{3Av} + \varepsilon_{Av}$

Una vez que ya se haya analizado el carácter e intensidad de la relación entre las variables, se puede proceder a estimar los parámetros de la ecuación de predicción. El criterio para obtener los coeficientes de regresión β_0, β_1, β_2,…, β_i es el de mínimos cuadrados. Este consiste en minimizar la suma de los cuadrados de los residuos de tal manera que la recta de regresión que se defina sea la que más se acerca a la nube de puntos observados y, en consecuencia, la que mejor los representa.

Ya conocido los modelo, se deben efectuar una serie de pruebas, para analizar la “bondad” de los mismos y así determinar si efectivamente pueden ser utilizados para modelar el agrado de un género determinado para cada usuario en particular.

Teniendo en cuenta que el análisis realizado solo está asociado a una muestra de 15 usuarios, se realiza además, la prueba no paramétrica *Shapiro-Wilk*, con el objetivo de comprobar la distribución normal (media 0, varianza 1) de los coeficientes asociados a cada uno de los modelos en estudio.

A continuación se muestra la información obtenida de la salida del software R para cada uno de los modelos.

El primer modelo presenta las siguientes características:

- ✓ El p-valor para la prueba F es de 3.154e-07, menor que el nivel de significación que es 0.05.
- ✓ El R cuadrado múltiple es de 0.9453, lo que está indicando que el modelo explica el 95% de la variabilidad de los datos, lo que lo hace muy bueno.
- ✓ Los residuos son relativamente pequeños siendo el primer cuartíl -0.61105 y el tercer cuartíl 0.75940, lo que indica que el modelo hasta el momento es adecuado.
- ✓ Los coeficientes β_2 y β_3 son significativos pues el *p-valor* de las pruebas *t* son menores que el nivel de significación que es 0.05 mientras que el coeficiente β_6 y el intercepto no lo son.
- ✓ No se viola ningún supuesto del modelo; aceptándose la hipótesis nula de que los residuos siguen una distribución normal.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.08341 -0.61105  0.04479  0.75940  0.80956

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.14922    0.77722   1.479  0.16729
misdatos[, 2] -0.28538    0.07526  -3.792  0.00298 **
misdatos[, 3]  0.76685    0.18110   4.234  0.00140 **
misdatos[, 6]  0.32348    0.18965   1.706  0.11610
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.799 on 11 degrees of freedom
Multiple R-squared: 0.9453,     Adjusted R-squared: 0.9304
F-statistic: 63.37 on 3 and 11 DF,  p-value: 3.154e-07

        Shapiro-Wilk normality test

data:  regresion$res
W = 0.8835, p-value = 0.05353
```

Figura 3: Gráficos de diagnosis del primer modelo.

De acuerdo con los parámetros estimados (figura 3), la ecuación de Regresión Lineal Múltiple ajustada por el método de mínimos cuadrados es:

$$D = -0.28538\,A + 0.76685\,C$$

Donde, manteniendo constante la variable Comedia (C), un incremento en el agrado de los usuarios por la variable Acción (A), implica la disminución del gusto de los usuarios a las películas del género Drama. En forma similar, manteniendo constante la variable Acción, un incremento del agrado de los usuarios por la variable Comedia, es acompañado por un incremento de la preferencia de los usuarios hacia el género Drama.

El segundo modelo presenta las siguientes características:

- ✓ El p-valor para la prueba F es de 1.040e-08, menor que el nivel de significación que es 0.05.
- ✓ El R cuadrado múltiple es de 0.9706, lo que está indicando que el modelo explica el 97% de la variabilidad de los datos, lo cual lo hace muy bueno.
- ✓ Los residuos son relativamente pequeños siendo el primer cuartíl -0.22411 y el tercer cuartíl 0.33884, lo que indica que el modelo hasta el momento es adecuado.
- ✓ Los coeficientes β_1 y β_4 son significativos pues el *p-valor* de las pruebas *t* son menores que el nivel de significación que es 0.05, mientras que el coeficiente β_5 y el intercepto no lo son.
- ✓ No se viola ningún supuesto del modelo; aceptándose la hipótesis nula de que los residuos siguen una distribución normal.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.32105 -0.22411 -0.03664  0.33884  0.90131

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.7326     0.6247   1.173 0.265678
misdatos[, 1]  -0.2715     0.0610  -4.451 0.000976 ***
misdatos[, 4]   0.8675     0.1289   6.731 3.24e-05 ***
misdatos[, 5]   0.2891     0.1374   2.104 0.059237 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.625 on 11 degrees of freedom
Multiple R-squared: 0.9706,     Adjusted R-squared: 0.9626
F-statistic: 121.3 on 3 and 11 DF,  p-value: 1.040e-08


        Shapiro-Wilk normality test

data:  regresion$res
W = 0.9497, p-value = 0.5197
```

Figura 4: Gráficos de diagnosis del segundo modelo.

De acuerdo con los parámetros estimados (figura 4), la ecuación de Regresión Lineal Múltiple ajustada por el método de mínimos cuadrados es:

$$A = -0.2715\ D + 0.8675\ T$$

Donde, manteniendo constante la variable Terror (T), un incremento en el agrado de los usuarios por la variable Drama (D), implica la disminución del gusto de los usuarios a las películas del género Acción. En forma similar, manteniendo constante la variable Drama, un incremento del agrado de los usuarios por la variable Terror, es acompañado por un incremento de la preferencia de los usuarios hacia el género Acción.

El tercer modelo presenta las siguientes características:

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.1737 -1.0596  0.2965  0.7638  1.3541

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.5328     0.5876   2.609   0.0229 *
misdatos[, 1]  0.5298     0.2061   2.570   0.0245 *
misdatos[, 6]  0.2341     0.2649   0.884   0.3942
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.033 on 12 degrees of freedom
Multiple R-squared:  0.83,      Adjusted R-squared: 0.8016
F-statistic: 29.29 on 2 and 12 DF,  p-value: 2.417e-05

        Shapiro-Wilk normality test

data:  regresion$res
W = 0.8798, p-value = 0.04714
```

Figura 5: Gráficos de diagnosis del tercer modelo

De acuerdo con los parámetros estimados (figura 5), la ecuación de Regresión Lineal Múltiple obtenida no se tomará en cuenta, al no ser influyente en la predicción, ya que el *p-valor* de la prueba de Shapiro-Wilk es menor que el nivel de significación que es 0.05, aceptándose la hipótesis alternativa de que los residuos no siguen una distribución normal, lo que implicaría una violación a este supuesto.

El cuarto modelo presenta las siguientes características:

- ✓ El p-valor para la prueba F es de 7.471e-07, menor que el nivel de significación que es 0.05.
- ✓ El R cuadrado múltiple es de 0.9047, lo que está indicando que el modelo explica el 90% de la variabilidad de los datos, lo cual lo hace aceptable.
- ✓ Los residuos son relativamente pequeños siendo el primer cuartíl -0.4648 y el tercer cuartíl 0.6860, lo que indica que el modelo hasta el momento es adecuado.
- ✓ El intercepto, al igual que el coeficiente β_2 son significativos pues el *p-valor* de las pruebas *t* son menores que el nivel de significación que es 0.05, mientras que el coeficiente β_5 no lo es.
- ✓ No se viola ningún supuesto del modelo; aceptándose la hipótesis nula de que los residuos siguen una distribución normal.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.2207 -0.4648 -0.2925  0.6860  1.3821

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.33557    0.49075   2.721 0.018557 *
misdatos[, 2] 0.67460    0.13929   4.843 0.000403 ***
misdatos[, 5] 0.09335    0.19628   0.476 0.642902
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8307 on 12 degrees of freedom
Multiple R-squared: 0.9047,     Adjusted R-squared: 0.8889
F-statistic: 56.99 on 2 and 12 DF,  p-value: 7.471e-07


        Shapiro-Wilk normality test

data:  regresion$res
W = 0.9517, p-value = 0.5523
```

Figura 6: Gráficos de diagnosis del cuarto modelo

De acuerdo con los parámetros estimados (figura 6), la ecuación de Regresión Lineal Múltiple ajustada por el método de mínimos cuadrados es:

$$T = 1.33557 + 0.67460\ A$$

Donde, un incremento en el agrado de los usuarios por la variable Acción (A), es acompañado por un incremento de la preferencia de los usuarios hacia el género Terror (T).

El quinto modelo presenta las siguientes características:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.5173  -0.9136  -0.1995   0.4594   2.0980

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.0332     0.8590   1.203    0.252
misdatos[, 2]    0.4724     0.3211   1.471    0.167
misdatos[, 4]    0.1982     0.4167   0.476    0.643

Residual standard error: 1.21 on 12 degrees of freedom
Multiple R-squared: 0.7616,     Adjusted R-squared: 0.7218
F-statistic: 19.16 on 2 and 12 DF,  p-value: 0.0001838


        Shapiro-Wilk normality test

data:  regresion$res
W = 0.9096, p-value = 0.1336
```

Figura 7: Gráficos de diagnosis del quinto modelo

De la salida del R (figura 7) se puede observar que el intercepto, al igual que los coeficientes que intervienen en el modelo no son significativos, pues los *p-valores* de las pruebas *t,* alcanzan valores superiores al nivel de significación que es de 0.05, lo que indica que el modelo no puede ser utilizable.

El sexto modelo presenta las siguientes características:

- ✓ El p-valor para la prueba F es de 3.843e-05, menor que el nivel de significación que es 0.05.
- ✓ El R cuadrado múltiple es de 0.8163, lo que está indicando que el modelo explica el 82% de la variabilidad de los datos, lo cual lo hace aceptable.
- ✓ Los residuos son relativamente pequeños siendo el primer cuartíl -0.62888 y el tercer cuartíl 0.31335, lo que indica que el modelo hasta el momento es adecuado.

- ✓ El coeficiente β_1 es significativo pues el *p-valor* de la prueba *t* es menor que el nivel de significación que es 0.05, mientras que el intercepto y coeficiente β_3 no lo son.
- ✓ No se viola ningún supuesto del modelo; aceptándose la hipótesis nula de que los residuos siguen una distribución normal.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.23739 -0.62888 -0.09309  0.31335  2.57434

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.3868     0.7685   0.503   0.6239
misdatos[, 1]   0.5169     0.2262   2.285   0.0413 *
misdatos[, 3]   0.2610     0.2953   0.884   0.3942
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.091 on 12 degrees of freedom
Multiple R-squared: 0.8163,     Adjusted R-squared: 0.7857
F-statistic: 26.66 on 2 and 12 DF,  p-value: 3.843e-05

        Shapiro-Wilk normality test

data:  regresion$res
W = 0.9053, p-value = 0.1147
```

Figura 8: Gráficos de diagnosis del sexto modelo

De acuerdo con los parámetros estimados (figura 8), la ecuación de Regresión Lineal Múltiple ajustada por el método de mínimos cuadrados es:

$$Av = 0.5169\ D$$

Donde, un incremento en el agrado de los usuarios por la variable Drama (D), es acompañado por un incremento de la preferencia de los usuarios hacia el género Aventura.

3.3 Conclusiones.

A partir del resultado del la aplicación del método matemático estudiado de Regresión Lineal Múltiple, a un conjunto de datos de prueba seleccionados, con el objetivo de obtener una fórmula matemática, que sirva para calcular las preferencias de un usuario determinado, en su interacción con la Plataforma VideoWeb, se puede concluir que la mayor parte de los modelos matemáticos obtenidos se ajustan a las condiciones necesarias que debe cumplir este tipo de predicción, poniéndose de manifiesto la bondad de ajuste de los modelos y de esta forma poder ampliar el campo de acción a otros modelos predictivos existentes en la plataforma.

Conclusiones Generales

Se alcanzaron las siguientes conclusiones como resultado del proyecto de investigación.

- ✓ Se presentó una visión general acerca del estado del arte de la Minería de Datos y su aplicación a problemas de predicción, enfocados a los sistemas de uso web, con el fin de personalizar los servicios brindados por la Plataforma VideoWeb.
- ✓ Se resaltó la importancia y el gran potencial que tiene el uso de la técnica estadística de Regresión Lineal Múltiple, para el descubrimiento de patrones ocultos en los datos, lo cual posibilita llevar hasta el nivel individual el servicio brindado por el sistema, para lograr que el cliente esté satisfecho con la información que consume.
- ✓ Se llevó a cabo un estudio de la estructura de almacenamiento de información disponible actualmente en la Plataforma VideoWeb, concluyendo que la misma era deficiente, al no contar con las condiciones necesarias que permitan el almacenamiento del conocimiento preciso para llevar a cabo un proceso de personalización. Con el fin de eliminar la insuficiencia identificada se realizó una propuesta de cambios.
- ✓ Con el estudio de algunas de las herramientas estadísticas existentes en el mundo para un mejor manejo de los datos, quedó demostrado que el software R es el más adecuado para la implementación del algoritmo seleccionado.
- ✓ A partir del resultado obtenido de la aplicación del método matemático estudiado, se comprobó la validez del trabajo realizado, obteniendo de los usuarios la información que más relación tenga con los gustos demostrados en sus visitas.
- ✓ Con el análisis del caso de estudio mostrado, se comprobó que no todos los modelos obtenidos, cumplen los supuestos necesarios para poder representar una predicción futura, de ahí que se recomienden otras técnicas de procesamiento de datos.

Recomendaciones

Al término de la presente investigación, las recomendaciones que puedan ser tomadas en cuenta para dar continuidad a este proyecto son las siguientes:

- ✓ Realizar la investigación con datos reales tomados del proyecto a través de encuestas.
- ✓ Extender este estudio a otros niveles de predicción existentes en la Plataforma VideoWeb, hace de este sistema un producto mucho más completo, lo que garantizaría una gran oportunidad de ventas, para las entidades que utilicen la aplicación con fines comerciales, pagar por ver.
- ✓ La incorporación de otros criterios para la selección de preferencias de usuario, beneficiará aún más la calidad del servicio brindado por el sistema a los clientes, ya que la investigación hasta el momento solo está guiando su análisis en base a la utilización que hacen los clientes de cada uno de los contenidos ofrecidos, obviando de esta forma otros conceptos y clasificaciones que engloba el proceso de personalización. Pudiendo asimismo ofrecer un trato personalizado e individualizado, refiriéndose al conocimiento expreso por el propio cliente como respuestas a formularios, perfiles de usuario y comportamiento histórico en sesiones. Resulta de vital importancia obtener el máximo de información posible de la forma menos intrusiva permisible, haciéndole ver al usuario que está siendo beneficiado con un servicio de personalización mucho más potente.
- ✓ Resolver el problema de obtención de la preferencia de géneros de películas usando modelos de ecuaciones estructurales, en particular la modelación de sendero por mínimos cuadrados parciales.

Glosario

Base de datos: son programas que administran información, con ellas se puede organizar y reorganizar los datos, además de que facilita su búsqueda.

Una **base de conocimiento** permite organizar, capturar, almacenar y administrar la información sobre un ámbito definido, de tal forma que se pueda hacer un mejor uso de la información dentro de una organización o sistema. (Merino, 2008)

Las siglas **CMS** (del término inglés *Content Management Systems*) identifican un conjunto de programas informáticos destinados a gestionar la presentación de los contenidos de una sede Web.

Correlación: es una medida de la relación entre dos o más variables. La correlación puede tomar valores entre –1 y +1. El valor de –1 representa una correlación negativa perfecta mientras un valor de +1 representa una correlación perfecta positiva. Un valor de 0 representa una falta de correlación.

Equiprobables: cuando la probabilidad de ocurrencia de dos sucesos es la misma.

KDD: es el proceso completo de extracción de conocimiento implícito en los datos, es denominado Descubrimiento de Conocimiento en Bases de Datos o KDD, por sus siglas en inglés "*Knowledge Discovery in Databases*". (Han, y otros, 2001).

Media: valor que se obtiene al sumar todos los elementos en un conjunto y dividirlos entre el número de elementos.

Métodos multivariados: herramientas estadísticas que estudian el comportamiento de tres o más variables al mismo tiempo.

Multicolinealidad: Problema estadístico que se presenta en el análisis de Regresión Lineal Múltiple, en el que la confiabilidad de los coeficientes de regresión se ve reducida debido a un alto nivel de correlación entre las variables independientes.

Plataforma Web: es un sistema formado por un conjunto de componentes hardware y software que proporcionan capacidades (o servicios) sobre las que se deberá apoyar cualquier aplicación software y cuyo funcionamiento es a través de internet o la Web. (Aportela Rodríguez, 2007)

Plataforma VideoWeb: es una plataforma web, ya que se conforma con el objetivo funcional principal de brindar servicios de audio y video por la web.

Regresión: proceso general que consiste en predecir el comportamiento de ciertas variables a partir de otras.

R: software estadístico.

Técnicas bivariadas: técnicas que exigen análisis explicativos simples, relacionando dos variables.

Varianza: promedio de la desviación al cuadrado de una variable respecto de su media.

Variable cuantitativa: Son las variables que se expresan mediante cantidades numéricas.

Bibliografía

Aportela Rodríguez, I. M. 2007. Intranets: las tecnologías de información y comunicación en función de la organización. 2007.

Arévalo, José Luis Díaz y García, Rafael Pérez. Estado del arte en la utilización de técnicas avanzadas para la búsqueda de información no trivial a partir de datos en los sistemas de abastecimiento de agua potable. España : s.n.

Ascacíbar, Dr. Fco. Javier Martínez de Pisón. Técnicas de Minería de Datos. Técnicas de Modelado Predictivo. s.l. : Universidad De la Rioja Grupo EDMANS (Engineering Data Mining And Numerical Simulation).

Castillo, José Antonio Sáez. 2009. Métodos Estadísticos con R y R Commander. Dpto. Estadística e Investigación Operativa. Universidad de Jaén : s.n., 2009.

Collaborative Filtering Using Restoration Operators. **Nakamura, A., Kudo, M. y Tanaka, A. 2003.** Springer-Verlag Berlin Heidelberg : s.n., 2003.

Emmanuel, Paradis. 2003. R para Principiantes. Universit Montpellier II, France: s.n., 2003.

Freund, jonh E., R. Miller, Irwin y Johnson, Richard. 2006. Libro Probabilidad y Estadística para Ingenieros. La Habana : Félix Varela, 2006.

Gómez, A. J. Arriaza, y otros. 2008. Esdística Básica con R y R-Commander. s.l. : UCA Universidad de Cadiz, 2008.

Han, J. y Kamber, M. 2001. Data Mining. Concepts and Techniques. s.l. : Morgan Kaufmann Publisher, 2001.

IWorld.com. **web, Técnicas y modelos de personalización de sitios. Sciences de l' Evolution, 2003.** Sciences de l' Evolution, 2003. F-34095 Montpellier cdex 05.

Jaramillo, Carlos Mario Soto. 2009. Incorporación de Técnicas Multivariantes en un Sistema Gestor de Bases de Datos. Colombia-Medellín : Universidad Nacional de Colombia, 2009.

Jhonson, Dallas E. 1998. Métodos Multivariados aplicados al Análisis de Datos. Kansas State University : International Thomson Editores, 1998.

Merino, Vicente Rodrigo Pesantez. 2008. Educación adaptativa en la web: estado del arte. loja – ecuador : universidad técnica particular de loja, 2008.

Molinero, Luis M. Abril, 2002. Construcción de modelos de regresión. España : s.n., Abril, 2002.

Montgomery, D.C. y Runger, G.C. 2003. Applied Statistics and Probability for Engineers. s.l. : John Wiley & Sons, 2003. ISBN 0-471-20454-4..

Neter, J., y otros. 1996. Applied Linear Statistical Models. s.l. : McGraw-Hill, 1996. ISBN 0-256-11736-5.

ORALLO, C. F. R. M. J. R. Q. J. H. 2004. Introducción a la Minería de Datos. s.l. : Prentice Hall, 2004.

Pacheco, Joaquín, y otros. 2005. Predicción de la insolvencia empresarial: uso de búsqueda Tabú para selección de ratios explicativos. España : IE-Working Paper, 2005. WPE5-34.

Pitarque, Alfonso, Ruiz, Juan Carlos y Roy, Juan Francisco. Las redes neuronales como herramientas estadísticas no paramétricas de clasificación. España : Universidad de Valencia.

S, Pértega Díaz y S., Pita Fernández. 20/08/2001. Técnicas de regresión: Regresión Lineal Múltiple. Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña. : s.n., 20/08/2001. Cad Aten Primaria 2000; 7: 173-176..

Salas, Christian. 2008. ¿Por qué comprar un programa estadístico si existe R? . Chile : s.n., 2008.

Vallejos, Sofia J. 2006. Minería de Datos. Argentina : s.n., 2006.

Varmuza, K. y Filzmoser, P. 2009. Introduction to multivariate statistical analysis in chemometrics. Boca Raton : CRC Press, 2009.

Weisberg, S. 2005. Applied Linear Regression. New Jersey : John Wiley & Sons, 2005.

Anexos

Anexo 1: Conjunto de datos generado para el análisis estadístico.

ID_Usuarios	Drama x_1	Acción x_2	Comedia x_3	Terror x_4	Suspenso x_5	Aventura x_6
1	4	3	3	3	2	2
2	5	10	6	9	8	4
3	1	8	3	7	6	2
4	7	4	8	5	3	6
5	9	1	7	2	4	8
6	3	4	3	3	2	2
7	10	5	9	6	4	8
8	8	1	7	3	2	6
9	4	7	5	8	6	3
10	1	9	2	7	8	4
11	4	3	3	3	2	2
12	5	10	6	9	8	4
13	1	8	3	7	6	2
14	7	4	8	5	3	6
15	9	1	7	2	4	8

Anexo 2: Código generado por R para la implementación del algoritmo.

```
> #Función para cargar los datos.
> Datos<-read.table("C:/Program Files/R/data/muestra_genero.txt",skip=1)
> Datos1<-Datos[,-c(1)]
> Datos1
```

	V2	V3	V4	V5	V6	V7
1	4	3	3	3	2	2
2	5	10	6	9	8	4
3	1	8	3	7	6	2
4	7	4	8	5	3	6
5	9	1	7	2	4	8
6	3	4	3	3	2	2
7	10	5	9	6	4	8
8	8	1	7	3	2	6
9	4	7	5	8	6	3
10	1	9	2	7	8	4
11	4	3	3	3	2	2
12	5	10	6	9	8	4
13	1	8	3	7	6	2
14	7	4	8	5	3	6
15	9	1	7	2	4	8

```
> #Función para calcular la matriz de correlación
> cor(Datos1)
```

	V2	V3	V4	V5	V6	V7
V2	1.0000000	-0.6170457	0.90492774	-0.42403562	-0.3864323	0.8968542
V3	-0.6170457	1.0000000	-0.33327590	0.95023513	0.8700936	-0.4630765
V4	0.9049277	-0.3332759	1.00000000	-0.09061804	-0.1699561	0.8581194
V5	-0.4240356	0.9502351	-0.09061804	1.00000000	0.8476790	-0.3024968
V6	-0.3864323	0.8700936	-0.16995612	0.84767899	1.0000000	-0.1285660
V7	0.8968542	-0.4630765	0.85811943	-0.30249678	-0.1285660	1.0000000

```
> #Primer modelo
> reg<-function(i,j,k,l){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k]+Datos1[,l],data=Datos1)
+ summary(r)
+ }
> reg(1,2,3,6)
Call:
lm(formula = Datos1[, i] ~ Datos1[, j] + Datos1[, k] + Datos1[, l], data = Datos1)
Residuals:
     Min       1Q   Median       3Q      Max
-1.08341 -0.61105  0.04479  0.75940  0.80956

Coefficients:
             Estimate   Std. Error  t value Pr(>|t|)
(Intercept)   1.14922    0.77722    1.479  0.16729
Datos1[, j]  -0.28538    0.07526   -3.792  0.00298 **
Datos1[, k]   0.76685    0.18110    4.234  0.00140 **
Datos1[, l]   0.32348    0.18965    1.706  0.11610
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.799 on 11 degrees of freedom
Multiple R-squared: 0.9453,     Adjusted R-squared: 0.9304
F-statistic: 63.37 on 3 and 11 DF,  p-value: 3.154e-07

> #Comprobación de normalidad
> reg<-function(i,j,k,l){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k]+Datos1[,l],data=Datos1)
+shapiro.test(r$res)
+ }
> reg(1,2,3,6)
Shapiro-Wilk normality test
data:  r$res
W = 0.8835, p-value = 0.05353
```

```
> #Segundo modelo
> reg<-function(i,j,k,l){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k]+Datos1[,l],data=Datos1)
+ summary(r)
+ }
> reg(2,1,4,5)
Call:
lm(formula = Datos1[, i] ~ Datos1[, j] + Datos1[, k] + Datos1[, l], data = Datos1)
Residuals:
     Min       1Q   Median       3Q      Max
-1.32105 -0.22411 -0.03664  0.33884  0.90131

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)   0.7326     0.6247      1.173    0.265678
 Datos1[, j] -0.2715     0.0610     -4.451    0.000976 ***
 Datos1[, k]  0.8675     0.1289      6.731    3.24e-05 ***
 Datos1[, l]  0.2891     0.1374      2.104    0.059237 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.625 on 11 degrees of freedom
Multiple R-squared: 0.9706,     Adjusted R-squared: 0.9626
F-statistic: 121.3 on 3 and 11 DF,  p-value: 1.040e-08

> #Comprobación de normalidad
> reg<-function(i,j,k,l){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k]+Datos1[,l],data=Datos1)
+shapiro.test(r$res)
+ }
> reg(2,1,4,5)
Shapiro-Wilk normality test
data:  r$res
W = 0.9497, p-value = 0.5197
```

```
> #Tercero modelo
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+ summary(r)
+ }
> reg(3,1,6)
Call:
lm(formula = Datos1[, i] ~ Datos1[, j] + Datos1[, k], data = Datos1)
Residuals:
    Min      1Q  Median      3Q     Max
-1.1737 -1.0596  0.2965  0.7638  1.3541

Coefficients:
             Estimate   Std.Error  t value  Pr(>|t|)
(Intercept)  1.5328     0.5876     2.609    0.0229 *
Datos1[, j]  0.5298     0.2061     2.570    0.0245 *
Datos1[, k]  0.2341     0.2649     0.884    0.3942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.033 on 12 degrees of freedom
Multiple R-squared:  0.83,      Adjusted R-squared: 0.8016
F-statistic: 29.29 on 2 and 12 DF,  p-value: 2.417e-05

> #Comprobación de normalidad
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+shapiro.test(r$res)
+ }
> reg(3,1,6)
Shapiro-Wilk normality test
data:  r$res
W = 0.8798, p-value = 0.04714
```

```
>#Cuarto modelo
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+ summary(r)
+ }
> reg(4,2,5)

Call:
lm(formula = Datos1[, i] ~ Datos1[, j] + Datos1[, k], data = Datos1)
Residuals:
    Min      1Q  Median      3Q     Max
-1.2207 -0.4648 -0.2925  0.6860  1.3821
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.33557    0.49075   2.721 0.018557 *
Datos1[, j]   0.67460    0.13929   4.843 0.000403 ***
Datos1[, k]   0.09335    0.19628   0.476 0.642902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8307 on 12 degrees of freedom
Multiple R-squared: 0.9047,     Adjusted R-squared: 0.8889
F-statistic: 56.99 on 2 and 12 DF,  p-value: 7.471e-07

> #Comprobación de normalidad
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+shapiro.test(r$res)
+ }
> reg(4,2,5)
Shapiro-Wilk normality test
data:  r$res
W = 0.9517, p-value = 0.5523
```

```
>#Quinto modelo
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+ summary(r)
+ }
> reg(5,2,4)
Call:
lm(formula = Datos1[, i] ~ Datos1[, j] + Datos1[, k], data = Datos1)
Residuals:
    Min      1Q  Median      3Q     Max
-1.5173 -0.9136 -0.1995  0.4594  2.0980

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0332     0.8590   1.203    0.252
Datos1[, j]    0.4724     0.3211   1.471    0.167
Datos1[, k]    0.1982     0.4167   0.476    0.643
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.21 on 12 degrees of freedom
Multiple R-squared: 0.7616,     Adjusted R-squared: 0.7218
F-statistic: 19.16 on 2 and 12 DF,  p-value: 0.0001838

> #Comprobación de normalidad
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+shapiro.test(r$res)
+ }
> reg(5,2,4)
Shapiro-Wilk normality test
data:  r$res
W = 0.9096, p-value = 0.1336
```

```
>#Sexto modelo
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+ summary(r)
+ }
> reg(6,1,3)
Call:
lm(formula = Datos1[, i] ~ Datos1[, j] + Datos1[, k], data = Datos1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23739 -0.62888 -0.09309  0.31335  2.57434
Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  0.3868     0.7685      0.503    0.6239
Datos1[, j]  0.5169     0.2262      2.285    0.0413 *
Datos1[, k]  0.2610     0.2953      0.884    0.3942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.091 on 12 degrees of freedom
Multiple R-squared: 0.8163,     Adjusted R-squared: 0.7857
F-statistic: 26.66 on 2 and 12 DF,  p-value: 3.843e-05

> #Comprobación de normalidad
> reg<-function(i,j,k){r<-lm(Datos1[,i]~Datos1[,j]+Datos1[,k],data=Datos1)
+shapiro.test(r$res)
+ }
> reg(6,1,3)
Shapiro-Wilk normality test
data:  r$res
W = 0.9053, p-value = 0.1147
```