Facultad de Ingeniería
División de Estudios de Postgrado
Maestría en Computación

BÚSQUEDA DE MODELOS PARA EL RECONOCIMIENTO DE PATRONES DE USO DE UN SITIO WEB A TRAVÉS DE LA MINERÍA DE DATOS

Autor: Hedry Camanes Fortoul Yegues
Tutor: Wilfredo Bolivar Maluenga

Mérida, Septiembre 2008

Agradecimientos

A Dios Todopoderoso, sin Él nada de esto sería posible.

A mis padres y mi tío “Maquique”, quienes cultivaron en mí el espíritu de superación y han sido pilares fundamentales de mi vida.

A mi esposo Giancarlo, por todo su amor, comprensión y tolerancia. Pero sobre todo por ser mi amigo incondicional en todo momento.

A la Universidad Nacional Experimental del Táchira, por ser la casa de estudios que me brindó la base de mi formación académica.

A la Universidad de Los Andes - División de estudios de Postgrado - Maestría en Computación, por permitirme adquirir nuevos conocimientos a través del recurso más valioso que posee, sus profesores.

Al profesor Wilfredo Bolivar por ser mi guía y apoyo permanente durante el desarrollo de este trabajo de investigación.

A la Coordinación de Control de Estudios de la UNET, por facilitarme información necesaria para llevar a cabo el desarrollo de mi tesis.

Para finalizar, quiero expresar un profundo agradecimiento a todos quienes con su ayuda, apoyo y comprensión me alentaron a lograr esta hermosa realidad.

A mis padres

A mi esposo

Resumen

La presente investigación tuvo como objetivo la búsqueda de modelos que facilitaran el reconocimiento de patrones de comportamiento de usuarios en un sitio web. El caso de estudio se enmarcó en el Módulo de Departamentos de la página de Control de Estudios y Evaluación de la Universidad del Táchira. Fue aplicada la metodología KDD (*Knowledge Discovery in Databases*) utilizando diversas técnicas de minería, entre las que se pueden destacar: el enfoque tradicional estadístico de regresión lineal, el análisis de secuencias utilizando Cadenas de Markov, la implementación del algoritmo de alineamiento global de secuencias de Needleman-Wunsch, Bayes Ingenuo y *Clustering*. *Para realizar las pruebas, se* utilizaron registros de navegación almacenados en una Base de Datos MySql, los cuales proporcionaron el identificador del usuario conectado, la página visitada, la dirección IP de origen de la conexión, entre otros. La aplicación de las técnicas mencionadas, se llevó a cabo de dos maneras; en primer lugar, a través de programas desarrollados en el lenguaje de programación Java y en segundo lugar, haciendo uso de herramientas automáticas de estadística y minería como SPSS (*Statistical Package for the Social Sciences*) y Weka (Waikato Environment for Knowledge Analysis). Los resultados permitieron obtener modelos que se ajustaron (en diferente grado) al objetivo planteado y fue la técnica de Bayes Ingenuo la que proporcionó, a partir de la muestra seleccionada de usuarios, un alto porcentaje de precisión para tratar de determinar a través de los registros de navegación en el sitio, si un usuario es quien dice ser.

Palabras Claves - Patrones de comportamiento, Minería de Datos, *Knowledge Discovery in Databases*, Cadenas de Markov, Alineamiento Global de Secuencias, Bayes Ingenuo, Clustering.

Índice General

Índice de tablas

Índice de figuras

CAPITULO I

Introducción

1.1 Introducción al Problema

Los Sistemas de Información y sus tecnologías han cambiado la forma en que operan las organizaciones. Su uso permite mejoras en el desarrollo de las actividades que estas llevan a cabo a través de la automatización de los procesos operativos y el suministro de una plataforma de información para la toma de decisiones [1].

El creciente proceso de globalización y el desarrollo de la nueva sociedad de la información han dado paso a lo que hoy en día se conoce como Gobierno Electrónico. Esta filosofía propone la modernización de los procesos de la gestión pública, así como la revisión, rediseño y optimización de los procesos como paso previo a la introducción de cualquier cambio en la tecnología o en las funciones de producción de las organizaciones públicas. De esta manera, el Gobierno Electrónico se convierte en instrumento fundamental para mejorar el desempeño de los actos del Estado [2].

Indudablemente, toda organización depende de la información; la utilización de ésta de manera adecuada y en el momento oportuno constituye el punto central en el desarrollo de la sociedad, por lo cual se hace necesario utilizar nuevas técnicas para el proceso de transformación de datos y para poder acceder a estos de una forma más directa, como por ejemplo: aplicaciones en ambiente WEB.

Las universidades no escapan a este planteamiento, puesto que las mismas manejan grandes volúmenes de información en lo que respecta al área académica y financiera, siendo éste un factor determinante para la evaluación del cumplimiento de sus metas y objetivos, lo que permite medir la eficacia económica, la gestión y el impacto social en las principales áreas que enmarcan la actividad universitaria.

Al desarrollar Sistemas de Información en ambiente Web que manejan datos críticos, se debe pensar detenidamente en los problemas de seguridad y en el modo en

que se reducirá el riesgo de posibles ataques, esto implica ofrecer métodos que aseguren la integridad de los datos, que en cierto modo, pueden verse afectados por intrusos en la red.

Existen diversas técnicas para mantener la seguridad en un Sitio Web. El uso de un sistema seguro de autenticación es, sin duda, un paso inicial a la hora de evitar esta clase de riesgos. No obstante, un sistema de autenticación seguro no servirá de mucho si los usuarios utilizan contraseñas en blanco o fáciles de adivinar. Una de las opciones que se podrían implementar, sería la auditoría de ciertos eventos que ocurren durante la sesión de un usuario, registrando transacciones susceptibles a la suplantación, junto con el nombre de usuario, la hora, la fecha y la información necesaria para identificar los detalles de dicha transacción [3].

Actualmente los Servidores Web donde reposan dichas aplicaciones, generan un gran volumen de datos provenientes del registro de las acciones que los usuarios realizan. Cada requerimiento de los clientes (browsers, agentes, etc.) queda registrado en los *logs* que se generan constantemente en el servidor [4]. Sin embargo, utilizar éstos registros resulta una tarea compleja debido a que el formato en el que se encuentran los datos en ocasiones requiere de una limpieza previa con el fin de obtener parámetros de interés para el proceso de minería; problemas como la dificultad en la detección del usuario y de las sesiones, hacen que los archivos de transacciones no sean tomados en cuenta como una herramienta viable para el estudio del uso de un sitio [5].

Existen aplicaciones web que además de generar dichos *logs*, guardan sus propios registros con información mejor estructurada, lo que facilita en gran medida el análisis de los datos que allí se encuentran. Tal es el caso de la Página Web de Control de Estudios de la UNET; la cual posee una Base de Datos donde se registran las acciones que cada usuario realiza dentro del sitio, permitiendo esto, proponer modelos que faciliten la identificación de patrones de comportamiento de usuarios en el Sitio (utilizando técnicas de Minería de Datos), para así intentar determinar si un usuario registrado es quien dice ser a través su información histórica.

1.2 Descripción del Trabajo

El desarrollo de aplicaciones web supone la exposición de gran cantidad de información restringida en ámbitos inseguros por definición. El tema de crear aplicaciones Web seguras es un tanto complejo, ya que requiere realizar un estudio para comprender los puntos vulnerables de la seguridad [6].

Una de las principales preocupaciones de los desarrolladores de aplicaciones web, es la posible vulnerabilidad de sus sistemas a los ataques de presuntos impostores. El peligro radica en que dichos ataques pueden tener efectos tan devastadores como la limitación de la disponibilidad del servicio, el acceso ilícito a datos privados y, en el peor de los casos, la pérdida del control de los equipos en beneficio de los usuarios malintencionados [3].

A pesar de que en los últimos años se ha observado un marcado incremento de los Sistemas de Seguridad de las aplicaciones Web, en su mayoría han sido vulnerados [7]. La minería de datos a través del Web Mining se ha convertido en una herramienta imprescindible para descubrir patrones en la estructura, el contenido y la utilización de los Sitios Web [8].

La Coordinación de Control de Estudios de la Universidad del Táchira no escapa de esta realidad y su sistema se encuentra susceptible a la entrada de usuarios no autorizados. A partir de esto se deriva la idea de buscar modelos para tratar de identificar patrones de uso de los usuarios registrados con el fin de detectar intrusos en el sitio; y así lograr tomar las medidas necesarias para mantener la integridad de las operaciones que realizan los Departamentos de la Universidad a través de esta página.

1.2.1 Justificación

El aumento y la gravedad de los ataques de seguridad de hoy en día hacen que la detección de intrusos sea una parte indispensable en cualquier sistema. A pesar de que los enfoques clásicos de seguridad en sistemas informáticos se basan en la aplicación de sistemas de autenticación de usuario (Nombre de usuario y contraseña), debemos estar claros que estos pueden sucumbir ante la acción de un pirata que trate de acceder al

archivo de contraseñas de una máquina, aprovechando un *bug* (Error de programación que genera problemas en las operaciones de una computadora) del servidor [9].

En tal sentido, se hace necesario encontrar la manera de detectar la entrada, a un sistema, de personas no autorizadas mediante la identificación de patrones de uso de los usuarios legítimos del mismo para intentar garantizar la integridad de las operaciones que se llevan a cabo dentro del sistema en estudio.

1.2.2 Hipótesis

Es posible encontrar patrones de comportamiento de usuarios en un sitio Web mediante el uso de técnicas de minería de datos.

1.2.3 Objetivos

Objetivo General

Búsqueda de modelos para el reconocimiento de patrones de uso de un Sitio Web a través de la Minería de Datos, específicamente, en el Site de Departamentos de la página de Control de Estudios y Evaluación de la Universidad del Táchira.

Objetivos Específicos

- Investigar acerca del uso del Web Mining y de algunas técnicas de Minería de Datos que permitan obtener modelos para identificar patrones de uso de un Sitio Web.
- Seleccionar las técnicas de minería que serán aplicadas.
- Identificar y depurar los datos que serán utilizados para el proceso de minería.
- Buscar modelos para identificar patrones de comportamiento de los usuarios a partir de las técnicas de minería seleccionadas.
- Realizar pruebas sobre los datos para validar los modelos encontrados.

1.2.4 Sumario

El Capítulo 2 describe la metodología seguida para la aplicación de las técnicas de minería seleccionadas. El Capítulo 3 muestra la descripción desde el punto de vista computacional de las técnicas de minería utilizadas para tratar de conseguir los modelos que identifican los usuarios. En el Capítulo 4 se muestran los resultados obtenidos de la aplicación de la metodología propuesta y la validación de los modelos encontrados.

CAPITULO II

Metodología

2.1 Introducción

El *Data Mining* es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes conjuntos de datos. Los algoritmos de *Data Mining* se enmarcan en el proceso completo de extracción de información conocido como KDD (*Knowledge Discovery in Databases*), que se encarga de preparar los datos y de interpretar los resultados obtenidos.

El análisis de la información recopilada en algunas ocasiones puede llevarse a cabo de forma manual, utilizando para ello algunas técnicas estadísticas. Sin embargo, cuando la cantidad de datos de los que se dispone aumenta, esta forma de estudio se puede complicar. Allí es donde entra en juego el conjunto de técnicas de análisis automático al que hace referencia el *Data Mining* o KDD [10].

2.2 Antecedentes

Existen algunos trabajos relacionados con Minería Web, específicamente algunos inherentes a la identificación de patrones de usabilidad:

- Julio Villena, *et al.* [11], desarrollaron un trabajo de minería de datos que estuvo orientado al análisis de tráfico de usuarios en un Sitio Web. Para ello se incorporó el uso de huellas (marcas generadas en los registros de navegación por los usuarios de un sitio al realizar la petición de páginas con elementos dinámicos, lo que permite que dicha petición se realice directamente al servidor web, reduciéndose la posibilidad de que esta sea servida por máquinas intermedias), que permitieran entender los patrones de acceso y comportamiento habitual de los usuarios dentro del sitio y sus tendencias de navegación.

- José Luís Ortega Priego [5], realizó un estudio acerca de la usabilidad y navegabilidad de la web del Cindoc a través de los archivos de transacciones web de su servidor central durante el mes de octubre de 2003. El objetivo era determinar pautas de navegación de los usuarios del sitio y fallos en el diseño del mismo. Este estudio además de permitirle a Cindoc verificar los contenidos más demandados en su sitio, le permitió concluir que el uso de metodologías basadas en minería web, permiten constatar el comportamiento de los usuarios en dicho sitio.
- Cristóbal Romero, *et al.* [12], realizaron un estudio que describe la utilización de técnicas de minería de datos en sistemas de e-learning. El objetivo de este trabajo fue descubrir posibles reglas de predicción que permitieran la adaptación del contenido de un curso web a partir del análisis de información previa del nivel de conocimiento de cada participante. El método de descubrimiento de reglas que se propuso en este estudio fue el de Algoritmos Evolutivos y en concreto la programación genética basada en gramática.
- Martinelli D., et al. [13], desarrollaron otro importante trabajo de investigación en el cual se propone la utilización de una red neuronal SOM (mapa auto-organizativo) para la identificación de hábitos de navegación de los usuarios de un sitio en base a las páginas visitadas. La investigación permitió comparar los resultados obtenidos utilizando la mencionada técnica y los arrojados por el algoritmo de las K-medias. El trabajo concluye que la red neuronal ofrece mejores resultados que los obtenidos con la técnica el algoritmo de clustering utilizado.
- Otro importante estudio encontrado, fue el desarrollado por la Ing. Yenit Juliana Guerrero [14], quien plantea el uso de cubos, patrones secuenciales, agrupamiento y análisis de caminos como herramientas para desarrollar un sistema hibrido de Minería de Datos para encontrar páginas frecuentemente visitadas por los usuarios al navegar por internet, con la finalidad de ofrecer recomendaciones en cuanto a la estructura del sitio y la disposición de sus contenidos.

2.3 Fuente de Datos

Con el objetivo de realizar la verificación de la hipótesis planteada, se consideró como fuente de datos para los modelos desarrollados, la Base de Datos (MySql) del

Módulo de Departamentos del sitio Web de la Coordinación de Control de Estudios de la Universidad del Táchira (base de datos destinada a registrar las acciones de cada usuario dentro del Sitio). Cada registro guardado en dicha Base de Datos proporciona información de sesión por usuario, es decir, es posible saber la hora de inicio y fin de sesión, las páginas visitadas, la hora en la que visitó cada link y la dirección IP desde donde se inicio la conexión. (Ver Tabla 2.1)

Campo	**Descripción**
ID_CAMPO	Identificador único (clave primaria de la tabla)
USUARIO	Identificador de acceso al sitio. Se corresponde con el número de cédula del responsable de la cuenta.
IP	Dirección IP de la máquina en la cual el usuario inició sesión
FECHA	Registro de aaaa/mm/dd en la que se realizó la transacción
HORA	Registro de hh:mm:ss en la que se realizó la transacción
PAGINA	Nombre de la página visitada
PARAMETROS	Información del módulo al que pertenece el registro y acción realizada dentro de la página visitada.

Tabla 2.1: Información proporcionada por la Base de Datos MySql

2.4 Método

La finalidad de éste trabajo es la aplicación de técnicas de minería de datos que permitan obtener modelos para reconocer patrones de comportamiento de los usuarios de un sitio web. En tal sentido se propone utilizar la metodología KDD (Ver Figura 2.1), que plantea la realización de las actividades que se presentan a continuación: [15]

- *Especificación del Problema.* Esta fase requiere del conocimiento de los datos y la manera en que fueron generados; saber que información está disponible y cuál debe ser obtenida.

- *Construcción de la Base de Datos para la Minería.* Esta fase está dirigida a la generación de una base de datos alterna con el propósito de llevar a cabo el proceso de minería sin correr el riesgo de alterar los datos en la base de datos original. Además de esto, debe realizarse una reestructuración de datos incompletos e inconsistentes. Esto último es extremadamente importante, ya que los datos ”sucios” implican un análisis inexacto y unos resultados, por tanto, incorrectos (Se considerará que los datos son ”sucios” o ”ruidosos” cuando exista una importante contribución aleatoria de los mismos, la cual no aporta conocimiento alguno, por ejemplo: registros repetidos, registros que previamente se consideró no incluir dentro del estudio).

- *Exploración de los Datos.* En esta fase se emplean algunas técnicas de visualización de datos, de búsqueda de relaciones entre atributos y otras medidas de exploración. La meta es identificar los campos con mayor potencial predictivo y los atributos útiles para el proceso.

- *Preparación de los datos para generar el modelo*. Esta fase incluye cuatro pasos importantes:
    - Selección de atributos: En un caso ideal se toman todos los atributos disponibles, alimentando con ellos los algoritmos de Minería de Datos y dejándolos encontrar las mejores predicciones. En la práctica esto no funciona bien debido a que el tiempo empleado para construir un modelo se incrementa con la cantidad de atributos y esto puede llevar a la creación de modelos erróneos.
    - Construcción de nuevos atributos: este paso permite crear nuevas parámetros derivados de datos en bruto.
    - Re-codificación de atributos: paso que permite que los campos puedan ser ajustados para que entren en un rango particular previamente definido.

- *Construcción del Modelo.* Es un proceso iterativo y será necesario explorar múltiples técnicas alternativas hasta encontrar las más útiles para alcanzar el objetivo planteado. Con base en los resultados obtenidos, se podrá decidir si crear otros modelos empleando la misma técnica con parámetros diferentes o intentar con otra técnica o algoritmo.

- *Evaluación del Modelo.* En esta fase se deben examinar los resultados arrojados por los modelos e interpretar sus significados. Para ello se pueden utilizar herramientas visuales que generen gráficos estadísticos.

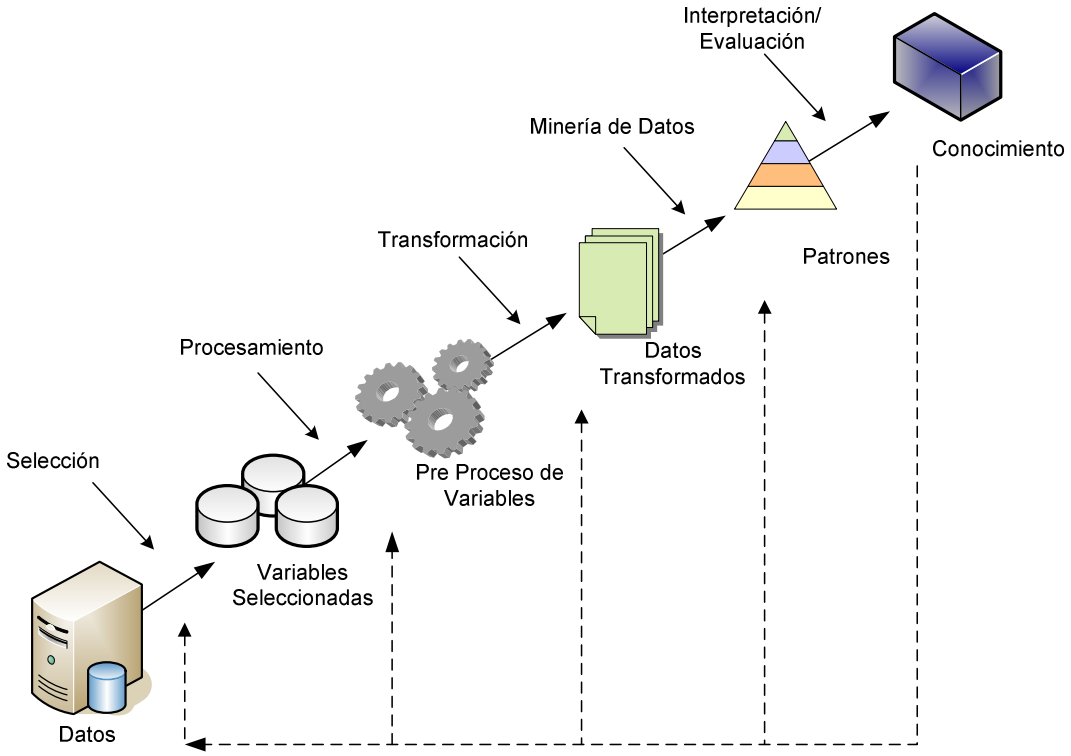


Figura 2.1: Proceso KDD

2.4.1 Especificación del Problema y Construcción de la Base de Datos para la Minería

Antes de iniciar el proceso de análisis fue necesario determinar las características de los datos con los que se contaba y las posibilidades que estos proporcionaban para alcanzar el objetivo planteado. En primer lugar se llevó a cabo una revisión de la cantidad de información disponible para iniciar el proceso de minería; y se pudo observar la presencia de cerca de 12 millones de registros en un lapso de 3 años de funcionamiento del sitio para todos los usuarios del módulo que abarca el trabajo de investigación. Estudiar toda la población, es posible, sea la manera más exacta de encontrar los

modelos buscados. Sin embargo, haciendo referencia al concepto de estadística inferencial definido mediante (2.1), se propuso el estudio de una parte de los datos, para luego generalizar los resultados obtenidos a la totalidad de los mismos.

> *"Generalización hacia las poblaciones de los resultados obtenidos en las muestras".[16]*

(2.1)

Antes de llevar a cabo la selección de los datos para el estudio, se calculó el número registros a tomar en cuenta para las pruebas, siguiendo la premisa planteada en [17], acerca de la obtención del tamaño de la muestra para poblaciones finitas y que formalmente está definido según (2.2)

$$n = \frac{N * \hat{p}(1-\hat{p})}{(N-1)(\varepsilon/Z)^2 + \hat{p}(1-\hat{p})}$$

(2.2)

Donde:

N= Tamaño de la población.

$\hat{p} =$ Proporción esperada.

$\varepsilon =$ Error máximo de estimación a permitir.

$Z =$ Z correspondiente a la puntuación en la curva normal estándar para un nivel de confianza elegido.

Debido a que el propósito de seleccionar muestras aleatorias consiste en obtener información acerca de los parámetros de comportamiento de la población, se propuso a continuación la estimación de estadísticos para datos no agrupados como: la media (2.3), la mediana (2.4) y la desviación estándar (2.5), las cuales facilitaron la aplicación de las primeras técnicas de análisis de los datos.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(2.3)

$$med = x_{\left(\frac{n+1}{2}\right)}$$

(2.4)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.5}$$

Donde:

n= representa el número de elementos que pertenecen a la muestra.

x_i= representa la observación i en la muestra seleccionada.

A partir de la definición descrita en (2.2), se procedió al cálculo del error máximo de estimación para un tamaño de muestra calculado n (2.6).

$$\varepsilon = Z * \sqrt{\frac{\frac{N * \hat{p}(1-\hat{p})}{n} - \hat{p}(1-\hat{p})}{N-1}} \tag{2.6}$$

Luego de establecido el tamaño de la muestra y determinado el error máximo esperado para dicha cantidad, se procedió a determinar la manera en la que serían seleccionados los candidatos de la población en estudio. Para ello se llevó a cabo un muestreo aleatorio simple. El mismo consistió en la generación de números aleatorios con una función random en un programa desarrollado en java, el cual permitió obtener la posición de los X_i elementos a tomar del universo de registros en la Base de Datos utilizada.

Una vez acordado el tamaño de la muestra para el estudio, se procedió a crear una Base de Datos alterna con las mismas características de la base de datos original; y a partir de esta, se comenzó el proceso de pruebas para obtener los modelos, sin poner en riesgo la integridad de los datos originales.

2.4.2 Exploración de los Datos

Los datos estadísticos obtenidos de muestras, experimentos o cualquier colección de mediciones, en ocasiones necesitan ser condensados a una forma más adecuada para encontrar la utilidad de los mismos. Algunas veces resulta satisfactorio presentar los datos tal como se encuentran y obtener información directamente de ellos; y otras veces

es necesario agruparlos y presentarlos gráficamente para una mayor comprensión de estos [17].

Con el objetivo de obtener una primera visualización de las características de los datos seleccionados para el estudio, se propuso, la generación de tablas de frecuencias de los atributos en la Base de Datos, así como diagramas de barra y de dispersión que mostraran información preliminar del comportamiento de los usuarios en el sitio.

Durante la fase de selección de atributos para el proceso de minería, se planteó el uso de nuevos parámetros basados en datos existentes; tal es el caso de el atributo grupo_hora. La clasificación propuesta se basó en el cómputo del número de clases a utilizar para la agrupación de los datos y los límites de cada una de las clases. Walpole, et al [18], propone para tal fin la ordenación de los elementos de la muestra seguido del cálculo del rango de los datos R (2.7), a continuación, la estimación del número de intervalos a utilizar (a lo cual denotaremos como K), para luego concluir con la extensión de cada intervalo W (2.8).

$$R = X_n - X_1 \tag{2.7}$$

Donde X_1 es el elemento menor y X_n es el elemento mayor y K el número de intervalos seleccionados (para el caso de estudio por convención K= 24).

$$W = \frac{R}{K} \tag{2.8}$$

2.4.3 Preparación de los datos para generar el modelo

A partir de la visualización gráfica inicial de los datos, fue posible detectar la presencia de valores fuera de rango o también denominados Outliers, definido como [18]:

> "*aquellas observaciones que tienen un comportamiento muy diferente con respecto al resto de los datos frente al análisis que se desea realizar sobre las observaciones experimentales*".

Con el objetivo de decidir el mantener o descartar estas observaciones, se llevó a cabo un estudio de los atributos, a fin de verificar el origen de estos valores y poder establecer un criterio al respecto.

2.4.4 Construcción del Modelo

A fin de verificar la posibilidad de obtener patrones de comportamiento de los usuarios del Sitio de Control de Estudios de la Universidad del Táchira, fueron seleccionadas una serie de técnicas de minería cuyas bases teóricas se describen a continuación:

2.4.4.1 Prueba 1: Análisis de Varianza (ANOVA): es una técnica estadística de contraste de hipótesis que permite analizar simultáneamente la influencia de dos o más factores de clasificación (variables independientes) sobre una variable respuesta continua.

Para llevar a cabo esta prueba se dividió el conjunto de datos en dos grandes grupos: grupo de prueba y grupo de entrenamiento. Para cada uno de los registros se llevó a cabo el cálculo del parámetro "duración" con base en los parámetros fecha y hora de los mismos. Esto dio como resultado el tiempo que determinado usuario permanece en cada página a lo largo de sus sesiones (conociéndose previamente que el inicio y el fin de una sesión estaba determinado por la visita de ciertas páginas).

El análisis buscaba determinar si el comportamiento de navegación de un usuario, en términos de duración por página durante un lapso de tiempo, podía guardar relación con el comportamiento (en los mismos términos) durante otro periodo similar para el usuario en cuestión.

Se asumió que los datos se encontraban normalmente distribuidos con base en lo descrito por el Teorema del Límite Central [18]; y *partir de esto, se procedió a estimar la varianza de la población* S^2 [17], para realizar el análisis*:*

- Varianza dentro de los grupos: representada por MSE y calculada como la media de las k varianzas muestrales. Definida formalmente según (2.9)

$$MSE = \frac{SSE}{k(n-1)} \tag{2.9}$$

donde SSE (2.10) corresponde con la suma de cuadrados del error y k(n-1) los grados de libertad del mismo.

$$SSE = SST - SSA \tag{2.10}$$

con SST (suma de cuadrados en muestras de distinto tamaño) definido según (2.11) y SSA según (2.14)

$$SST = \sum_{i=1}^{k} \quad \sum_{j=1}^{n_i} y_{ij}^2 - C \tag{2.11}$$

C denominado como término de corrección (2.12)

$$C = \frac{T^2}{kn} \tag{2.12}$$

y T representando el total de los tratamientos.

- Varianza entre grupos: representada por MSA y calculada a partir de la varianza de las medias muestrales. Definida formalmente según (2.13)

$$MSA = \frac{SSA}{k-1} \tag{2.13}$$

donde SSA (2.14) corresponde con la suma de cuadrados de los tratamientos y k-1 los grados de libertad del mismo.

$$SSA = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C \tag{2.14}$$

La distribución muestral del cociente de dos estimaciones independientes de la varianza de una población normal es una distribución F con los grados de libertad correspondientes al numerador y denominador respectivamente [17]. Entonces se tiene (2.15)

$$F = \frac{MSA}{MSE} \tag{2.15}$$

Al contrastar el estadístico $F_{calculado}$ ($^{MSA}/_{MSE}$) con el valor de F para un nivel de confianza elegido ($\mathrm{F}_{\alpha(\mathrm{k}-1,(\mathrm{n}-1)\mathrm{k})}$), se rechaza la hipótesis de que las k muestras estudiadas provienen de una misma población si se cumple que $F_{calculado} > \mathrm{F}_{\alpha(\mathrm{k}-1,(\mathrm{n}-1)\mathrm{k})}$.

2.4.4.2 Prueba 2: Cadenas de Markov: el objetivo de aplicar esta prueba a los datos de la muestra consistió en verificar si existía alguna relación entre las saltos de páginas (secuencia de páginas visitadas) durante las sesiones de un usuario en un lapso de tiempo dado. Es decir, poder predecir a partir de datos históricos de cada usuario, la probabilidad de que éste visite cierta página en determinado instante.

Kolman, et al. [19], define formalmente una cadena de Markov como una serie de eventos en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior.

Suponiendo que el sistema tiene n estados posibles (páginas que conforman el sitio en estudio), para cada i=1,2,…,n y cada j=1,2,…,n, sea p_{ij} la probabilidad de que si el sistema se encuentra en la página i en cierto periodo de observación, estará en el página j en el siguiente; p_{ij}, podrá definirse como una probabilidad de transición que debe cumplir con las reglas:

$$0 \leq p_{ij} \leq 1 \tag{2.16}$$

$$\sum_{j=0}^{m} p_{ij}^{(n)} = 1 \qquad \text{para toda i; n=0,1,2,…} \tag{2.17}$$

El método utilizado para esta fase de estudio, se basó en una representación Markoviana de primer orden (n=1) a través de una matriz de transición:

P =

Pagina	**0**	**1**	**2**	**…**	**m**
0	p_{00}^{n}	p_{01}^{n}	p_{02}^{n}	⋯	p_{0m}^{n}
1	p_{10}^{n}	p_{11}^{n}	p_{12}^{n}	⋯	p_{1m}^{n}
2	p_{20}^{n}	p_{21}^{n}	p_{22}^{n}	⋯	p_{2m}^{n}
…	…	…	…	…	…
m	p_{m0}^{n}	p_{m1}^{n}	p_{m2}^{n}	⋯	p_{mm}^{n}

Una vez obtenida la matriz de transición de primer orden para la muestra seleccionada, se propuso la utilización del estadístico Z (2.18) descrito en [17], para una prueba relativa a diferencia entre dos proporciones, con el fin de verificar si la proporción entre dos estados i y j en la matriz obtenida, era igual a la proporción entre los mismos estados i y j en la matriz histórica (en un mismo periodo de tiempo) para un usuario específico.

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \approx N(0,1) \tag{2.18}$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{2.19}$$

Donde

x_1 = número de eventos observados en una muestra de tamaño n_1.

x_2 = número de eventos observados en una muestra de tamaño n_2.

$\hat{p}_1 =$ Probabilidad de transición del estado i al estado j en la matriz calculada,

$\hat{p}_2 =$ Probabilidad de transición del estado i al estado j en la matriz histórica.

$\hat{q} =$ 1- $\hat{p}$

Luego de este procedimiento se obtuvo una matriz de aceptación y rechazo (2.20) de donde se logró hacer inferencias acerca de la similitud de las matrices comparadas.

$$para\ \ p_1 \neq\ p_2\ con\ un\ nivel\ de\ significancia\ \alpha\ la\ región\ crítica\ es$$

$$Z < -Z\alpha_{/2} \quad \text{ó} \quad Z > Z\alpha_{/2} \tag{2.20}$$

2.4.4.3 Prueba 3: Alineamiento de Secuencias: La comparación de secuencias es utilizada comúnmente para predecir la similitud de estructura y función entre proteínas. Tales comparaciones requieren en general, un alineamiento de las secuencias que maximice el número de residuos idénticos y que minimice el número de inserciones y sustracciones requeridas para conseguir el alineamiento [20].

Según Devlin [20], dos secuencias se consideran homólogas cuando pueden alinearse en una elevada proporción.

Para determinar si una secuencia de ADN es homóloga a otra, diversos autores han propuesto la utilización de programas de comparación basados en algoritmos de alineamiento enmarcados en dos grandes métodos:

- **Alineamiento local:** trata de encontrar los dos fragmentos de ambas secuencias que tienen un alineamiento con una puntuación máxima.
- **Alineamiento global:** busca el alineamiento de las secuencias completas con una puntuación máxima.

Un procedimiento para generar alineamientos entre secuencias requiere un método para establecer equivalencias entre caracteres y algún criterio para elegir entre todas las posibles soluciones al problema planteado.

Dos secuencias alineadas tendrán mayor similitud mientras mayor sea el número de matchs (coincidencias) y menor el número de mismatch (no coincidencias) y gaps (inserción de espacios). A esto se le conoce como scoring de alineamiento; y no es más que un sistema de puntuaciones que permite calcular un número, que a mayor valor, generará un mayor nivel de significancia.

Ejemplo:

Secuencia 1	A T G C T G G T A	| = coincidencia (match)
	| | + | + + |	+ = cambio (mismatch)
Secuencia 2	A T C - - G C A A	- = Inserción-deleción (Gap)

Uno de los algoritmos más importantes de alineación global [21], es el algoritmo de Needleman-Wunsch implementado mediante la programación dinámica. Este, realiza la comparación de dos cadenas $A\ y\ B$ de tamaño $m\ y\ n$ respectivamente, formadas por elementos de un alfabeto finito de símbolos.

El primer paso que debe llevarse a cabo para comenzar el proceso de alineamiento [22], es el establecimiento de una función de similitud $S(a_i, b_j)$ entre los elementos $a_i\ y\ b_j$ de las secuencias a alinear. Para ello se busca una matriz de similitud basado en un sistema de puntuación que puede ser planteado como (2.21):

Dadas dos secuencias A y B

$$A = a_1 a_2 a_3 \dots a_m$$
$$B = b_1 b_2 b_3 \dots b_n$$

$$S[ij] = \begin{cases} X\ si\ a_i = b_j \\ Y\ si\ a_i \neq b_j \\ Z\ si\ gap \end{cases}$$

$$Con\ \ X > Y > Z$$

(2.21)

A continuación se forma la matriz H (2.22), también denominada matriz de scoring con base en:

$$H_{(i,j)} = \max \begin{cases} H(i-1, j-1) + S(a_i, b_j) \\ H(i-1, j) + W \\ H(i, j-1) + W \end{cases}$$

(2.22)

Donde W representa la penalización por presencia de GAPs (inserciones y sustracciones realizadas durante el proceso de alineamiento).

Una vez obtenida la Matriz de Scoring, el algoritmo plantea la recuperación de la solución mediante un proceso conocido como Backtracking; que consiste en tomar la última coincidencia en el alineamiento realizado y comenzar a buscar el camino que maximice la función mediante el recorrido de los vecinos de la celda $H_{(i,j)}$, es decir, se observan los valores de alineamiento que se encuentran en $H_{(i-1,j)}$, $H_{(i-1,j-1)}$, $H_{(i,j-1)}$, y se selecciona el vecino que presente el más alto puntaje, hasta llegar a un elemento en la primera fila o la primera columna de H. Por convención general, para el llenado de la matriz, el algoritmo establece que el valor en $H_{(0,0)}$ será 0.

Sabiendo que el mejor alineamiento se dará cuando todos los elementos en ambas secuencias sean idénticos, se podrá estimar el porcentaje de similitud que representa el score obtenido con base en el máximo alineamiento.

2.4.4.4 Prueba 4: Bayes Ingenuo: Una vez aplicados algunos métodos estadísticos tradicionales, se planteó la utilización de algunas técnicas de clasificación entre las que se encuentra Naive Bayes, el cual se caracteriza [23], por realizar predicciones en base a asociaciones y relaciones encontradas en datos históricos. Este algoritmo se basa en el Teorema de Bayes considerando la definición de probabilidad total. Expresa la probabilidad de que un elemento d_i, (representado por un vector de atributos independientes), pertenezca a una clase c, dada. (2.23)

$$P(c|d_i) = \frac{P(c)P(d_i|c)}{P(d_i)} \qquad i = 1, 2, 3, \dots, n \tag{2.23}$$

Donde $P(d_i)$ es la probabilidad de escoger aleatoriamente un elemento d_i y $P(c)$ es la probabilidad de que al tomar un d_i cualquiera éste pertenezca a la clase c.

A partir de esto es posible obtener una matriz de confusión definida por Sobrino [24], como un arreglo bidimensional cuadrado donde cada columna representa una categoría resultado de la clasificación y cada fila hace referencia a una clase real. Así pues, los elementos de la diagonal principal representan las unidades clasificadas correctamente y

el resto de los elementos, indican no sólo los clasificados de manera incorrecta, sino la clase a la que fueron erróneamente asignados.

2.4.4.5 Prueba 5: Clustering: La siguiente técnica de descubrimiento aplicada sobre el conjunto de datos es el Clustering. Esta, es una de las principales herramientas utilizadas en el proceso de minería de datos que permite la obtención de grupos y la identificación de características interesantes en los datos. Consiste en la agrupación de una colección de datos no etiquetados en un conjunto de grupos de modo tal, que los objetos que pertenecen a un grupo sean homogéneos entre sí [25]. Expresado en términos de variabilidad se hablaría de minimizar la variabilidad dentro de los grupos para al mismo tiempo maximizar la variabilidad entre los distintos grupos. (Ver Figura 2.2).

El método de Clustering utilizado en esta fase de estudio fue el de particionamiento de K-medias. Este algoritmo hace referencia a la existencia de K clases o patrones, siendo necesario, por tanto, el conocimiento a priori del número de clases existentes; éste valor es determinante para el rendimiento del algoritmo, puesto que un valor de K superior al número real de clases, dará lugar a clases ficticias, mientras que un K inferior producirá menos clases de las verdaderas [26].

Está dado por un conjunto de n puntos de datos en un espacio *d*-dimensional de tipo real R^d y un K número entero. El problema consiste en determinar el conjunto de de K puntos en R^d, llamados centroides, para así minimizar la distancia media cuadrada de cada punto a su centro más cercano [27].

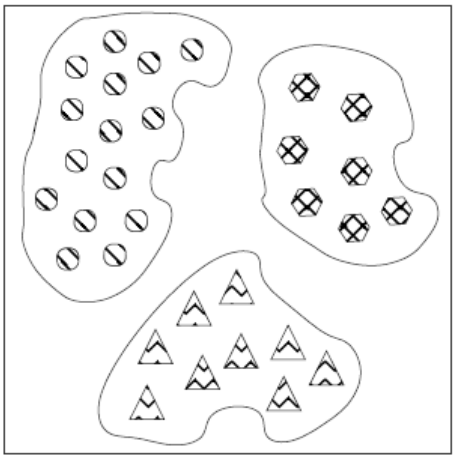


Figura 2.2: Ejemplo de Clustering

Dado el número de clases K, el conjunto de patrones de entrada y los centros de las clases, los pasos para la aplicación del método son:

- Se inicializan de forma aleatoria los centros de las K clases.
- Se asignan los N_i patrones de entrada a cada clase i del siguiente modo:
    - El patrón $X_{(n)}$ pertenece a la clase i si:

$$\left\|X_{(n)} - C_i\right\| < \|X(n) - C_s\| \quad \forall s \neq i \; con \; s = 1,2,\dots,K \tag{2.24}$$

Donde:

$X_{(n)}$ = Patrón de entrada n.

C_i = Clase i.

De esta forma cada clase tendrá asociado un determinado número de patrones de entrada, aquellos más cercanos al centro de la clase.

- Se calcula la nueva posición de los centros de las clases como la media de todos los patrones que pertenecen a su clase:

$$C_{ij} = \frac{1}{N_i} \sum_{n=1}^{N} Min\, X_j\,(n) \qquad para\; j = 1,2,3,\dots,p \quad i = 1,2,3,\dots,K \tag{2.25}$$

Donde:

C_{ij} = Nueva posición de los centros de las clases.

K = Número de clases.

N_i = Patrones de entrada a cada clase i.

- Se repiten los pasos 2.24 y 2.25 hasta que las nuevas posiciones de los centros no se modifiquen respecto a su posición anterior, es decir hasta que:

$$\left\|C_i^{nuevo} - C_i^{anterior}\right\| < \varepsilon \quad \forall i = 1,2,\dots.K \tag{2.26}$$

La principal desventaja de éste método es su dependencia de los valores iniciales asignados a cada centro. Sin embargo, Han, *et.al.* [28], indica es un algoritmo bastante eficiente en problemas de clasificación, pues converge en pocas iteraciones.

2.4.5 Evaluación de Modelos

El desarrollo de esta fase se realizó de manera iterativa al concluir cada una de las técnicas de minería seleccionadas, esto con la finalidad de verificar la aplicabilidad de cada una de ellas en la identificación de patrones de usuarios con base en la información proporcionada por sus registros de navegación. Para tal fin, fueron utilizados los datos de diferentes periodos de tiempo de los mismos usuarios que formaron parte de la muestra.

La evaluación de los modelos se ejecutó en dos etapas. En primer lugar, se dividió la muestra por usuario en dos grandes grupos; un grupo de entrenamiento y otro de prueba, cada uno de ellos correspondiente a registros de un lapso académico de estudio (2007-1 y 2007-3). La aplicación de las técnicas efectivas en este caso pudieran indicar un alto grado de correspondencia de los datos de prueba con los de entrenamiento; lo que significaría que los atributos utilizados para el análisis pudieran dar muestra de que un usuario puede ser identificado a través de estos. En segundo lugar, teniendo conocimiento de los datos pertenecientes a cada usuario, se realizó una mezcla de los registros de navegación de estos (asignación de registros de unos usuarios a otros), con el fin de obtener secuencias que no correspondieran con su comportamiento habitual dentro del Sitio. Con los métodos más eficaces el resultado esperado sería un bajo porcentaje de correspondencia de los registros con el usuario clasificador; esto, por la sencilla razón de que se estaría en presencia de información falsa de determinado usuario, lo que pudiera considerarse un caso de suplantación. Los resultados de la validación del modelo de cada experimento son incluidos en el Capítulo 4, donde además se especifica la manera en que estos fueron realizados.

Como complemento a la validación realizada, al final del estudio fue verificada la técnica de minería con mejor rendimiento durante las fases de prueba, con nuevos registros de navegación de usuario para un nuevo periodo académico (Lapso 2008-1), así como con registros de usuarios no incluidos inicialmente en el estudio.

CAPITULO III

Entorno de Software de Evaluación

3.1 Lineamientos generales

En este capítulo se describe el proceso de desarrollo computacional de las aplicaciones para la evaluación de los registros de navegación de los usuarios en el sitio web de la Coordinación de Control de Estudios y Evaluación de la UNET. Con base en la metodología descrita en el capitulo anterior y utilizando algunas técnicas de minería, se propuso, el desarrollo de programas y la utilización de otras herramientas que permitieran evaluar de manera automática los mencionados registros. Para el caso de los programas desarrollados, se propuso el uso del lenguaje multiplataforma java [29], a través de la interfaz de programación Eclipse (versión 3.2), así como la incorporación de las bibliotecas mysql-connector-java.jar, weka.jar y Jfreechart.jar para los procesos de conexión a base de datos, implementación de las técnicas y generación de salidas respectivamente.

3.2 Proceso de transformación de datos

Una vez configurada la herramienta de programación y antes de comenzar las pruebas con las técnicas seleccionadas, se inicio un proceso de enriquecimiento de los datos, que consistió en la transformación de algunos atributos (clasificaciones, re-codificaciones); y la generación de otros a partir de los ya existentes. Para ello, se desarrolló una aplicación cuyo objetivo fue tomar los datos proporcionados por la Base de Datos original para la creación de nuevas tablas con los datos transformados, de tal manera, de poder disponer de la información necesaria al momento de iniciar la fase de implementación de los métodos.

3.3 Herramientas de procesamiento automático

Para el desarrollo de las pruebas de algunos de los métodos propuestos, fueron utilizadas herramientas automáticas estadística y minería de datos. Tal es el caso de SPSS (para el análisis tradicional estadístico de regresión lineal - ANOVA) y Weka (para la aplicación de Bayes Ingenuo y Clustering). (Ver Figura 3.1).

SPSS (*Statistical Package for the Social Sciences*) es una potente herramienta de tratamiento de datos y análisis estadístico que funciona bajo el Sistema Operativo Windows y proporciona opciones de procesamiento a través de ventanas desplegables y cuadros de diálogo. Su utilización suministró información inicial del comportamiento general de los datos a través de gráficos. Weka (Waikato Environment for Knowledge Analysis), es un conjunto de bibliotecas JAVA para la extracción de conocimiento desde Bases de Datos [30]. Los algoritmos que proporciona la herramienta pueden ser aplicados directamente a un conjunto de datos o llamados directamente desde código Java. Contiene instrumentos de pre-procesamiento, clasificación, regresión, clustering, reglas de asociación y visualización de datos. Para efectos del presente trabajo de investigación se utilizaron las versiones 15 (SPSS) y 3.5.8 (Weka en modo explorador).

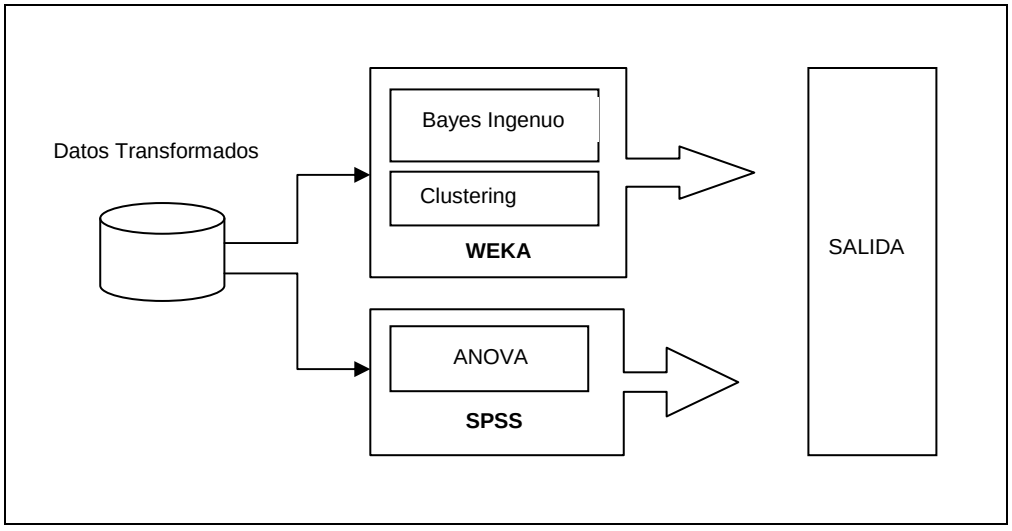


Figura 3.1 Procesamiento de datos a través de Herramientas Automáticas (SPSS y WEKA)

3.4 Implementación de técnicas en Java

Así como en algunos casos la evaluación de las técnicas se realizó a través de herramientas automáticas, en otros, estas fueron implementadas utilizando el lenguaje de programación Java. Cada una de estas aplicaciones estuvo compuesta por tres clases: la primera de ellas, la clase principal correspondiente a cada procedimiento (Cadenas de Markov, Alineamiento de Secuencias mediante el algoritmo de Needleman-Wunsch), una clase "métodos" (donde fueron construidas las consultas y otros procedimientos necesarios) y finalmente una clase "conex", en donde entre otras cosas, se especificó el nombre de la Base de Datos a utilizar, el usuario y la clave de conexión. (Ver Figura 3.2).

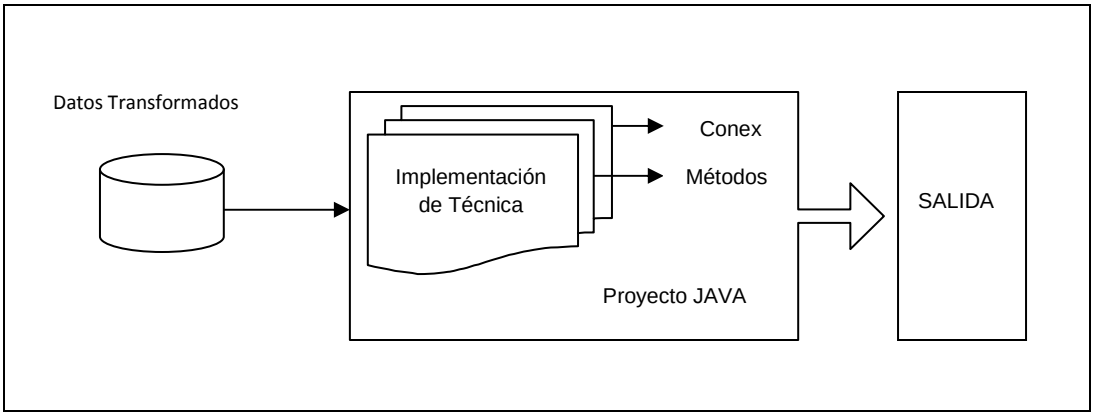


Figura 3.2 Procesamiento de datos a través de aplicaciones desarrolladas en Java (Cadenas de Markov y Alineamiento de secuencias mediante algoritmo de Needleman-Wunsch)

Cada una de estas aplicaciones tuvo como objetivo fundamental evaluar (a través de los métodos propuestos), la posibilidad de verificar a través de los registros de navegación de un usuario, si éste es quien dice ser.

CAPITULO IV

Resultados y Discusión

4.1 Introducción

El análisis y las pruebas se realizaron con aproximadamente 39.000 registros provenientes de sesiones de 11 usuarios del sitio durante los lapsos académicos 2007-1 (fecha de inicio: 19/03/2007, fecha de finalización: 27/07/2007) y 2007-3 (fecha de inicio: 10/09/2007, fecha de finalización: 11/04/2008), esto, con el objetivo de evaluar periodos similares en donde los usuarios tuvieran relativamente el mismo uso del sistema en estudio.

La primera revisión realizada a los datos consistió en la elaboración de gráficos que mostraran las características más importantes de los parámetros involucrados en el estudio. Algunos resultados de dicha visualización se muestran a continuación:

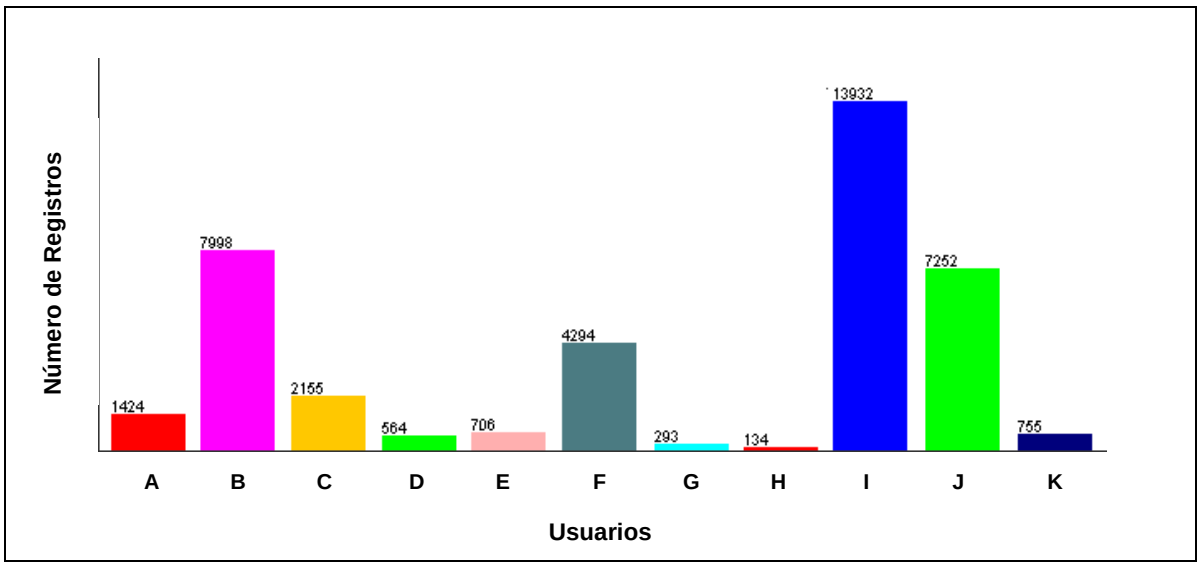


Figura 4.1 Diagrama de barras - registros de datos por usuario

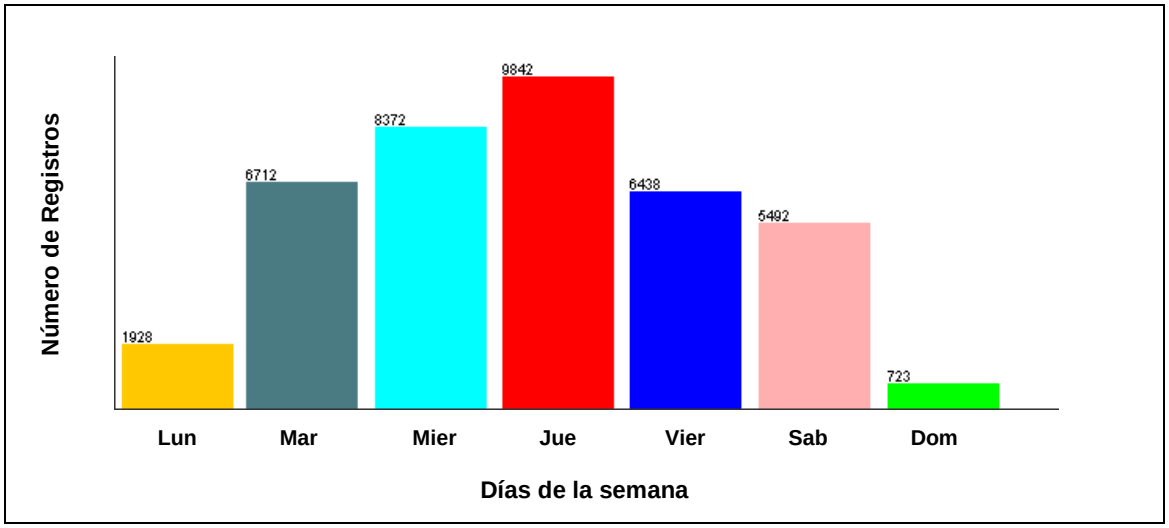


Figura 4.2 Diagrama de barras - registros de conexión por día de la semana

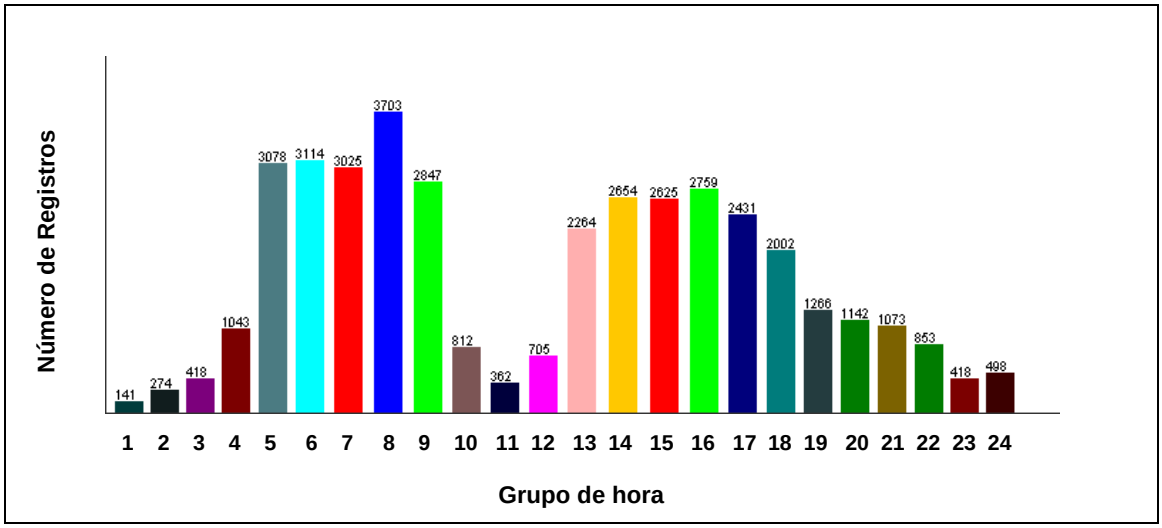


Figura 4.3 Diagrama de barras - registros de conexión por grupo de hora

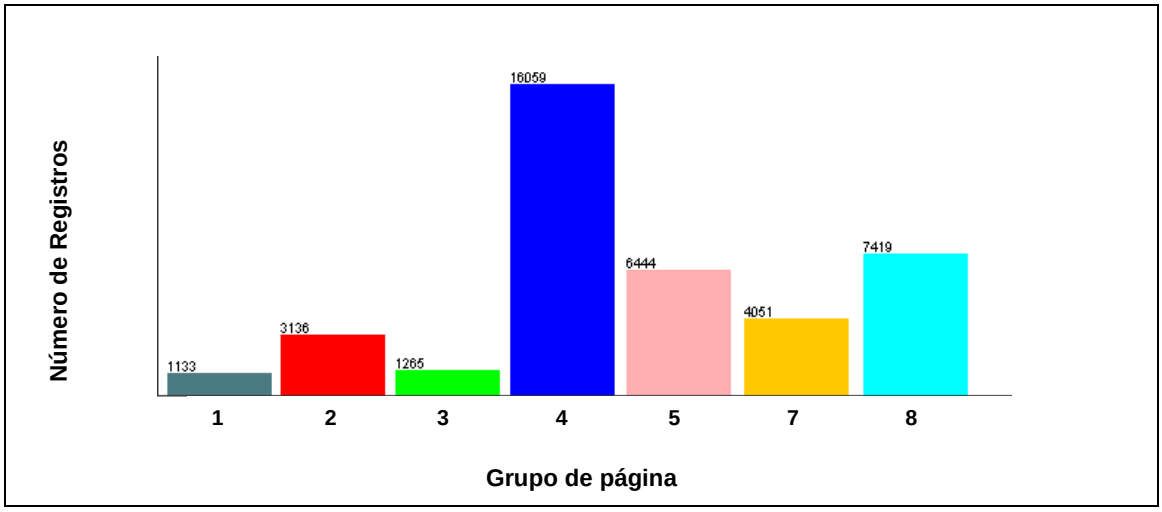


Figura 4.4 Diagrama de barras - registros de conexión por grupo de página

Las Figuras 4.1, 4.2, 4.3 y 4.4 muestran las frecuencias de algunas características individuales presentes en los registros de navegación de los usuarios involucrados en el estudio; claramente se pueden observar ciertos comportamientos generales en los datos. Por ejemplo, se puede determinar que los días de la semana con mayor uso del Sitio corresponde con los días lunes, martes y miércoles en horas de la mañana (grupos de hora 5, 6, 7 y 8 que comprende el rango entre las 8:37 am y 11: 28 am - Ver Tabla 4.5). Del mismo modo se puede observar que las paginas más visitadas por los usuarios se encuentran en el grupo de página 4, lo que corresponde con el submenú Proceso – Secciones (Ver Tabla 4.6).

Debido a que el objetivo del proceso de minería consiste en la obtención de información oculta en los datos, a continuación se presenta en detalle los resultados obtenidos en la aplicación de diversas técnicas de análisis aplicadas.

4.2 Resultados Prueba 1: Análisis de Varianza (ANOVA):

Para llevar a cabo esta primera prueba, se realizó el cálculo de un nuevo parámetro denominado duración en base a los campos: fecha y hora. Dicho parámetro proporcionó información acerca del periodo de tiempo (en segundos) que un usuario permanecía en cada página a lo largo de su sesión. Con base en este nuevo elemento, se llevó a cabo el análisis de varianza entre muestras de un mismo usuario y entre

muestras de usuarios diferentes, con el objetivo de establecer hipótesis acerca de las muestras estudiadas.

Para cada uno de los usuarios involucrados en el estudio se separaron los datos correspondientes a cada periodo académico y la comparación se hizo de acuerdo a estos lapsos. Los resultados obtenidos se muestran a continuación:

Tabla 4.1 Resultados de Análisis de Varianzas de un factor entre muestras de usuarios iguales

Datos de bitácora Módulo Departamentos del Site de Control de Estudios de la UNET
Rango de estudio: Registros de: Lapso 2007-1 (Muestra 1) / Lapso 2007-3 (Muestra 2)
Variable de estudio: Duración de usuario por página
Valor crítico de F para $\alpha = 0,05 \approx 3,85$

Usuario		N° Registros Muestra 1	Promedio Tiempo (seg.)	Varianza	Suma de Cuadrados entre Grupos	Suma de Cuadrados dentro de Grupos	F Calculado	Probabilidad
A	Muestra 1	880	592,567	5902260,578	279420,489	9229272219,454	0,0430517	0,835657
	Muestra 2	544	563,737	7442329,966				
B	Muestra 1	4607	77,848	1030382,836	4981865,054	9732745383,841	4,0928834	0,043097
	Muestra 2	3391	128,351	1471033,050				
C	Muestra 1	1281	276,984	2513054,863	2568452,812	5898809907,936	0,9374567	0,333041
	Muestra 2	874	347,296	3072279,132				
D	Muestra 1	359	444,768	3468629,697	698960,600	1618529015,732	0,242699	0,622455
	Muestra 2	205	371,580	1846860,705				
E	Muestra 1	426	260,542	2223788,375	47650755,528	7356935990,082	4,559796	0,033075
	Muestra 2	280	791,614	22981454,947				
F	Muestra 1	2685	132.333	1068480,963	3777013,014	5525096241,603	2,934055	0,086801
	Muestra 2	1609	193,604	1652545,607				
G	Muestra 1	176	303,278	1070230,007	22219371,664	2021490006,246	3,198550	0,074743
	Muestra 2	117	865,555	15812066,852				
H	Muestra 1	83	516,228	2066328,422	20578,231	237502947,238	0,011437	0,914995
	Muestra 2	51	490,705	1361280,331				
I	Muestra 1	7747	146,850	963946,770	8548302,879	18579791834,423	6,408998	0,011365
	Muestra 2	6185	196,706	1797066,647				
J	Muestra 1	3983	145,244	2098003,258	1579394,220	12106741917,342	0,945804	0,330822
	Muestra 2	3269	115,585	1148253,654				
K	Muestra 1	430	480,106	14240129,690	6790685,379	6530169944,771	0,783040	0,376495
	Muestra 2	325	288,569	1299858,974				

Tabla 4.2 Resumen de resultados del Análisis de Varianzas de un factor entre muestras de usuarios distintos

Datos de bitácora Módulo Departamentos del Site de Control de Estudios de la UNET
Rango de estudio: Registros de: Lapso 2007-1 (Cruce de datos de usuarios)
Variable de estudio: Duración de usuario por página
Valor crítico de F para $\alpha = 0,05 \approx 3,85$

Usuario		N° Registros Muestra 1	Promedio Tiempo (seg.)	Varianza	Suma de Cuadrados entre Grupos	Suma de Cuadrados dentro de Grupos	F Calculado	Probabilidad
A	Muestra A	4712	167,491	1884546,555	18718923,701	13624042166,495	12,801209	0,000348
	Muestra B	4607	77,848	1030382,836				
	Muestra A	4712	167,491	1884546,555	12074900,071	12094809049,330	5,981138	0,014488
	Muestra C	1281	276,984	2513054,863				
	Muestra A	4712	167,491	1884546,555	3382691,933	9823208883,382	1,768618	0,183612
	Muestra E	426	260,542	2223788,375				
	Muestra A	4712	167,491	1884546,555	9919558,739	9047537754,293	5,254959	0,021927
	Muestra H	83	516,228	2066328,422				
B	Muestra B	4607	77,848	1030382,836	44838226,177	5987712774,662	37,172283	1,16321678 E-09
	Muestra D	359	444,768	3468629,697				
	Muestra B	4607	77,848	1030382,836	5035774,208	7613746247,851	4,821646	0,028135
	Muestra F	2685	132,333	1068480,963				
	Muestra B	4607	77,848	1030382,836	13755213,432	12212675029,652	13,912136	0,000192
	Muestra I	7747	146,850	963946,770				
	Muestra B	4607	77,848	1030382,836	63639126,139	10854958979,931	29,51858	5,79891E-08
	Muestra K	430	480,106	14240129,690				
C	Muestra C	1281	276,984	2513054,863	7894106,232	4458479657,498	2,900214	0,08875
	Muestra D	359	444,768	3468629,697				
	Muestra C	1281	276,984	2513054,863	13282628,465	9325725862,766	2,434128	0,11890
	Muestra K	430	480,106	14240129,690				

Usuario		N°Registros Muestra 1	Promedio Tiempo (seg.)	Varianza	Suma de Cuadrados entre Grupos	Suma de Cuadrados dentro de Grupos	F Calculado	Probabilidad
D	Muestra D	359	444,768	3468629,697	30911047,089	4109572336,810	22,881068	1,80543E-06
	Muestra F	2685	132,333	1068480,963				
	Muestra D	359	444,768	3468629,697	244328,496	7350785068,889	0,0261586	0,871555
	Muestra K	430	480,106	14240129,690				
E	Muestra E	426	260,542	2223788,375	227466,098	1132400311,097	0,120522	0,728590
	Muestra G	176	303,278	1070230,007				
	Muestra E	426	260,542	2223788,375	5219350,391	8411841746,540	5,069913	0,024371
	Muestra I	7747	146,850	963946,770				
	Muestra E	426	260,542	2223788,375	5115890,045	9299359033,558	2,42443	0,119527
	Muestra J	3983	145,244	2098003,258				
F	Muestra F	2685	132,333	1068480,963	11865380,860	3037241835,650	10,805739	0,001024
	Muestra H	83	516,228	2066328,422				
	Muestra F	2685	132,333	1068480,963	267342,230	11222051878,818	0,158803	0,690273
	Muestra J	3983	145,244	2098003,258				
G	Muestra G	176	303,278	1070230,007	4210976,053	7654021938,158	4,357858	0,03687
	Muestra I	7747	146,850	963946,770				
	Muestra G	176	303,278	1070230,007	4209537,702	8541539225,176	2,048699	0,15241
	Muestra J	3983	145,244	2098003,258				
H	Muestra H	83	516,228	2066328,422	11204486,314	7636170617,451	11,48595	0,00070
	Muestra I	7747	146,850	963946,770				
	Muestra H	83	516,228	2066328,422	90776,052	6278454567,729	0,007388	0,93153
	Muestra K	430	480,106	14240129,690				
I	Muestra I	7747	146,850	963946,770	6787,933	15820980660,619	0,005031	0,94345
	Muestra J	3983	145,244	2098003,258				
	Muestra I	7747	146,850	963946,770	45244328,005	13575747323,879	27,24508	1,83636E-07
	Muestra K	430	480,106	14240129,690				

En la Tabla 4.1 se encuentran reflejados algunos de los resultados obtenidos por el análisis de varianza realizado sobre la duración por página de los usuarios, basado en la hipótesis de igualdad de medias poblacionales entre los datos de estudio (4.1). Al revisar algunos resultados, como por ejemplo, el caso del usuario A, se puede afirmar que existe una alta probabilidad (0,835657) para considerar que las medias poblacionales son iguales y por lo tanto inferir que no hay evidencia para pensar que ambas muestras no provienen de la misma población (en este caso muestras de un mismo usuario).

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \qquad (4.1)$$

Donde H_0 (para el presente caso de estudio) supone que la media de los tiempos que un usuario permanece en cada página del sitio es igual para las muestras estudiadas.

Sin embargo, aún cuando los resultados de dicha tabla muestran el resultado de la comparación entre muestras de usuarios iguales, se puede observar la presencia de valores del estadístico $F_{calculado} > F_{0,05}$ ($F_{calculado} > 3{,}85$), lo que proporciona indicios que muestran, que para algunos casos (Ejemplo: pruebas de usuarios B, E, G, I) debe ser rechazada la hipótesis nula acerca de la igualdad de las medias poblacionales, es decir, no es posible confiar totalmente de los resultados arrojados por éste estadístico para afirmar la similitud entre los tiempos de navegación por página de todos los individuos involucrados en el estudio, debido a que se pudiera estar incurriendo en un error tipo I (el rechazo de la hipótesis nula siendo esta verdadera) o un error tipo II (no rechazar la hipótesis nula siendo esta falsa).

Luego de haber obtenido los resultados de las comparaciones entre muestras de usuarios iguales, se procedió a realizar el mismo análisis, esta vez, cruzando información de tiempo de permanencia por página entre muestras de usuarios distintos. Un resumen de la información obtenida en esta fase se muestra en la Tabla 4.2. La información mostrada proporciona indicios de que los tiempos comparados efectivamente pertenecen a usuarios distintos. Tal es el caso del usuario B cuyos valores del estadístico $F_{calculado}$ para las combinaciones probadas hacen que se rechace la hipótesis de igualdad de medias con un alto grado de significancia. No obstante, al igual que en el primer análisis de varianza realizado, algunas comparaciones ofrecieron resultados dudosos (Ejemplo:

usuario A), que en ciertos cruces arrojó valores del estadístico $F_{calculado}$ (1,768618) que apuntaban a la hipótesis alterna.

4.3 Resultados Prueba 2: Cadenas de Markov

Una vez realizado un análisis de varianza entre las medias muestrales, se aplicaron cadenas de Markov de primer orden sobre los saltos de páginas en las sesiones de usuario durante los periodos de estudio (Lapsos Académicos 2007-1, 2007-3), con la finalidad de verificar el parecido (en términos de proporción) de los saltos observados, una vez más, partiendo de la hipótesis de igualdad en las mismas ($H_o: \theta_0 = \theta_1 \;\; Vs \;\; H_1: \theta_0 \neq \theta_1$).

Para ello fue desarrollada una aplicación en java que permitiera la generación de matrices de transición (4.2) por usuario para cada periodo de estudio.

Páginas	P1	P2	P3	…	Pn
P1					
P2					
P3					
…					
Pn					

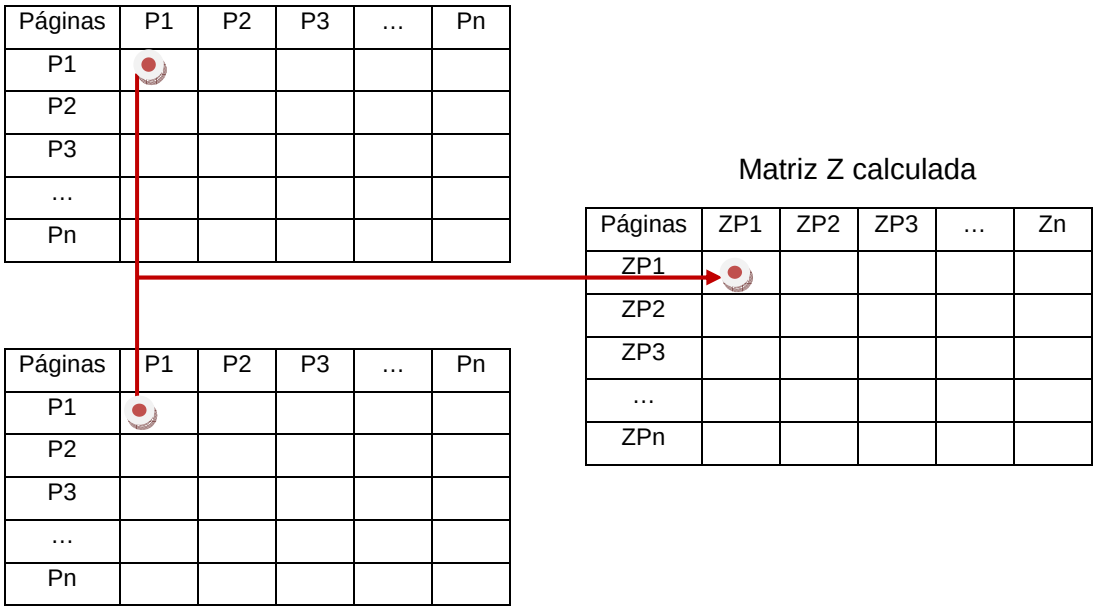
$$\sum_{i=0}^{n} P_{i,0} = 1$$

(4.2)

El resultado al finalizar el proceso, son 2 matrices de transición (una para cada período de estudio) por cada usuario. El procedimiento seguido para la comparación de ambas tuvo basamento en hipótesis relativas a proporciones para cada transición en la matriz (4.3)

A partir de la matriz Z calculada y tomando en cuenta sólo los elementos para los cuales existe información (posiciones que dan muestra de transiciones entre páginas), fue calculado un porcentaje que indica, del total de las transiciones tomadas, la porción que representan aquellas transiciones que inducen a no rechazar H_o (igualdad entre las proporciones comparadas); y otro porcentaje que indica lo contrario, de acuerdo a cierto nivel de significancia. La comparación se llevó a cabo entre usuarios iguales y variando el

nivel de significancia (α) entre 0.01 y 0.05. Los resultados obtenidos se muestran en las Tabla 4.3 y 4.4.



(4.3)

Usuario	**Porcentaje** $(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}})$	**Porcentaje** $(Z < -Z_{\frac{\alpha}{2}}\ ,\ Z > Z_{\frac{\alpha}{2}})$
A	75,00	25,00
B	26,09	73,91
C	61,90	38,10
D	52,38	47,62
E	11,76	88,24
F	0,00	100,00
G	11,76	88,24
H	15,79	84,21
I	46,15	53,85
J	5,00	95,00
K	52,38	47,62

Tabla 4.3 Resumen de resultados del Análisis de Proporciones entre muestras de usuarios iguales con $\alpha = 0.01$

Usuario	Porcentaje $(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}})$	Porcentaje $(Z < -Z_{\frac{\alpha}{2}}\ ,\ Z > Z_{\frac{\alpha}{2}})$
A	65,00	35,00
B	21,74	78,26
C	52,38	47,62
D	52,38	47,62
E	11,76	88,24
F	0,00	100,00
G	0,00	100,00
H	5,26	94,74
I	26,92	73,08
J	5,00	95,00
K	28,57	71,43

Tabla 4.4 Resumen de resultados del Análisis de Proporciones entre muestras de usuarios iguales con $\alpha = 0.05$

Como se observa en las Tablas 4.3 y 4.4, existe una alta proporción de usuarios cuyas matrices de transición entre páginas para el lapso académico 2007-1 difiere de sus matrices de transición para el siguiente lapso de estudio. Por tal motivo se rechaza la hipótesis nula que plantea la igualdad de proporción de saltos entre páginas para las sesiones de los usuarios analizados.

4.4 Resultados Prueba 3: Análisis de Secuencias

Las pruebas realizadas en esta fase del estudio se basaron (como ya se mencionó) en la implementación de un algoritmo de alineamiento global de secuencias denominado "Algoritmo de Needleman-Wunsch" utilizado para el alineamiento de secuencias de proteínas.

La implementación (al igual que el método anterior) fue desarrollada en el lenguaje de programación Java. Los Datos de entrada para ésta fueron las secuencias de páginas por sesión de usuario. Dichas secuencias reflejan las direcciones seguidas por el usuario antes y después de estar en una página específica (4.4).

Patrón de Secuencia A_1 = *< p1, p2, p3, p4, p1, p6, p3, p18, p1, p24, p3, p5 >*

Patrón de Secuencia A_2 = *< p1, p2, p7, p15, p1, p24, p3, p6, p1, p2, p3 >*

Donde A_1 y A_2 son patrones de navegación del usuario A

(4.4)

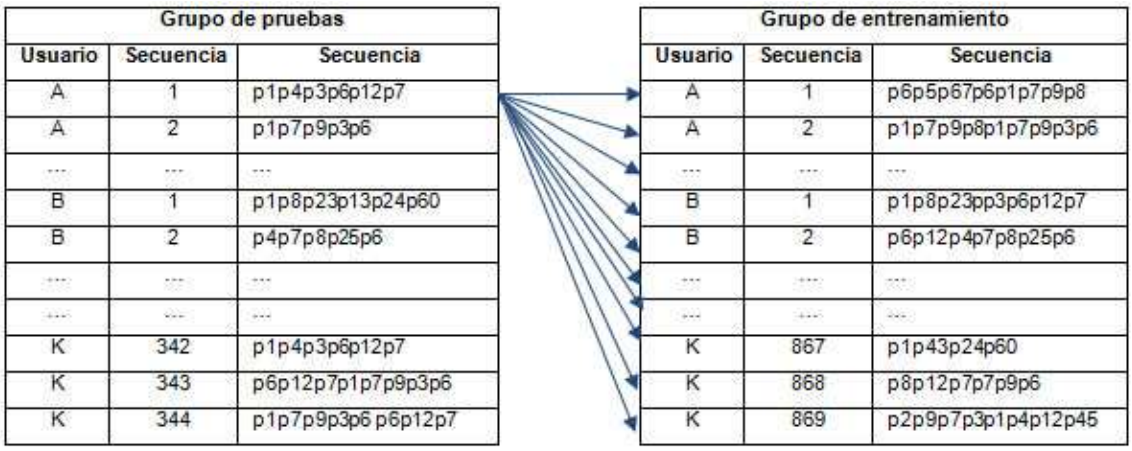
Antes de Comenzar la implementación, fue necesario el establecimiento de la función de similitud (4.5) necesaria para el cálculo de la puntuación del alineamiento encontrado

$$S[ij] = \begin{cases} 1\ si\ a_i = b_j \\ 0\ si\ a_i \neq b_j \\ -1\ si\ gap \end{cases}$$

(4.5)

Para detectar la presencia de una nueva sesión de usuario, se llevó a cabo el cálculo del tiempo en segundos que un usuario permanecía en cada página durante todo el recorrido en el Sitio. La principal razón, es debido a que el Sistema en estudio se encuentra configurado para que la sesión finalice automáticamente luego de 180 segundos de inactividad en la página (parámetro session.cache_expire = 180 en el archivo de configuración php.ini). Lo que puede originar que existan registros de navegación para un mismo usuario durante tiempos prolongados sin la presencia de la huella en la página de cierre de sesión.

Una vez detectadas las sesiones por usuario, se inició el proceso de prueba del algoritmo sobre las secuencias encontradas. Para ello se propuso la separación de dichas secuencias en dos grandes grupos. El primero de ellos (denominado grupo de entrenamiento) estuvo conformado por cerca del 70% del total de las secuencias de cada usuario, la diferencia fue considerada como el grupo de pruebas tomado para validar el método. Dichas pruebas consistieron en la comparación (alineación) de cada secuencia perteneciente al grupo de pruebas con cada secuencia presente en el grupo de entrenamiento; proporcionando esto, un puntaje de alineación durante cada iteración (Ver Figura 4.5).



Grupo de pruebas		
Usuario	Secuencia	Secuencia
A	1	p1p4p3p6p12p7
A	2	p1p7p9p3p6
...	...	...
B	1	p1p8p23p13p24p60
B	2	p4p7p8p25p6
...	...	...
...	...	...
K	342	p1p4p3p6p12p7
K	343	p6p12p7p1p7p9p3p6
K	344	p1p7p9p3p6 p6p12p7

Grupo de entrenamiento		
Usuario	Secuencia	Secuencia
A	1	p6p5p67p6p1p7p9p8
A	2	p1p7p9p8p1p7p9p3p6
...	...	...
B	1	p1p8p23pp3p6p12p7
B	2	p6p12p4p7p8p25p6
...	...	...
...	...	...
K	867	p1p43p24p60
K	868	p8p12p7p7p9p6
K	869	p2p9p7p3p1p4p12p45

Figura 4.5 Proceso de alineamiento entre registros de prueba y de entrenamiento

Debido a que cada secuencia en el grupo de pruebas genera, según lo descrito, un puntaje de alineación con cada secuencia en el grupo de entrenamiento, se propuso el cálculo de un promedio entre los valores arrojados por usuario. Es decir, haciendo referencia a la **Figura 4.5,** el proceso de alinear la secuencia 1 del usuario A en el grupo de pruebas con todas las secuencias de A en el grupo de entrenamiento produciría un promedio de alineamiento con A (P_A); asimismo, al alinear la misma secuencia A del grupo de pruebas con todas las secuencias de B en el grupo de entrenamiento produciría un promedio de alineamiento con B (P_B). El proceso se repite y al culminar se tendrá un valor de alineamiento de dicha secuencia con todos los usuarios del grupo de entrenamiento. A partir de estos promedios, se selecciona el mayor valor de alineamiento para conocer el usuario cuyas secuencias en promedio (en el grupo de entrenamiento) tiene un mayor parecido con la secuencia analizada.

Los resultados observados en esta prueba (Ver Tabla 4.5), muestran que el alineamiento de secuencias de navegación a través del algoritmo de Needleman-Wunsch (para los datos estudiados) arroja una matriz de confusión cuyos máximos valores asociados a cada usuario no se encuentran en la diagonal principal. Un resumen de estos (Ver Tabla 4.6) indican, que a pesar de que se puede obtener un porcentaje de alineación del usuario H (65,12%), en general, el porcentaje de alineación correcta de las secuencias tomadas para el estudio se encuentra alrededor del 22%.

	Alineado como usuario:										
Secuencia de Usuario	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
A	12	0	2	1	1	0	4	22	2	2	3
B	3	21	0	0	0	0	12	11	0	0	0
C	3	0	1	2	2	0	7	24	3	5	1
D	3	0	2	2	1	0	13	11	1	7	6
E	0	0	0	3	16	0	10	17	0	0	1
F	10	4	4	2	2	0	7	9	0	2	5
G	1	0	4	0	8	0	13	13	4	3	3
H	0	0	0	0	6	0	1	28	3	0	5
I	5	0	3	1	0	0	11	21	4	1	4
J	5	0	2	0	4	0	4	14	0	8	5
K	3	0	0	0	0	0	18	17	1	1	5

Tabla 4.5 Matriz de confusión: prueba de alineamiento de secuencias a través del algoritmo de Needleman-Wunsch

Secuencias de usuario	Alineaciones correctas	% de alineaciones correctas	Alineaciones incorrectas	% de alineaciones incorrectas
A	12	24,49	37	75,51
B	21	44,68	26	55,32
C	1	2,08	47	97,92
D	2	4,35	44	95,65
E	16	34,04	31	65,96
F	0	0,00	45	100,00
G	13	26,53	36	73,47
H	28	65,12	15	34,88
I	4	8,00	46	92,00
J	8	19,05	34	80,95
K	5	11,11	40	88,89

Tabla 4.6 Resumen de resultados: prueba de alineamiento de secuencias a través del algoritmo de Needleman-Wunsch

4.5 Resultados Prueba 4: Bayes Ingenuo

Antes de dar inicio a esta prueba se realizó la clasificación de algunos atributos (hora y página), así como la generación de otros nuevos. Para el primer caso, se obtuvo una agrupación de las horas de navegación (Ver Tabla 4.7) y de las páginas del Sitio (Ver Tabla 4.8), esta última con base en los sub-módulos internos del mismo. Para el segundo caso, se obtuvo un nuevo atributo a partir de la fecha de navegación, el cual hace referencia a los días de la semana en que un usuario inició sesión.

Clase	Rango de Hora
1	00:00 - 05:25
2	05:26 - 06:14
3	06:15 - 07:53
4	07:54 - 08:36
5	08:37 - 09:19
6	09:20 - 10:02
7	10:03 - 10:45
8	10:46 - 11:28
9	11:29 - 12:11
10	12:12 - 12:56
11	12:57 - 13:44
12	13:45 - 14:34
13	14:35 - 15:17
14	15:18 - 16:00
15	16:01 - 16:43
16	16:44 - 17:26
17	17:27 - 18:09
18	18:10 - 18:52
19	18:53 - 19:35
20	19:36 - 20:18
21	20:19 - 21:01
22	21:02 - 21:44
23	21:45 - 22:30
24	22:21 - 23:59

Tabla 4.7 Clasificación de Horas de Navegación

Clase	Grupo de Página
1	Información Alumnos
2	Información Administrativa
3	Consultas- Secciones
4	Procesos - Secciones
5	Consultas - Permisos
6	Procesos - Permisos
7	Otras opciones

Tabla 4.8 Clasificación de Páginas del Sitio

Luego de esto, los datos fueron introducidos en la herramienta Weka para realizar la prueba en cuestión. Los atributos tomados en cuenta en esta fase fueron: dirección IP, día en que se estableció la conexión, rango de hora, páginas visitadas (con base en la clasificación previamente realizada), duración y usuario. Este último tomado como el

atributo clasificador. La matriz de confusión (Ver Tabla 4.9) y la información adicional arrojada por la aplicación (Ver Tabla 4.10) utilizando un 70% de los datos para el proceso de entrenamiento, se muestra a continuación:

	Clasificado como usuario:										
Usuario	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
A	4178	1	0	25	0	0	3	0	0	4	1
B	25	5	0	11	0	0	0	0	0	0	0
C	0	0	48	7	9	0	1	5	0	8	0
D	9	0	0	1157	0	0	0	26	9	39	0
E	3	0	14	44	141	0	6	0	0	0	0
F	0	0	1	1	0	165	0	0	0	0	0
G	28	0	0	0	0	0	612	0	6	0	0
H	2	0	0	8	0	0	0	2414	0	9	0
I	0	0	0	11	0	0	31	1	395	16	0
J	1	0	3	64	5	3	0	59	3	2010	0
K	80	0	6	10	0	1	0	0	1	0	127

Tabla 4.9 Matriz de Confusión - Variable clasificadora: Usuario

```
Time taken to build model: 0.03 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          11252     (94.9376 %)
Incorrectly Classified Instances          600     (5.0624 %)
Kappa statistic                          0.9349
Mean absolute error                      0.0148
Root mean squared error                  0.0849
Relative absolute error                 10.4073 %
Root relative squared error             31.8308 %
Total Number of Instances                11852
```

Tabla 4.10 Sumario de salida - Variable clasificadora: Usuario

La matriz de confusión muestra para este caso, un alto grado de clasificación correcta de cada usuario. Del 30% de los datos utilizados para el proceso de prueba, el algoritmo logró clasificar correctamente el 94,93% de estos, lo que pudiera considerarse

como una salida satisfactoria para la identificación de usuarios con base en los atributos mencionados anteriormente.

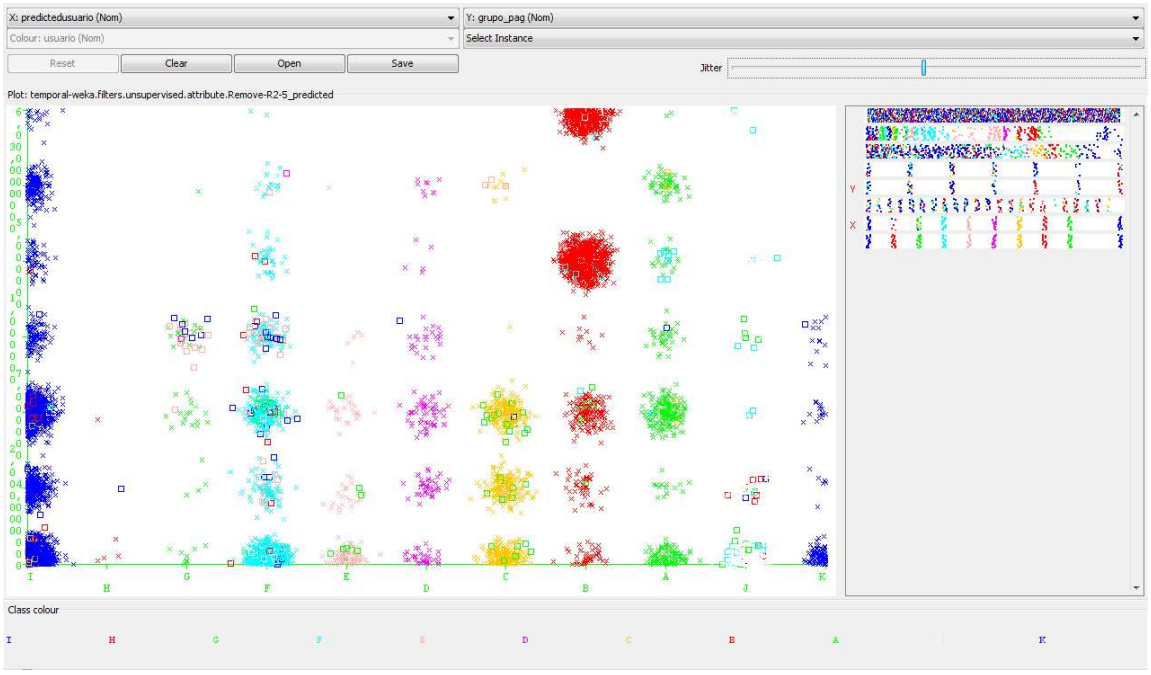


Figura 4.6 Visualización del error de clasificación sobre los datos Bayes Ingenuo: Variable clasificadora: usuario

Gráficamente también es posible visualizar el poder de clasificación que posee la técnica de Bayes Ingenuo sobre los datos en estudio. En la imagen (Ver Figura 4.6), se encuentra graficado el error de clasificación de la variable predicha usuario, sobre la información suministrada; y puede observarse una baja presencia de puntos de distinto color (error de clasificación) sobre las aglomeraciones de puntos de una misma coloración (clasificaciones realizadas correctamente).

4.6 Resultados Prueba 5: Clustering

Para llevar a cabo esta parte del estudio, fue utilizado el método de Clustering de SimpleKMeans proporcionado por la herramienta Weka en modo explorador, con un porcentaje de división de los datos (porcentaje de entrenamiento) de 70%. Las pruebas realizadas incluyeron la variación de los atributos (duración_seg, dia, grupo_pag, grupo_hora y usuario) incluidos en cada corrida.

Debido al conocimiento a priori que debe tener el algoritmo del número inicial de clusters, se propuso la asignación de un valor de K (número de clusters) igual al número de usuarios involucrados en el estudio (11), ya que este valor (teóricamente) debería representar la cantidad de divisiones adecuadas en caso de que el método se ajuste al objetivo planteado.

Algunos de las salidas arrojadas por la herramienta se muestran en las tablas 4.11, 4.12 y 4.13. Los resultados arrojados por los reportes del K-medias, en todos los casos muestran un porcentaje de clasificación incorrecta de los datos que se encuentra sobre el 75% (Ver Figura 4.7) bajo los atributos de clasificación suministrados. Sin embargo, esto no indica ineficiencia del algoritmo, sino simplemente la existencia de objetos que a pesar de su pertenencia a una clase, comparte la mayoría de sus características con las de objetos de otras clases, motivo por el cual son clasificados en una clase diferente de acuerdo a los criterios de clasificación del propio algoritmo.

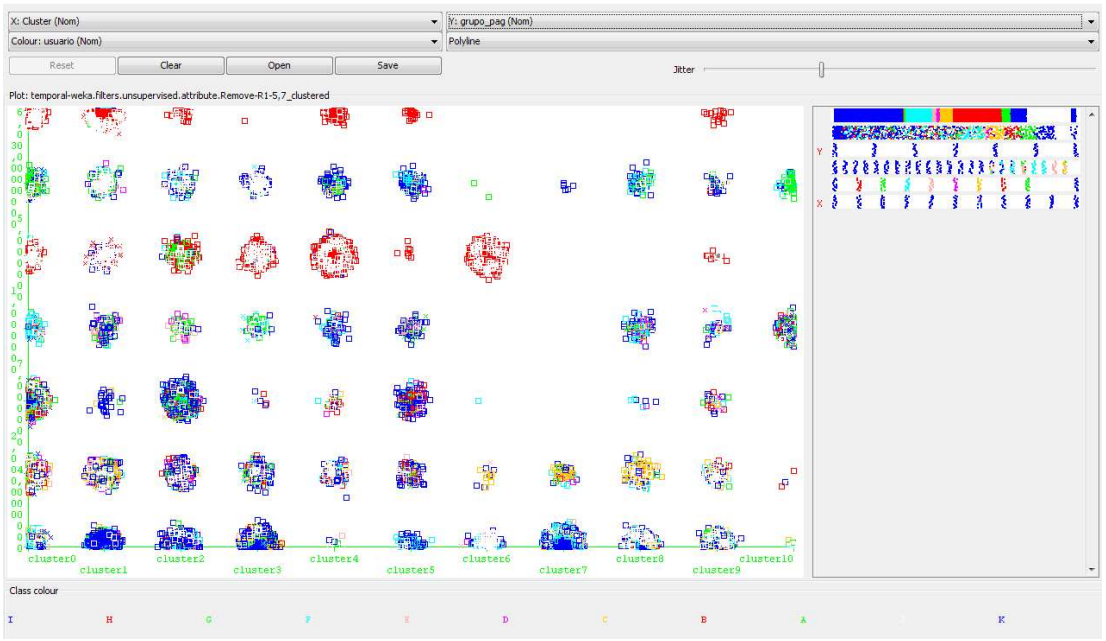


Figura 4.7 Visualización de Clusters encontrados sobre los datos
Variable clasificadora: usuario

Tabla 4.11 Resumen de resultados prueba de Clustering – Weka

Datos de bitácora Módulo Departamentos del Site de Control de Estudios de la UNET
Número de Instancias: 39507
Número de Iteraciones: 4
Cantidad de Atributos: 5
Variable Clasificadora: usuario
Variables Clasificadas: duracion_seg / dia / grupo_pag / grupo_hora
Instancias clasificadas de manera incorrecta: 30918 (78.2595%)

	Asignado a Cluster:										
Usuario	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
I	2950	1475	1961	2119	201	1939	174	1108	1392	529	84
H	38	0	40	7	13	0	1	7	21	7	0
G	98	10	37	38	5	15	23	15	38	5	9
F	771	467	411	1115	215	624	19	260	158	220	34
E	151	85	61	211	24	51	0	27	53	5	38
D	167	107	115	42	5	49	3	33	8	23	12
C	565	125	377	362	35	141	129	139	224	57	1
B	718	2692	488	179	1904	39	1779	47	45	64	43
A	389	120	324	60	16	156	73	41	145	43	57
J	833	1669	495	786	579	772	495	243	1112	234	34
K	132	205	63	92	17	123	0	90	25	7	1

Tabla 4.12 Resumen de resultados prueba de Clustering – Weka

Datos de bitácora Módulo Departamentos del Site de Control de Estudios de la UNET
Número de Instancias: 39507
Número de Iteraciones: 3
Cantidad de Atributos: 4
Variable Clasificadora: usuario
Variables Clasificadas: dia / grupo_pag / grupo_hora
Instancias clasificadas de manera incorrecta: 30734 (77.7938%)

	Asignado a Cluster:										
Usuario	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
I	2722	1253	1410	2416	64	1195	2167	1622	925	85	73
H	49	0	31	6	14	0	5	23	6	0	0
G	105	12	19	39	2	20	14	60	4	18	0
F	789	408	320	839	180	552	408	165	596	36	1
E	136	95	81	200	14	23	55	55	2	38	7
D	149	120	123	64	0	37	23	8	29	11	0
C	475	72	349	411	25	228	205	309	78	0	3
B	739	2555	1231	177	2759	85	52	348	15	37	0
A	342	137	291	140	2	167	44	181	55	63	2
J	784	1529	589	561	851	73	957	879	703	30	296
K	143	81	52	198	0	26	237	5	13	0	0

Tabla 4.13 Resumen de resultados prueba de Clustering – Weka

Datos de bitácora Módulo Departamentos del Site de Control de Estudios de la UNET
Número de Instancias: 39507
Número de Iteraciones: 3
Cantidad de Atributos: 4
Variable Clasificadora: usuario
Variables Clasificadas: duracion_seg / grupo_pag / grupo_hora
Instancias clasificadas de manera incorrecta: 30519 (77.9426%)

	Asignado a Cluster:										
Usuario	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
I	3069	1182	2477	2378	240	378	39	2899	740	440	90
H	47	2	32	36	5	0	0	1	0	11	0
G	82	24	59	49	4	0	0	0	24	26	25
F	714	200	488	910	213	62	31	1216	249	94	117
E	114	50	105	92	19	40	22	208	20	21	15
D	209	96	66	51	17	24	3	37	14	29	18
C	676	160	389	485	9	29	8	198	142	55	4
B	690	2388	414	377	2158	81	1761	30	27	47	25
A	487	130	289	205	37	53	2	106	20	35	60
J	852	1519	276	656	803	27	503	2050	318	202	46
K	165	25	62	123	8	9	0	287	13	14	49

4.7 Resumen y Comparación de Resultados

En la primera etapa del trabajo de investigación se llevó a cabo una exploración de los datos utilizando un enfoque tradicional estadístico a través del análisis de varianza de un factor (ANOVA) sobre el parámetro duración, correspondiente al tiempo de permanencia de un usuario en cada página del sitio. Los resultados de esta fueron mostrados desde dos perspectivas, la primera de ellas haciendo uso de información de usuarios iguales, en donde se observó que del número total de resultados obtenidos para el estadístico $F_{calculado}$, 8 de 11 veces (72,72%) el valor se encontró dentro del límite de aceptación de la hipótesis nula ($F_{calculado}$< 3,85), pudiendo esto permitir hacer inferencias acerca de que las muestras provenían de la misma población. Sin embargo al realizar el mismo análisis, sólo que sobre la mezcla de información de usuarios (lo que ocurriría cuando el comportamiento de un usuario no es el habitual o cuando se está en presencia de un impostor), del número total de resultados obtenidos para el estadístico $F_{calculado}$, 10 de 23 veces (43,47%) el valor se encontró dentro del límite de aceptación, representando esto un porcentaje considerable de falsos positivos para el caso de la comprobación de la hipótesis nula.

Durante la segunda etapa, el análisis se enfocó en la verificación del parecido (en términos de proporción) de los saltos entre páginas registrados en la bitácora de navegación (Base de Datos utilizada para el estudio) de los usuarios del Sitio. Para ello se utilizaron cadenas de Markov de primer orden y adicionalmente se llevó a cabo el cómputo del estadístico Z (denominado para este caso Matriz $Z_{calculada}$) proveniente de la diferencia de proporciones de las matrices markovianas obtenidas en cada periodo de estudio (2007-1 y 2007-3). Los resultados obtenidos en este caso, muestran que 10 de 11 casos (90,90%), las proporciones que se encontraron dentro del rango de aceptación del mencionado estadístico (-2,57 <Z< 2,57 utilizando un $\alpha = 0{,}01$) no superan el 70%.

La tercera etapa se basó en el alineamiento de secuencias de navegación por sesión de usuario utilizando para ello el algoritmo de Needleman-Wunsch. Los resultados indican que el promedio total de alineamiento de secuencias de los usuarios se encuentra alrededor del 22%; y que 10 de 11 casos (90,90%), muestran un alineamiento por debajo del 50%, razón por la cual no se recomendó la técnica para tratar de determinar si las secuencias de saltos de un usuario durante una sesión poseen cierto grado de similitud.

La siguiente técnica puesta a prueba fue la de Bayes Ingenuo a través de la herramienta de minería Weka. Los resultados obtenidos muestran un grado de clasificación correcta de los usuarios de 94,93 % de acuerdo a los atributos tomados en cuenta (IP, día, rango_hora, grupo_página, duración y usuario). Esto es posible visualizarlo en la matriz de confusión, la cual muestra los mayores valores de clasificación asociados a cada usuario ubicados en la diagonal principal para el caso del análisis de los registros de usuarios reales. Mientras que para el caso de registros de impostores, la traza de la matriz de confusión no sigue el comportamiento descrito.

Con base en la asunción de independencia de características planteada por la técnica de Bayes Ingenuo [31] y junto al nivel de predicción alcanzado con la misma en los experimentos realizados, se puede interpretar que los atributos tomados en cuenta (IP, día, rango_hora, grupo_página, duración y usuario) son independientes entre sí y podrían ser utilizados para describir patrones de comportamiento de los usuarios del Sitio de Control de Estudios de la Universidad del Táchira.

Aunque es posible que en la realidad los atributos seleccionados guarden algún tipo de dependencia entre sí, asumir independencia entre ellos para este caso, permitió generar un modelo que describe el comportamiento de los usuarios registrados en el sitio.

A partir de lo anterior se puede deducir por ejemplo, que la dirección ip desde la cual un usuario se conecta al sistema no necesariamente depende de la hora o el día en que se realizó dicha conexión. Mientras que la identidad de un usuario si depende de una o más de estas variables.

A continuación se planteo realizar pruebas utilizando la técnica de Clustering a la espera de mejorar los resultados alcanzados con el método anterior. Para ello se utilizó la herramienta de minería Weka. Los resultados obtenidos utilizando un total de 11 clusters y alternando el número de atributos incluidos en la prueba, se observó un porcentaje de clasificación incorrecta que se encontró sobre el 75%.

La tabla a continuación (Ver Tabla 4.14) muestra la recopilación de la precisión alcanzada por cada una de las técnicas utilizadas durante el proceso de pruebas.

Prueba	Precisión alcanzada
ANOVA	72,72 %
Cadenas de Markov	Aprox. 10 %
Alineamiento de Secuencias mediante algoritmo de Needleman-Wunsch	Aprox. 10 %
Bayes Ingenuo	94,93%
Clustering	Aprox. 25 %

Tabla 4.14 Precisión alcanzada por las técnicas de minería empleadas en la fase de pruebas

4.8 Validación de Modelo

Una vez culminado el proceso de pruebas y con base en los resultados obtenidos en cada fase, se propuso la validación del modelo que mostró el mejor desempeño utilizando para ello registros de navegación de un lapso distinto a los utilizados para los experimentos. Debido a que la técnica con mejor resultado mostrado fue la de Bayes Ingenuo, a continuación se procedió a tomar los datos del lapso 2008-1 (periodo académico comprendido entre 22-04-08 y 08-08-08), con el fin de aplicarles el modelo obtenido durante el entrenamiento.

Como ya se hizo mención, la primera etapa de este proceso de validación consistió en probar el modelo obtenido con los nuevos registros de los usuarios. Los resultados obtenidos se observan a través de una matriz de confusión y un resumen proporcionado por la herramienta (Ver Tablas 4.15 y 4.16). Se puede observar un elevado porcentaje de clasificación correcta (94.38%) de los registros de los usuarios que formaron parte del trabajo de investigación.

	Clasificado como usuario:										
Usuario	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
A	233	3	0	0	1	43	0	0	0	0	0
B	3	2798	0	0	9	0	1	0	0	0	0
C	0	0	50	0	0	0	0	0	0	0	0
D	0	0	0	56	1	0	0	0	0	0	0
E	0	0	13	0	1	0	0	0	0	0	0
F	1	26	0	0	0	611	1	0	0	3	0
G	0	3	0	0	0	6	13	0	0	1	0
H	0	0	0	0	0	0	0	4	0	0	0
I	0	16	2	0	24	15	1	0	2494	1	1
J	0	178	0	0	2	0	0	0	0	885	0
K	0	0	0	0	0	15	2	0	52	1	5

Tabla 4.15 Matriz de Confusión para prueba de validación utilizando Bayes Ingenuo
Variable clasificadora: Usuario

Time taken to build model: 0.29 seconds
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances	7150	(94.3894 %)
Incorrectly Classified Instances	425	(5.6106 %)
Kappa statistic	0.9213	
Mean absolute error	0.019	
Root mean squared error	0.0948	
Relative absolute error	13.6123 %	
Root relative squared error	36.2016 %	
Total Number of Instances	7575	

Tabla 4.16 Sumario de salida para prueba de validación utilizando Bayes Ingenuo
Variable clasificadora: Usuario

La segunda etapa de este proceso consistió en la mezcla de los registros de los usuarios a fin de tener combinaciones de estos y verificar así la capacidad de la técnica para indicar la baja relación de esta nueva información (creada a partir de información falsa) con el modelo entrenado a partir de datos reales. La mezcla se llevó a cabo a través

de una pequeña aplicación en Java que permitió asignar a cada registro de la información original un nuevo usuario seleccionado al azar, esto con la finalidad de conservar el carácter aleatorio del experimento. La matriz de confusión y el resumen de salida (Ver Tablas 4.17 y 4.18) muestran para este caso un bajo porcentaje de clasificación correcta de los usuarios (2.81%).

	Clasificado como usuario:										
Usuario	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
A	0	232	13	9	5	62	1	0	93	82	0
B	58	1	1	0	1	79	0	0	177	132	0
C	175	22	0	0	2	442	1	0	179	184	1
D	1	171	0	0	6	15	1	1	282	206	0
E	2	202	1	0	2	77	13	3	516	131	0
F	0	154	0	0	2	0	0	0	125	45	0
G	1	716	0	0	1	0	0	0	161	5	0
H	0	84	0	0	4	0	0	0	383	12	0
I	0	1027	0	0	6	15	1	0	210	2	3
J	0	412	0	0	3	0	1	0	228	0	2
K	0	3	50	47	6	0	0	0	192	92	0

Tabla 4.17 Matriz de Confusión para prueba de validación (datos alterados) utilizando Bayes Ingenuo - Variable clasificadora: Usuario

```
Time taken to build model: 0.28  seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         213    (2.8119 %)
Incorrectly Classified Instances      7362        (97.1881 %)
Kappa statistic                         -0.0781
Mean absolute error                      0.1759
Root mean squared error                  0.4074
Relative absolute error                108.2837 %
Root relative squared error            134.8627 %
Total Number of Instances             7575
```

Tabla 4.18 Sumario de salida para prueba de validación (datos alterados)

utilizando Bayes Ingenuo - Variable clasificadora: Usuario

Gráficamente también es posible observar que la técnica en presencia de datos alterados de usuario (valores inusuales de sus atributos), posee una baja capacidad de clasificación correcta. En la imagen (Ver Figura 4.8), se puede visualizar una aglomeración de puntos de distintos colores en diversos sectores, lo que da muestra del alto grado de error de clasificación.

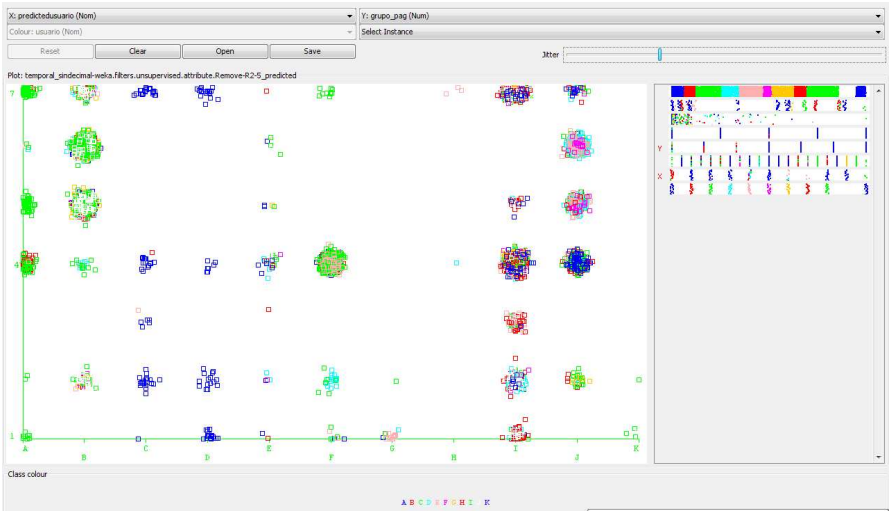


Figura 4.8 Visualización del error de clasificación sobre los datos alterados con Bayes Ingenuo Variable clasificadora: usuario

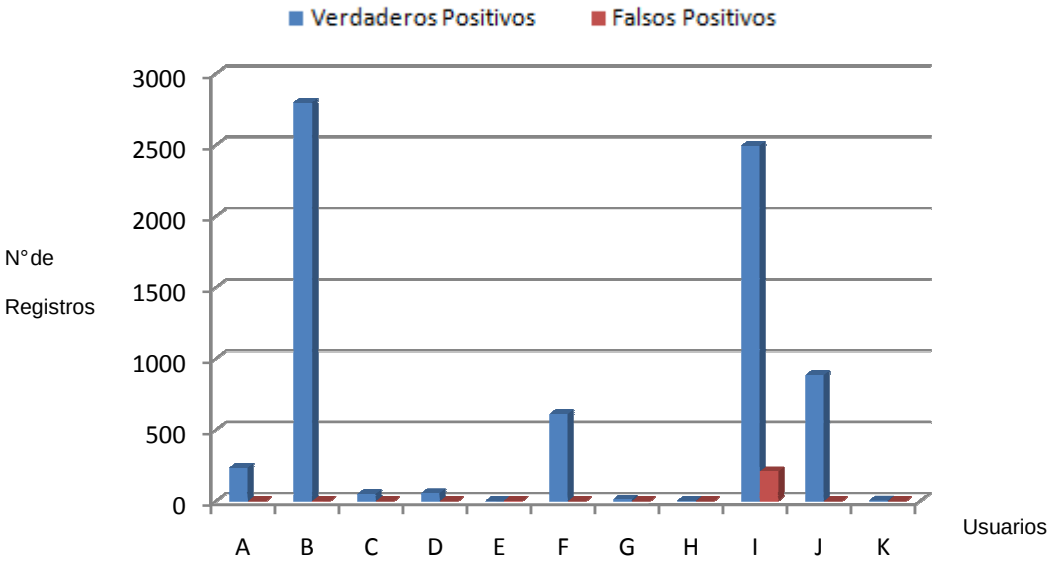


Figura 4.9 Comparación de falsos positivos / verdaderos positivos de prueba de aceptación de usuarios con Bayes Ingenuo

La figura 4.9 muestra el porcentaje de verdaderos positivos (área azul) y falsos positivos (área roja) obtenidos de la diagonal principal de las tablas 4.15 y 4.16 respectivamente y que representan las veces en las que cada usuario fue clasificado correctamente (para el primer caso) y las veces en las que los datos de un impostor fueron tomados como los de el usuario real (en el segundo caso).

Para complementar la fase de validación, se tomaron registros de usuarios que no formaron parte del estudio y se etiquetaron como registros de usuarios con los cuales se generó el modelo, con el fin de probar el comportamiento de este en presencia de huellas de navegación no habituales. Los resultados obtenidos se muestran en las tablas 4.19 y 4.20. A diferencia de la salida obtenida ante datos propios de cada usuario (Ver Tabla 4.15), la matriz de confusión de esta prueba no posee una aglomeración clara en su diagonal principal, es decir, el modelo no logra clasificar claramente el usuario de acuerdo a los registros suministrados El porcentaje de clasificación correcta para este caso es de 3.36%, lo que permite indicar que la técnica seleccionada (Bayes Ingenuo) proporciona modelos que facilitan el alcance del objetivo planteado.

	A	B	C	D	E	F	G	H	I	J	K
A	62	2	213	168	8	15	35	52	368	346	144
B	39	14	181	558	45	5	23	48	1686	4062	1277
C	154	0	10	1909	1	0	0	30	15	30	6
D	33	0	35	1	2	29	32	5	109	276	42
E	16	1	6	42	1	10	14	55	126	309	94
F	38	6	136	137	16	27	75	32	1528	940	1349
G	13	1	36	17	0	7	0	16	58	62	83
H	11	6	0	0	9	0	0	4	0	104	0
I	4492	74	1	3	5191	0	1	4	574	3539	53
J	68	611	430	3464	2	44	72	61	809	578	919
K	22	1	12	102	1	68	13	13	114	350	48

Tabla 4.19 Matriz de Confusión para prueba de validación (nuevos datos alterados) utilizando Bayes Ingenuo - Variable clasificadora: Usuario

Time taken to build model: 0.31 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances	1319	(3.3657 %)
Incorrectly Classified Instances	37870	(96.6343%)
Kappa statistic	-0.0914	
Mean absolute error	0.1732	
Root mean squared error	0.3267	
Relative absolute error	103.3033 %	
Root relative squared error	110.5094 %	
Total Number of Instances	39189	

Tabla 4.20 Sumario de salida para prueba de validación (nuevos datos alterados) utilizando Bayes Ingenuo - Variable clasificadora: Usuario

CAPITULO V

Conclusiones y Perspectivas

El objetivo del trabajo de investigación se enmarcó en la búsqueda de modelos que permitieran determinar la existencia patrones de comportamiento de los usuarios del Sitio Web de Control de Estudios de la UNET a través de sus registros de navegación. El proceso de selección de las técnicas, se realizó de acuerdo a las características de los datos y su evaluación se llevó a cabo de manera incremental, es decir, el proceso se inició con el enfoque tradicional estadístico y paulatinamente, a medida que fueron obtenidos los resultados de cada prueba, se experimentó con otras técnicas para tratar de encontrar aquellas que proporcionaran el mayor rendimiento en términos de tratar de descubrir si un usuario es quien decía ser con base en la información almacenada.

Para realizar el proceso de pruebas con cada técnica, fue necesario el tratamiento de los datos originales. En primer lugar, adecuarlos al tipo de prueba y en segundo lugar, obtener parámetros que proporcionaran nueva información para encontrar los modelos buscados. Esto fue logrado con el desarrollo de una aplicación en Java que se encargó de tomar los datos originales y de generar la nueva información en tablas adicionales, que luego serviría de entrada a las técnicas seleccionadas.

Durante las pruebas de cada modelo se utilizaron registros de navegación de periodos académicos similares (2007-1 y 2007-3), esto con el objetivo de evaluar comportamientos en lapsos de tiempo relativamente parecidos. Las técnicas evaluadas en el presente trabajo fueron el Análisis de Varianza de un solo factor (ANOVA), Cadenas de Markov, Alineamiento de Secuencias mediante el algoritmo de Needleman-Wunsch, Bayes Ingenuo y Clustering. Cada una de ellas proporcionó una salida que permitió evaluar el rendimiento de los modelos encontrados. Los resultados obtenidos variaron por cada técnica, sin embargo, la que obtuvo una mayor precisión fue la de Bayes Ingenuo; con un poder de clasificación correcta de aproximadamente el 94%, valor que para efectos de la investigación puede ser considerado como satisfactorio para alcanzar el objetivo planteado.

Una vez evaluados los resultados, se propuso la validación del modelo encontrado con información de un periodo académico distinto a los involucrados en la fase de prueba. Este proceso se fundamentó en la verificación tanto con datos reales de navegación de usuario, como con datos alterados (lo que podría considerarse como la entrada de un intruso). Los resultados obtenidos muestran que la técnica de Bayes Ingenuo, es capaz de clasificar correctamente cada usuario en una alta proporción, razón por la cual se recomendó el uso de esta para tratar de determinar la autenticidad de un usuario a través de sus huellas de navegación en el Sitio.

Debido a que el trabajo realizado permitió conseguir una técnica que permite generar modelos para identificar patrones de comportamiento de usuarios dentro del Sitio de Control de Estudios de la UNET con un porcentaje de precisión cercano al 94%, se recomienda la implementación de todo el proceso (transformación de datos, verificación del patrón y generación de resultados), en una aplicación que realice el chequeo de la información que va generando el usuario al ir navegando por el Sitio, para intentar detectar a tiempo cualquier anomalía y tomar acciones que puedan prevenir cualquier acción no deseada.

Para finalizar se propone (para trabajos futuros) profundizar en el estudio de las técnicas seleccionadas en los siguientes aspectos:

- La revisión detallada del Alineamiento de Secuencias de Needleman-Wunsch, ya que por las características del problema, pudieran encontrarse mejores resultados utilizando una función de similitud dinámica que se ajuste al comportamiento histórico de cada usuario dentro del Sitio.
- Verificar los resultados obtenidos con las Cadenas de Markov utilizando otra técnica de comparación de las matrices obtenidas.
- Incluir un nuevo parámetro dentro del estudio que refleje la acción que está ejecutando el usuario dentro del Sitio.
- Proponer pruebas formales (como la curva característica de operación) para verificar el ajuste del modelo seleccionado.

De igual modo, incluir nuevas técnicas de minería de datos para el estudio del problema.

Bibliografía

[1] K. Meter and G. Cambridge, *Shaping the future: Business Design through Information Technology*. Harvard Business School Press, 1991.

[2] M. S. G. de Chile, "Gobierno electrónico", in *Modernización de los Servicios Públicos*, Chile.

[3] M. Powell, "Cómo proteger un servicio web xml de intrusos, segunda parte", *Microsoft Corporation*, Enero 2003.

[4] G. D. Koblinc, "*Web mining*", in Departamento de Ciencias de la Computación Universidad de Buenos Aires, Argentina.

[5] J. Ortega, "*Session analysis of cindoc's web: an approach to web usage mining",* España, 2005.

[6] M. S. D. N. en Español, "Procedimientos de seguridad básicos para aplicaciones web", *Microsoft Corporation*, 2006.

[7] M. J. Ranum, "Intrusion detection: challenges and myths", *Network Flight Recorder, Inc*, Marzo 2000.

[8] O. Etzioni, "*The world wide web: Quagmire or gold mine", Communications of the ACM*, 1996.

[9] T. F. Lunt, "*Detecting intruders in computer systems", in Sixth Annual Symposium and Technical Displays on Physical and Electronic Security*, 1990.

[10] U.Fayyad, G. Piatetsky, and P. Smyth, "*The kdd process for extracting useful knowledge from volumes of data", in Communications of the ACM*, 1996.

[11] J. Villena, J. C. González, E. Barceló, and J. R. Velasco, "Minería de uso de la web mediante huellas y sesiones", Madrid, España, 2002.

[12] C. Romero, S. Ventura, and C. Hervás, "Descubrimiento de reglas de predicción en sistemas de e-learning utilizando programación genética", in Universidad de Córdoba, Campus Universitario de Rabanales, Córdoba, España.

[13] D. Martinelli, H. Merlino, P. Britos, R. García-Martinez "Identificación de Hábitos de Uso de Sitios Web Utilizando SOM", in *Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires*, Buenos Aires, Argentina.

[14] Y. Guerrero, "Uso de técnicas inteligentes en Minería Web", in Universidad de los Andes, Mérida, 2005

[15] E. Herbert, "*Introduction to data mining and knowledge discovery", in Two Cows Corpration*, USA, 1999

[16] G. Canavos, “Probabilidad y Estadística – Aplicaciones y Métodos”, in México 1988.

[17] I. Miller, J. Freund and R. Johnson, “Probabilidad y Estadística para Ingenieros”, in Prentice Hall, 1997.

[18] R. Walpole, R. Myers, S. Myers and K. Ye, “Probabilidad y estadísticas para ingeniería y ciencias”, in México 2007.

[19] B. Kolman, D. Hill and V. Ibarra, “Algebra Lineal”, Pearson Educación, 2006.

[20] T. Devlin, “Bioquímica: Libro de texto con aplicaciones clínicas”, Reverté, 2004.

[21] R. Virgili, “Genoma humano: Nuevos avances en investigación, diagnóstico y tratamiento”, Edicions Universitat Barcelona, 2006.

[22] N. Gautham, “*Bioinformatics: database and algorithms*”, Alpha Science, 2006.

[23] M. Martí, G. Dalla Corte and J. Llisterri Boix. “Tratamiento del Lenguaje Natural”, Edicions Universitat Barcelona, 2002.

[24] J. Sobrino, N. Raissouni, and N. Kerr. “Teledetección”, Universitat de Valencia, 2000.

[25] P. Tan, M. Steinbach, and V. Kumar. “*Introduction to Data Mining”*, Addison Wesley, 2006.

[26] D. Maravall, “Reconocimiento de Formas y Visión Artificial”, Addison-Wesley Iberoamericana, 1994.

[27] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko and A. Wu, “*An Efficient K-Means Clustering Algorithm: Análysis and Implementation”*, IEEE *Transactions of Pattern Analysis and Machine Intelligence*, Vol 24, N° 7, 2002.

[28] Han, J., Kamber, M. y Tung, A.K.H. (2001). *Spatial clustering methods in data mining: A survey*. H. Miller and J. Han, editors, Taylor and Francis.

[29] H. Schildt and J. Holmes, “El arte de programar en Java”, México 2004.

[30] The University of Waikato: Weka, Nueva Zelanda. [web en línea]. Disponible desde Internet en: http://www.cs.waikato.ac.nz/ml/weka/

[31] Domingos, P. & Pazzani, M. (1996), Beyond independence: conditions for the optimality of the simple Bayesian classifier, in L. Saitta, ed., ‘Machine Learning: Proceedings of the Thirteenth International Conference’, Morgan Kaufmann, pp. 105-112. http://citeseer.ist.psu.edu/domingos96beyond.html