



UNIVERSIDAD  
NACIONAL  
DE LOJA



*Área de la Energía, las Industrias y los Recursos Naturales No Renovables*

---

CARRERA DE INGENIERÍA EN SISTEMAS

# "Desarrollo de un sistema inteligente para la clasificación de documentos ya digitalizados aplicando redes neuronales supervisadas"

*"Tesis previa a la Obtención del título de Ingeniero en Sistemas"*

**Autor:** Jiménez Ochoa Doris Yadira

**Director:** Ing. Paz Arias Henry Patricio, Mg. Sc

Loja-Ecuador

2015

# **CERTIFICACIÓN DEL DIRECTOR**

Ing. Henry Patricio Paz Arias

**DOCENTE DE LA CARRERA DE INGENIERÍA EN SISTEMAS**

## **CERTIFICA:**

Que la Egresada **Doris Yadira Jiménez Ochoa** autora del presente trabajo de titulación, cuyo tema versa sobre "**DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA CLASIFICACIÓN DE DOCUMENTOS YA DIGITALIZADOS APLICANDO REDES NEURONALES SUPERVISADAS**", ha sido dirigido, orientado y discutido bajo mi asesoramiento y reúne a satisfacción los requisitos exigidos en una investigación de este nivel por lo cual autorizo su presentación y sustentación .

Loja, Junio 2015

A handwritten signature in blue ink, appearing to be 'H. Paz Arias', written over a horizontal dotted line.

Ing. Henry Patricio Paz Arias

**DIRECTOR DEL TRABAJO DE TITULACIÓN**

## **AUTORÍA**

Yo **DORIS YADIRA JIMENÉZ OCHOA** declaro ser autora del presente trabajo de tesis y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos de posibles reclamos o acciones legales por el contenido de la misma.

Adicionalmente acepto y autorizo a la Universidad Nacional de Loja, la publicación de mi tesis en el Repositorio Institucional – Biblioteca Virtual.

**Firma:**



**Cédula:** 1104605710

**Fecha:** 16 de Junio 2015

**CARTA DE AUTORIZACIÓN DE TESIS POR PARTE DE LA AUTORA, PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO.**

Yo **DORIS YADIRA JIMENEZ OCHOA** ,declaro ser autor de la tesis titulada: **"DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA CLASIFICACIÓN DE DOCUMENTOS YA DIGITALIZADOS APLICANDO REDES NEURONALES SUPERVISADAS"**, como requisito para optar al grado de: **INGENIERÍA EN SISTEMAS**; autorizo al Sistema Bibliotecario de la Universidad Nacional de Loja para que con fines académicos, muestre al mundo la producción intelectual de la Universidad, a través de la visibilidad de su contenido de la siguiente manera en el Repositorio Digital Institucional:

Los usuarios pueden consultar el contenido de este trabajo en el RDI, en la redes de información del país y del exterior, con las cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia de la tesis que realice un tercero.

Para constancia de esta autorización, en la ciudad de Loja, **dieciséis** días del mes de **Junio** del dos mil quince.



**Firma:**

**Autor:** Doris Yadira Jiménez Ochoa

**Cédula:** 1104605710

**Dirección:** Catamayo (Isidro Ayora de 18 de Agosto)

**Correo Electrónico:** ya\_dy\_j20@ hotmail.com

**Teléfono:** 2677-026

**Celular:** 0994498936

**DATOS COMPLEMENTARIOS**

**Director de Tesis:** Ing. Henry Patricio Paz Arias, Mg. Sc

**Tribunal de Grado:** Ing. Luis Roberto Jácome Galarza, Mg. Sc

Ing. Waldemar Victorino Espinoza Tituana, Mg. Sc

Ing. Carlos Miguel Jaramillo Castro, Mg. Sc

## **AGRADECIMIENTO**

Mi agradecimiento a las Autoridades de la Universidad Nacional de Loja, de manera especial al Personal Docente de la Carrera de Ingeniería en Sistemas del Área de la Energía las Industrias y los Recursos Naturales No Renovables, por sus valiosos conocimientos durante el proceso académico.

Al Mg.SC Henry Paz, mi director de tesis cuya capacidad profesional y calidad humana me motivaron siempre a la culminación de mi tesis, con comentarios y apoyo desde el mismo momento de la elección del tema hasta su culminación. El me enseñó que es posible aprender algo más si se dispone de ilusión.

A mi familia y padres, los cuales me han animado a que termine mi carrera y que siga adelante con mis sueños, por el apoyo incondicional y por los valores que han inculcado en mí para crecer como persona y el haberme querido y apoyado siempre.

Por último, a todos aquellas personas que de forma desinteresada me han apoyado y me ayudado dedicándome tiempo y paciencia para que este trabajo fuese posibles y concluido con los mejores resultados y demostrando que la victoria no sabe igual si no se lucha, así como a todos los autores científicos de las áreas consultadas que han posibilitado el sustento para esta investigación.

Doris Yadira Jiménez Ochoa

## **DEDICATORIA**

A Dios nuestro Señor, que me permite vivir cada día y que ha sido mi motor principal durante este tiempo, y con su ayuda logré a culminar uno de mis sueños anhelados.

De manera especial a mis padres, que siempre ha estado ahí motivándome, a culminar mis estudios para que llegue a ser una profesional con éxito.

A mis hermanos, y familia y esposo quienes por medio de su comprensión pudieron entender mi dedicación a este proyecto apoyándome en todo momento.

A mis amigos y compañeros que me brindaron su apoyo, amistad y cariño desinteresado, y a todos aquellos que tengan interés con mi proyecto les sirva para futuras investigaciones.

Doris Yadira Jiménez Ochoa

## **CESIÓN DE DERECHOS**

**Doris Yadira Jiménez Ochoa**, declaro ser autora intelectual del presente trabajo de titulación, autorizo a la Universidad Nacional de Loja, al Área de Energía, las Industrias y los Recursos Naturales No Renovables y por ende a la Carrera de Ingeniería en Sistemas a hacer uso del mismo en lo que estime sea conveniente.

Para constancia firmamos a continuación.



Doris Yadira Jiménez Ochoa

CI: 1104605710

## Índice de Contenidos

<b>CERTIFICACIÓN DEL DIRECTOR</b> .....	II
<b>AUTORÍA</b> .....	III
<b>AGRADECIMIENTO</b> .....	V
<b>DEDICATORIA</b> .....	VI
<b>CESIÓN DE DERECHOS</b> .....	VII
<b>a. Título</b> .....	12
<b>b. Resumen</b> .....	13
<b>c. Introducción:</b> .....	15
<b>d. Revisión de literatura</b> .....	18
CAPÍTULO I: ESTUDIO SOBRE LA CLASIFICACIÓN DE INFORMACIÓN .....	18
CAPÍTULO II: APLICACIONES WEB.....	28
CAPÍTULO III: REDES NEURONALES SUPERVISADAS.....	49
<b>e. Materiales y Métodos</b> .....	72
<b>f. Resultados</b> .....	75
1. Análisis .....	75
2. Diseño .....	85
3. Desarrollo .....	96
4. Pruebas .....	100
5. Implementación .....	106
<b>g. Discusión</b> .....	110
1. Desarrollo de la propuesta alternativa .....	110
2. Valoración técnica económica ambiental .....	112
<b>h. Conclusiones</b> .....	115
<b>i. Recomendaciones</b> .....	116
<b>j. Bibliografía</b> .....	117
<b>K. Anexos</b> .....	122
Anexo 1. Artículo .....	122
Anexo 2: Sumarios de la clasificación de Dewey .....	134
Anexo 3: Vocabulario utilizado para cada categoría .....	137
Anexo 4: Declaración de Confidencialidad.....	144

Anexo 5: Certificado de traducción del Resumen .....	146
Anexo 6: Certificado conferido por la Bibliotecaria del funcionamiento de la Aplicación .....	147
Anexo 7: Licencia Creative Commons del Normativo .....	148

## Índice de Figuras

FIGURA 1.DIVISIÓN DE LAS FORMAS DE CLASIFICAR OBJETOS.....	19
FIGURA 2. CLASIFICACIÓN SUPERVISADA .....	21
FIGURA 3. REPRESENTACIÓN DE LOS METADATOS DE UNA PÁGINA WEB.....	22
FIGURA 4.EJEMPLO DE TIPOS DE METADATOS .....	25
FIGURA 5.ESQUEMA BÁSICO DE UNA APLICACIÓN WEB.....	29
FIGURA 6. ESQUEMA GENERAL DE UNA APLICACIÓN WEB .....	29
FIGURA 7. ARQUITECTURA: TODO EN UN SERVIDOR.....	31
FIGURA 8.ARQUITECTURA: SERVIDOR DE DATOS SEPARADOS.....	31
FIGURA 9.ARQUITECTURA: TODO EN UN SERVIDOR CON SERVICIO DE APLICACIONES.....	32
FIGURA 10.ARQUITECTURA: SERVIDOR DE DATOS CON SERVICIO DE APLICACIONES .....	32
FIGURA 11.ARQUITECTURA: TODO SEPARADO.....	33
FIGURA 12.FUNCIÓN DEL PROTOCOLO HTTP .....	33
FIGURA 13.LENGUAJES TANTO CLIENTE-SERVIDOR.....	35
FIGURA 14.ESQUEMA DEL FUNCIONAMIENTO DE HIBERNATE .....	43
FIGURA 15.ESQUEMA DEL FUNCIONAMIENTO MVC .....	44
FIGURA 16.FUNCIONAMIENTO DE UN BD EMPLEANDO LA ARQUITECTURA CLIENTE -SERVIDOR.....	45
FIGURA 17.EJEMPLO DE UNA RED TOTALMENTE CONECTADA .....	56
FIGURA 18.ARQUITECTURA DE UNA RED PERCEPTRÓN .....	58
FIGURA 19.PERCEPTRÓN SIMPLE .....	59
FIGURA 20. ARQUITECTURA DEL PERCEPTRÓN MULTICAPA .....	60
FIGURA 21.ESTRUCTURA DE UN PERCEPTRÓN MULTICAPA.....	61
FIGURA 22.ESTRUCTURA DE UN APRENDIZAJE SUPERVISADO .....	62
FIGURA 23.ESTRUCTURA DE UN APRENDIZAJE NO SUPERVISADO .....	64
FIGURA 24.RED BAYESIANA PARA EL RECONOCIMIENTO DE MULTIPALABRAS .....	66
FIGURA 25.PREPROCESAMIENTO DE LOS TEXTOS DE LA MUESTRA PARA APLICACIÓN DE LAS TÉCNICAS DE CLASIFICACIÓN. ....	68
FIGURA 26.PERCEPTRÓN MULTICAPA.....	69
FIGURA 27.COMPARACIÓN MEDIANTE TEST DE KRUSKAL WALLIS .....	70
FIGURA 28.ÁREAS DE APLICACIÓN MÁS RECIENTES (2001) .....	71
FIGURA 29.REPOSITORIO DE CIENCIA - SCIENCE LIBRARY .....	76
FIGURA 30.METADATOS DEL REPOSITORIO DE CIENCIA - SCIENCE LIBRARY .....	77
FIGURA 31.BIBLIOTECA VIRTUAL UNIVERSAL.....	78
<b>FIGURA 32.METADATOS DE LA BIBLIOTECA VIRTUAL UNIVERSAL.....</b>	<b>78</b>
FIGURA 33.BIBLIOTECA UTPL .....	79
FIGURA 34.METADATOS BIBLIOTECA UTPL.....	80
FIGURA 35.REDE DE REVISTAS CIENTÍFICAS DE AMÉRICA LATINA Y EL CARIBE, ESPAÑA Y PORTUGAL.....	81
FIGURA 36.METADATOS RED DE REVISTAS CIENTÍFICAS DE AMÉRICA LATINA Y EL CARIBE, ESPAÑA Y PORTUGAL .....	81
FIGURA 37.MODELO DE CLASES.....	85
FIGURA 38. MODELO DE LA BASE DE DATOS .....	86
FIGURA 39.VENTANA BUSCAR LIBROS .....	87
FIGURA 40.VENTANA CREAR LIBRO .....	88
FIGURA 41.VENTANA MOSTRAR LIBRO.....	88
FIGURA 42.GRUPOS PRINCIPALES DE SISTEMA DE CLASIFICACIÓN DEWEY .....	91
FIGURA 43.VOCABULARIO DE STEMMING .....	93
FIGURA 44. DIAGRAMA EJEMPLO DE FLUJO DE ENTRADA Y TRANSFORMACIÓN REALIZADA .....	94
FIGURA 45.RESULTADO DE LA RED EN RWEKA .....	96

FIGURA 46 DISEÑO DEL ALGORITMO PARA LECTURA DE METADATOS .....	97
FIGURA 47 MÉTODO DEL FUNCIONAMIENTO DE LA RED NEURONAL MLP .....	99
FIGURA 48 RESULTADO DE LA MUESTRA 1 .....	103
FIGURA 49. RESULTADOS DE LA MUESTRA 2 .....	104
FIGURA 50. RESULTADO ENTRE LA CLASIFICACIÓN DEL SISTEMA Y DE LA BIBLIOTECA .....	109

## Índice de Tablas

TABLA I LENGUAJES DE PROGRAMACIÓN .....	37
TABLA II REPOSITARIOS DE ESTUDIO .....	82
TABLA III METADATOS PARA LA BASE DE DATOS.....	84
TABLA IV PASOS PARA UNA CLASIFICACION MLP .....	101
TABLA V TALENTO HUMANO .....	112
TABLA VI RECURSOS MATERIALES.....	113
TABLA VII RECURSOS TÉCNICOS.....	113
TABLA VIII RECURSOS TECNOLÓGICOS .....	113
TABLA IX SERVICIOS.....	114
TABLA X COSTO TOTAL.....	114

## **a. Título**

**"DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA CLASIFICACIÓN DE DOCUMENTOS YA DIGITALIZADOS APLICANDO REDES NEURONALES SUPERVISADAS"**

## **b. Resumen**

En este trabajo se propone un modelo de Redes Neuronales Artificiales con aprendizaje supervisado: Perceptrón Multicapa utilizando como criterio de clasificación el área disciplinar, de esta manera se propone el desarrollo de una nueva aplicación web que utilizando técnicas de sistemas inteligentes implemente un sistema de clasificación de documentos de tipo supervisado para agrupar los documentos por categorías según corresponda.

El criterio para realizar la clasificación de documentos está basado por 14 categorías de esta manera el sistema formará grupos de documentos afines a la categoría correspondiente, además se puede contrastar sus resultados con el ingreso de 100 documentos de la biblioteca del Área de la Energía las Industrias y los Recursos no Renovables de la carrera de Ingeniería en Sistemas.

EL dominio de esta aplicación es utilizar un corpus de 5000 documentos del repositorio de la Universidad Autónoma de la ciudad de México (Red de Revistas Científicas de América Latina y el Caribe, España y Portugal) en distintos formatos (Microsoft Word, PDF, texto plano). En el sistema se han implementado analizadores que extraen metadatos (título, autor, palabras claves, descripción, fecha entre otros) así mismo se ha desarrollado una variedad de algoritmos de aprendizaje supervisados aplicados a la categorización de documentos, del mismo modo se desarrolló un lematizador lo cual ayuda al recall (reduce la palabra a su raíz) a los documentos de consulta y eliminar los artículos.

En particular, en los experimentos de nuestro dominio el lenguaje R con su paquete RWEKA ha demostrado tener buenas propiedades para experimentar y demostrar resultados de la asertividad y error de la clasificación de los documentos.

El objetivo final es construir una aplicación web para categorizar documentos de forma automática de acuerdo a las categorías que se planteó.

## Summary

A model of artificial neural networks with supervised learning is proposed in this research work, Multilayer Perceptron using criterion for classifying the area as discipline, thus proposes the development of a new web application that using techniques of intelligent systems implement a type supervised document classification system to group documents into categories as appropriate. The criteria for classifying documents is based for 14 categories in this way system formed groups of related documents to the corresponding category, also their results can be contrasted with the income of 100 documents from the library of the energy Area industries and non-renewable resources of the engineering systems.

The domain of this application is to use a body of 5000 documents in the repository of the University of the city of Mexico(Network of scientific journals of Latin America and the Caribbean, Spain and Portugal) in different formats(Microsoft Word, PDF, text plate).The system have implemented scanners that extract metadata (title, author, keywords, description, date among others), has also developed a variety of supervised learning algorithms applied to the categorization of documents, in the same way a stemmer was developed which helps the recall (reduces the word to its root), to consultation documents and delete articles.

In particular, in the experiments of our domain language R with your package RWEKA has shown good properties to experiment and demonstrate results of the assertion and error of classification of the documents.

Our ultimate goal is to build a web application to categorize documents automatically, this system allows any user to upload books and automatically classifies them according to the existing categories. In particular, in the experiments of our domain language R with your package RWEKA has shown good properties to experiment and demonstrate results of the assertion and error of classification of the documents.

## **c. Introducción:**

El aumento exponencial de la información disponible en formato digital durante los últimos años y las expectativas de crecimiento futuro hacen necesaria la organización de la información con el fin de mejorar la búsqueda y acceso a la información. Con este fin adquiere importancia la investigación de la clasificación automática de textos.

La clasificación automática de textos está basada en sus inicios en técnicas de ingeniería del conocimiento, donde un experto definía de forma manual las reglas que cada documento debía cumplir para pertenecer a una u otra categoría. No obstante el gran costo que suponía esto junto con los avances que se habían realizado en el área de la inteligencia artificial, dieron lugar en la década de los 80 a la utilización de técnicas de aprendizaje automático [1].

La categorización de documentos de texto es una aplicación de la minería de texto que asigna a los documentos a una categoría, etiquetas o clases, basadas en el contenido [2]. Es un componente importante de muchas tareas de organización y gestión de la información. El enfoque tradicional para la categorización de textos en que los expertos definían manualmente las reglas de clasificación ha sido reemplazado por otro basado en técnicas de aprendizaje automático o en combinaciones con otras técnicas. Se define la categorización (o clasificación) de textos como la actividad de etiquetar textos en lenguaje natural con categorías temáticas de un conjunto predefinido.

Cuando se utiliza aprendizaje automático el objetivo es aprender a clasificar a partir de ejemplos que permitan hacer la asignación a la categoría automáticamente. Durante el aprendizaje o entrenamiento del sistema se evalúan las condiciones de pertenencia a cada una de las categorías. Para realizar el entrenamiento es necesario disponer de conocimiento previo de expertos en forma de decisiones de categorización asignadas a cada uno de los documentos. Este conocimiento corresponde a un conjunto de documentos preclasificados de modo que el sistema pueda leer la categoría o grupo de pertenencia de cada uno de los documentos.

El desarrollo de la ciencia y tecnología viene avanzando aceleradamente, la información en cada área de conocimiento se incrementa en forma exponencial y su tratamiento así como su almacenamiento se hace más compleja. El explosivo crecimiento de la información disponible en documentos digitales en el área de informática y sistemas, ha

hecho necesario desarrollar nuevos instrumentos y herramientas que faciliten la realización de procesos de búsquedas de forma eficiente y efectiva, así como la administración de estos recursos. Es frecuente que para facilitar la búsqueda de información se proceda a la categorización de los documentos en un conjunto acotado de clases. Estas clases permiten representar áreas específicas del conocimiento y son generalmente consolidadas por expertos [3].

En nuestra vida cotidiana categorizar es fundamental para la comprensión de ideas para saber qué hacer, a quién enviar la información, dónde guardarla. El etiquetado o categorización de hasta miles de documentos son realizados por personas especialista en el área de interés de los documentos, y por tanto es un proceso muy costoso. Esto ha dado pie a nuevos métodos o herramientas automáticas para la gestión o categorización de documentos.

El contexto del presente trabajo tiene como finalidad crear una aplicación web que permita categorizar automáticamente documentos utilizando sistemas expertos de tal manera que realice un proceso más rápido en un menor tiempo y costo.

En lograr este propósito se ha seguido las siguientes fases: análisis, diseño, desarrollo, integración y pruebas [4].

**Análisis:** El primer paso consistió en la revisión y colección de documentos relacionados al tema en cuestión para poder conocer y entender el estado del arte en los sistemas de redes neuronales supervisadas desarrollados hasta el momento, basado en ello se consiguió el entendimiento de las técnicas utilizadas en los sistemas expertos y el mejoramiento a la hora de la implementación.

**Diseño:** Se realizó el diseño el diagrama de clases para identificar los principales clases del sistema y la relación entre cada una de ellas, además describe la estructura del sistema lo que hace y sus métodos, se elaboró el diseño de prototipos de pantallas para ver de forma básica las interfaces que pueden tener la aplicación y por último el diseño de la arquitectura de la red multicapa.

**Desarrollo:** El objetivo final es obtener una implementación práctica de un sistema inteligente para la clasificación de documentos. En un primer paso se llevó a cabo el

desarrollo del algoritmo de la lectura de los metadatos de los documentos, luego se construye un lematizador para la eliminación de artículos y el desarrollo de la red neuronal multicapa que realiza la clasificación de documentos en base a un corpus de datos.

**Integración:** Se integró los algoritmos de: lectura de metadatos y el de la red de Perceptrón multicapa lo cual esta integración nos permite a obtener nuestro objetivo final que es de desarrollo de un sistema inteligente aplicando redes neuronales para la clasificación de documentos.

**Pruebas:** En cuanto a la valoración de los resultados realizaremos pruebas con 1000 documentos para entrenar y constatar el margen de asertividad y error de la clasificación que presenta el sistema, posteriormente con 3000 libros para mejorar y reducir el margen de error y finalmente en la implementación con pruebas de 100 libros de la Universidad Nacional de Loja lo cual de esta manera podremos medir su nivel de eficiencia.

Aunque la clasificación de documentos sea un tema arduo debido a varios parámetros implicados (lematización, eliminación de artículos, metadatos.), llega a ser de interés cada vez mayor en diversos campos de aplicaciones.

Por último, en este trabajo de titulación (TT), hay que indicar que utiliza una estructura secuencial. Es decir las diferentes fases del proyecto se irán explicando según se hayan ido desarrollando.

Cada una de las fases principales consta de una sección de resultados, que nos permitirá valorar el éxito alcanzado en cada una de ellas.

Al final se incluirán unas conclusiones finales para poder hacer balance de todo el proyecto y definir si los objetivos se han cumplido.

## **d. Revisión de literatura**

### **CAPÍTULO I: ESTUDIO SOBRE LA CLASIFICACIÓN DE INFORMACIÓN**

Los avances tecnológicos de los últimos años han provocado un aumento exponencial de la cantidad de información producida a gestionar. El proceso de digitalización y la transformación de documentos que se está llevando a cabo son dos claros ejemplos de la revolución de la información [5].

La Categorización de Documentos (Document Clustering) puede definirse como la tarea de separar documentos en grupos. El criterio de agrupamiento se basa en las similitudes existentes entre ellos [6]. El concepto de clasificación de documentos se refiere al problema de encontrar para cada documento la clase a la que pertenece.

Un criterio de agrupamiento utilizado es dividir los documentos en una jerarquía de temas. Un ejemplo de esto último son las categorías que presenta el buscador Yahoo®. Un documento en particular se podría encontrar por ejemplo dentro de “Tecnología → Informática → Internet → Buscadores”. El problema que presenta esta técnica es la dificultad de encontrar la categoría que mejor describa a un documento. Los documentos no tratan un sólo tema y aunque lo hicieran el enfoque con el que el tema es tratado puede hacer que el documento encuadre en otra categoría. Esto hace que la categorización de documentos sea una tarea compleja y subjetiva, ya que dos personas podrían asignar el mismo documento a categorías diferentes, cada una aplicando un criterio válido.

#### **1.1. La necesidad de la agrupación automática de documentos**

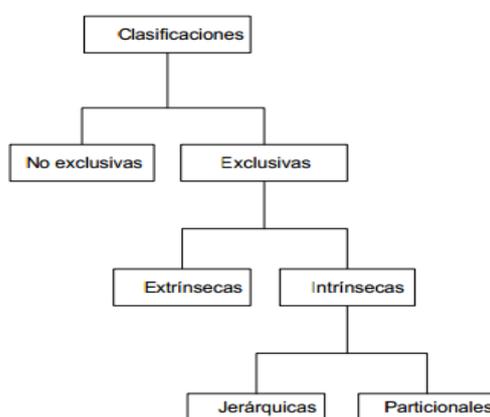
La clasificación automática de documentos ha sido ampliamente estudiada por diversos investigadores. Su utilidad se basa en la posibilidad de poder efectuar posteriormente una adecuada recuperación, asumiendo que aquellos textos que tratan de la misma materia están clasificados juntos o en apartados cercanos. Diversas técnicas han sido propuestas desde hace ya bastantes años. Así Fairthorne y Hayes sugirieron separadamente la posibilidad de utilizar sistemas de clasificación como un modo de aumentar la eficacia en la recuperación de información. Aunque el propio Salton cree interesante la agrupación de documentos, estima que resta efectividad a la recuperación.

Se puede definir la Clasificación Automática de Textos, también denominada Categorización de Textos o Topic Spotting, como la tarea de asignar automáticamente un conjunto de documentos a una o más categorías preexistentes a través de un conjunto predefinido de documentos pre- categorizados sobre los que el sistema lleva a cabo un proceso de aprendizaje supervisado. En el contexto de recuperación de información Van Rijsbergen formula la denominada “Hipótesis del Agrupamiento” [7]. Básicamente la hipótesis del agrupamiento sostiene que “Los avances tecnológicos de los últimos años han provocado un aumento exponencial de la cantidad de información producida a gestionar. El proceso de digitalización y la transformación de documentos que se está llevando a cabo son dos claros ejemplos de la revolución de la información.

La tarea de encontrar grupos de documentos con características comunes no sólo es compleja sino que además consume tiempo. Un bibliotecario que tratara de clasificar manualmente un documento tendría que leerlo para comprender su significado y luego asignarle una categoría utilizando su experiencia y sentido común. El costo y el tiempo asociados a este proceso han incentivado la investigación de técnicas que permitan automatizar la tarea. Debido al incremento en los volúmenes de información disponibles en forma electrónica y a la necesidad cada vez mayor de encontrar la información buscada en un tiempo mínimo éstas técnicas han estado recibiendo creciente atención.

## 1.2. Métodos para categorizar documentos

Las formas de clasificación de objetos tales como asignar clases predeterminadas a cada elemento o agruparlos en forma significativa, son susceptibles de dividirse según el esquema de la Figura 1.



**Figura 1. División de las formas de clasificar objetos**

- **No exclusivas:** Un mismo objeto puede pertenecer a varias categorías, clases o grupos.
- **Exclusivas:** Cada objeto pertenece solamente a una categoría, clase o grupo.
- **Extrínsecas (supervisadas):** Las clases a las que pertenecen los objetos están predefinidas, y se conocen ejemplos de cada una o algunos de los objetos ya están clasificados y son utilizados por el algoritmo para aprender a clasificar a los demás.
- **Intrínsecas (no supervisadas):** La clasificación se realiza en base a las características propias de los objetos, sin conocimiento previo sobre las clases a las que pertenecen.
- **Jerárquicas:** Los métodos jerárquicos consiguen la categorización final mediante la separación (métodos divisivos) o la unión (métodos aglomerativos) de grupos de documentos. Así estos métodos generan una estructura en forma de árbol en la que cada nivel representa una posible categorización de los documentos.
- **Particionales (no jerárquicas):** Los métodos no jerárquicos también llamados particionales o de optimización llegan a una única categorización que optimiza un criterio predefinido o función objetivo, sin producir una serie de grupos anidados.

### 1.3. Clasificación supervisada o categorización

En esta modalidad se parte de una serie de clases o categorías prediseñadas a priori en las cuales hay que colocar a cada uno de los documentos conocida como categorización, esta facilita la recuperación limitando las búsquedas a las clases o categorías que el usuario elige. Figura 2.



**Figura 2. Clasificación supervisada**

Aunque hay gran cantidad de algoritmos capaces de hacer clasificación supervisada, la idea o enfoque básico es muy parecido conseguir de alguna manera construir un patrón representativo de cada una de las clases o categorías y aplicar alguna función que permita estimar el parecido o similitud entre el documento a clasificar y cada uno de los patrones de las categorías. El patrón más parecido al documento es el que nos indica a qué clase debemos asignar ese documento como se muestra en la Figura 2.

## **1.4. Metadatos**

### **1.4.1. Que son los metadatos**

La literatura especializada define los Metadata o Meta Tags, como "Datos acerca de los datos" o "información acerca de la información". A través de esta metodología es posible describir el contenido de un recurso de aprendizaje. Otra definición es: información sobre objetos web comprensible por máquinas.

El término metadatos no tiene una definición única. Según la definición más difundida de metadatos es que son «datos sobre datos». También hay muchas declaraciones como «informaciones sobre datos», «datos sobre informaciones» e «informaciones sobre informaciones». [8]

Los metadatos se definen como datos que describen el contexto, contenido y la estructura de los documentos, así como su gestión en el tiempo [9].

Tradicionalmente los metadatos son utilizados en ambientes de bibliotecas y sistemas documentales. Un caso común es asociar autor, título y fecha a una publicación particular con el objetivo de organizarlo bajo una estructura definida. Esto sirve para minimizar esfuerzos de organización y facilitar su mantenimiento.

Existen normas que definen diferentes estructuras de metadatos, las cuales persiguen objetivos diferentes, por ejemplo:

- Dublin Core (DC). Datos sobre páginas del espacio web.
- Consortium for the Interchange of Museum Information (CIMI) Informaciones sobre museos.
- Machine Readable Card (MARC) Catalogación bibliográfica.
- Federal Data Geographic Committee (FGDC) Descripción de datos geo espaciales.

Estas normas definen lenguajes que especifican la sintaxis para generar estructuras y proveen especificaciones semánticas necesarias que explican el significado de las expresiones sintácticas.

Actualmente a los efectos de ayudar a solucionar el problema de sobrecarga de información derivada de la constante expansión del espacio web se promueve el uso de la denominada web semántica. La cual tiene entre sus objetivos modificar la forma en que se presenta la información en el espacio web en pos de facilitar el procesamiento automático de la misma, y de esta forma establecer facilidades para lograr un factible procesamiento, integración y reutilización de la información contenida en tal espacio. Los metadatos juegan un rol importante en el área mencionada, debido a que proveen una categorización semántica de su contenido, permitiendo razonar de forma automática sobre la información.

En el siguiente figura 3 muestra un ejemplo, se presentan los metadatos asociados a una página web.

```
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta name="Description" content="Revista Novatica de la Asociacion de
Tecnicos de Informática, núm. 157, may.-jun. 2002, Recuperación de Información
y la Web / Novatica, ATI's Journal and Magazine, #157, May-Jun. 2002 issue.
Monograph: Information Retrieval and the Web">
<meta name="GENERATOR" content="Mozilla/4.72 [en] (Win95; I) [Netscape]">
<meta name="Author" content="Rafael Fernandez Calvo">
<meta name="KeyWords" content="ATI, España, Spain, Informática, Computer,
Association, Publicaciones, Revista, Journal, Magazine, Novatica, Novatica, num.
157, mayo-junio, 2002 Recuperación de Información y la Web , #157, May-Jun.
2002 issue, Monograph, Information Retrieval, Web">
```

**Figura 3. Representación de los metadatos de una página web**

Otra clase de definiciones trata de precisar el término como «descripciones estructuradas y opcionales que están disponibles de forma pública para ayudar a localizar objetos» o «datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas».

#### 1.4.2. Estructura de metadatos

Cada estándar de metadatos define la estructura de éstos, lo que implica que no haya una estructura única de metadatos [10]. No obstante los metadatos están estructurados por un mínimo de elementos tales como: título, autor, fecha de creación, etc.

La estructura de metadatos consta principalmente de los siguientes elementos.

- **Accesibilidad:** indica la disponibilidad y usabilidad del recurso a grupos específicos.
- **Destinatario:** la persona (o personas) para quien ha sido destinado el recurso.
- **Agregación:** el nivel de posición de jerarquía del recurso.
- **Audiencia:** categoría de usuario al cual se provee el recurso.
- **Contribuidor:** entidad responsable de hacer contribuciones al contenido del recurso.
- **Cobertura:** la extensión o alcance del contenido del recurso.
- **Creador:** entidad responsable de la creación del contenido del recurso.
- **Fecha:** fecha asociada con un evento en el ciclo de vida del recurso.
- **Descripción:** descripción textual del contenido de un recurso.
- **Disposición:** instrucciones de retención y disposición para el recurso.
- **Formato:** la manifestación física o digital del recurso.
- **Identificador:** referencia no ambigua al recurso dentro de un contexto determinado.
- **Lenguaje:** lenguaje del contenido intelectual del recurso.
- **Locación:** ubicación física del recurso.
- **Mandato:** mandato legal o de otro tipo bajo el cual el recurso ha sido producido.

- **Preservación:** información para apoyar la preservación del recurso a largo plazo.
- **Relación:** referencias a recursos relacionados.
- **Derechos:** información acerca de los derechos sobre la producción del recurso.
- **Fuente:** referencia al recurso sobre el cual el recurso presente deriva.
- **Estado:** posición o estado del recurso.
- **Asunto:** tópico del contenido del recurso.
- **Título:** nombre asignado al recurso.
- **Tipo:** naturaleza o género del contenido del recurso.

### 1.4.3. Tipos de metadatos

Existen diversos tipos de metadatos cada uno con su propio formato para describirlos. Según la información que proporcionen existen metadatos sobre: el contenido, aspectos formales, derechos de autor y el contexto.

Según la función que proporcionan se pueden clasificar en [11]:

**Descriptivos.** Describen e identifican recursos de información. Permite a los usuarios la búsqueda y recuperación de la información. Contienen atributos físicos (medios, condición de las dimensiones) y atributos bibliográficos (título, autor, creador, idioma, palabras claves).

**Estructurales.** Facilitan la navegación y la presentación de los recursos. Proporcionan información sobre la estructura interna de los documentos, así como la relación entre ellos como página, sección, capítulo, partes, índice, tablas de contenido etc.

**Administrativos.** Facilitan la gestión de conjuntos de recursos. Incluye la gestión de derechos y sobre control de acceso y uso. Incluyen datos como tipo y modelo de escáner, resolución, profundidad, espacio de color, formato de archivo etc.

En la Figura 4 se muestra un ejemplo de extracción de metadatos de una obra.

No. de OD: 779	
<b>Título:</b> Huellas : búsquedas en artes y diseño / N° 5 (2006) <b>Editor:</b> Universidad Nacional de Cuyo. Facultad de Artes y Diseño	
<b>Contenido</b> Hacia una nueva semiótica, Guebbe, Gabriela, p. 9-16 Obras para piano y orquesta en la Argentina, De Marinis, Dora, p. 17-40 Cuicani, Plana, Beatriz, p. 41-52 Renovación de la música en Argentina, Loyola, María Enriqueta, p. 53-66 Forma móvil: Intermission 6, de Morton Feldman, Solare, Juan María, p. 67-74 El teatro mendocino en el siglo XIX, Navarrete, José Francisco, p. 75-84 Homenaje al Prof. Carlos Agustín Gómez, Coll, Roxana, p. 85-92	<b>Formato PDF</b> 83 Kb 124 Kb 579 Kb 3,4 Mb 501 Kb 92,9 Kb 1,3 Mb

◆ Metadatos administrativos ◆ Metadatos estructurales ◆ Metadatos descriptivos

**Figura 4. Ejemplo de tipos de metadatos**

### 1.4.5. Clasificación de metadatos

Los metadatos se clasifican usando tres criterios:

**Contenido.** Subdividir metadatos por su contenido es lo más común. Se puede separar los metadatos que describen el recurso mismo de los que describen el contenido del recurso. Es posible subdividir estos dos grupos más veces por ejemplo para separar los metadatos que describen el sentido del contenido de los que describen la estructura del contenido o los que describen el recurso mismo de los que describen el ciclo vital del recurso.

**Variabilidad.** Según la variabilidad se puede distinguir metadatos mutables e inmutables. Los inmutables no cambian no importa qué parte del recurso se vea por ejemplo el nombre de un fichero. Los mutables difieren de parte a parte por ejemplo el contenido de un vídeo.

**Función.** Los datos pueden ser parte de una de las tres capas de funciones: subsimbólicos, simbólicos o lógicos. Los datos subsimbólicos no contienen información sobre su significado. Los simbólicos describen datos subsimbólicos, es decir añaden sentido. Los datos lógicos describen cómo los datos simbólicos pueden ser usados para deducir conclusiones lógicas, es decir añaden comprensión.

## 1.5. Bibliotecas Digitales

Los repositorios están constituidos por información nacida en soportes físicos (textos, imágenes fijas, videos, sonido) que ha sido digitalizada y también por documentos nacidos en formato electrónico. Todos ellos se denominan "objetos digitales" y amplían

enormemente las posibilidades de recuperación y consulta de un documento, además de preservar los originales de su manipulación.

Biblioteca Digital es una colección estructurada de documentos digitales, desarrollada según una política y un esquema conceptual que ofrece a sus usuarios servicios de valor añadido, fundamentados precisamente en la colección o en aspectos relacionados con la misma[11].

Un repositorio digital es un sitio web centralizado donde se almacena y mantiene información digital, habitualmente una imagen, documentos Word, Excel, documentos digitalizados, libros electrónicos, una página HTML etc. [9].

Un repositorio corresponde a un sitio (lógico y físico) centralizado donde la información es almacenada y administrada. Un repositorio puede ser un lugar donde múltiples bases de datos o ficheros se ubican para ser distribuidos sobre una red de computadoras.

#### **1.5.1. Los metadatos en el entorno digital**

En el entorno digital la gestión de los documentos electrónicos es una tarea técnicamente especializada. La gestión de la autenticidad, fiabilidad, integridad y disponibilidad de los documentos digitales al largo plazo es realmente más compleja que la de los documentos analógicos. En un entorno digital las características de los documentos deben estar más explícitamente documentadas que en un entorno analógico.

La gestión de los documentos siempre ha implicado la gestión de los metadatos, sin embargo el entorno digital precisa una expresión diferente de los requisitos tradicionales, y de unos mecanismos diferentes en la identificación, captura, asignación y uso de lo metadatos. En el entorno digital los documentos autorizados son aquellos que se acompañan de metadatos que definen sus características críticas.

Estas características deberían estar explícitamente documentadas, porque no están implícitas como en algunos procesos basados en papel, en el entorno digital es esencial asegurar que la creación y la incorporación de los metadatos de gestión de documentos estén implantadas en los sistemas que crean y gestionan los documentos.

En un entorno digital, los contornos físicos de los objetos digitales son difusos y sus documentos y componentes pueden estar almacenados separadamente. Los vínculos

entre los metadatos, los documentos, las partes que lo componen, sus estructuras, etc., se pueden romper con bastante facilidad. En los documentos en papel los contornos de los documentos son físicos y el contexto se puede deducir de su medio de gestión y almacenamiento. Sin embargo en un entorno digital los documentos son virtuales, pueden estar formados por otros documentos que se encuentran almacenados en lugares diferentes, además sus entornos y estructuras pueden ser difusos. Por estas razones un entorno digital, son esenciales la existencia de sistemas robustos que gestionan metadatos [9].

### **1.5.2. Objetivos de los metadatos para la gestión de documentos**

Las organizaciones necesitan sistemas de información que capturen y gestionen información contextual que ayude al entendimiento, uso, acceso y gestión de sus documentos a lo largo del tiempo. Esta información es crítica para firmar la autenticidad, fiabilidad, integridad, disponibilidad y valor probatorio de los documentos [12]. En su conjunto, esta información se conoce como metadatos para la gestión de documentos

Los metadatos para la gestión de documentos pueden usarse en una organización para distintos fines relacionados con la identificación, autenticación, descripción, localización y gestión de sus recursos de forma sistemática y consistente para cumplir con los requisitos propios de la organización y de la sociedad y hacer frente a sus responsabilidades [9].

Las aplicaciones informáticas de gestión de documentos y los sistemas de gestión que incluyen la funcionalidad de gestión de documentos gestionan los documentos mediante la incorporación y gestión de metadatos sobre ellos y el contexto de su creación y uso.

### **1.5.3. Beneficios de los usos de metadatos en gestión de documentos**

Según ISO 23081[12] los metadatos sirven para:

- Proteger, mantener y asegurar el valor de los documentos como prueba.
- Asegurara la accesibilidad y uso de los documentos a largo plazo.
- Favorece la comprensibilidad de los documentos.
- Asegurar la autenticidad, fiabilidad e integridad de los documentos.
- Contribuir a gestionar los derechos de autor.
- Proteger la confidencialidad de los documentos.

- Favorecer la recuperación, sostenibilidad e interoperabilidad de los documentos a través de los sistemas que los gestiona.
- Proporcionar vínculos entre los documentos y el contexto de su creación y uso.
- Mantener su estructura y legibilidad de una forma fidedigna.
- Soportar una migración eficiente y completa.

## **CAPÍTULO II: APLICACIONES WEB**

Internet y la Web han influido enormemente tanto en el mundo de la información como en la sociedad en general. Si nos centramos en la Web en pocos menos de 10 años ha transformado los sistemas informáticos, ha roto las barreras físicas (debido a la distancia), económicamente y lógicas (debido al empleo de distintos sistemas operativos, protocolos, etc.), y ha abierto todo un abanico de nuevas posibilidades. Una de las áreas que más expansión está teniendo en la Web en los últimos años son las aplicaciones Web.

Las aplicaciones Web permiten la generación automática de contenido, la creación de páginas personalizadas según el perfil del usuario o el desarrollo del comercio electrónico. Además una aplicación web permite interactuar con los sistemas informáticos de gestión de una empresa, como puede ser gestión de clientes, contabilidad o inventario a través de una página web.

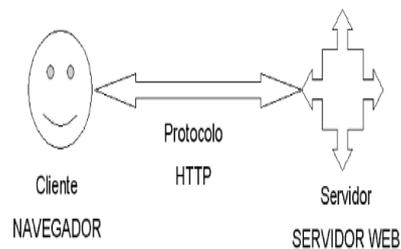
Las aplicaciones web se encuadran dentro de las arquitecturas cliente –servidor: un ordenador solicita servicios (el cliente) y el otro está a la espera de recibir solicitudes y las responde (el servidor).

### **2.1. Que es una aplicación Web**

Una aplicación Web es un tipo especial de aplicación cliente—servidor, en la cual el cliente o usuario empleando un navegador Web cualquiera accede a la aplicación mediante la dirección en la que está ubicado el respectivo servidor Web. El acceso a este servidor se realiza ya sea a través de internet u intranet [13].

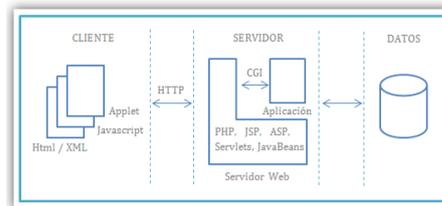
La comunicación entre el cliente y servidor se da mediante el protocolo HTTP que forma parte de la amplia de protocolos de comunicación TCP/IP que son empleados en el

internet como se muestra en la Figura 5. Estos protocolos permiten la conexión de sistemas heterogéneos, lo que facilita el intercambio de información en la World Wide Web y es la manera en la que se transfiere las páginas Web entre servidores y clientes [14].



**Figura 5. Esquema básico de una aplicación web**

En la Figura 6 se puede observar los bloques que definen a los elementos cliente-servidor y su comunicación, la cual puede ser empleando el protocolo HTTP o HTTPS dependiendo del nivel de seguridad que se requiera el sistema. Además se muestran cada tipo de tecnología involucrada en la generación e interacción de documentos Web.



**Figura 6. Esquema general de una aplicación web**

Las aplicaciones web son utilizadas generalmente para implementar web mail, subastas y ventas en línea, wikis, foros de discusión, redes sociales, juegos etc.

### 2.1.1. Características de una aplicación Web

- La Portabilidad de la aplicación es dinámica, pudiendo de esta manera ejecutarse en cualquier plataforma, hablamos de dispositivos móviles, computadoras que alojen cualquier sistema operativo e inclusive consolas de videojuegos.

- No se necesita instalar la aplicación en el lado del cliente, este accede simplemente a través del navegador Web de su preferencia.
- Las actualizaciones que se requieran en la aplicación se verán reflejadas directamente en la interfaz de usuario, basta con actualizarlas en el servidor.
- El cliente y el servidor pueden re presentarse como una sola entidad y también como entidades separadas, realizando actividades o tareas independientes.
- Al implementar una aplicación web, no se requieren de sofisticados equipos, lo que implica una reducción de costos a nivel de infraestructura.
- No ocupan espacio en el disco duro del Usuario.
- Los recursos del equipo del cliente (usuario), no son consumidos, es el servidor el que asume todos los procesos.

Las Aplicaciones web se emplean en tres entornos informáticos muy similares que suelen confundirse entre sí internet, intranet y extranet [15] [16].

**INTERNET:** desde 1988 la internet ha ido creciendo rápidamente ahora más de 1000 países están conectados a este nuevo medio para intercambiar todo tipo de información, la internet posee un diseño descentralizado, cada ordenador en la internet es independiente.

Existen una gran variedad de formas de acceder a la internet, este método más común es obtener acceso a través de proveedores de servicios de internet, cuando nos referimos a internet es un conjunto de dos o más redes de ordenadores interconectados entre sí.

**INTRANET.** Una intranet es una red de ordenadores basado en los protocolos que gobiernan internet (IP) que pertenece a una organización y que es accesible únicamente por los miembros de la organización, empleados o personas con autorización.

**EXTRANET:** es una intranet a la que pueden acceder particularmente personas autorizadas ajenas a la organización o empresa propietaria de la intranet

## 2.2. Arquitectura de las Aplicaciones Web

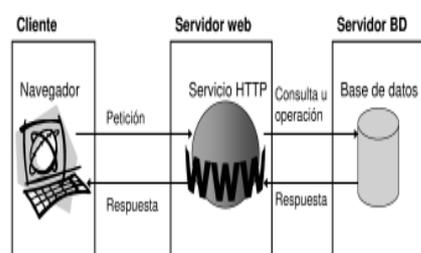
Existen diversas variantes de la arquitectura básica según como se implementen las diferentes funcionalidades de la parte del servidor [16] Las arquitecturas más comunes son:

1. **Todo en un servidor:** en un único ordenador aloja el servicio de HTTP, la lógica de negocio y la lógica de datos y los datos. El software que ofrece el servicio de HTTP gestiona también la de negocio. Las tecnologías que emplean esta arquitectura ASP Y PHP. Figura 7



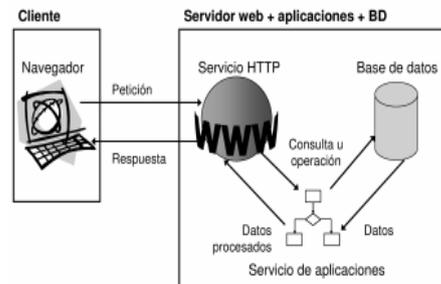
**Figura 7. Arquitectura: todo en un servidor**

2. **Servidor de datos separados:** a partir de la arquitectura anterior, se separa la lógica de datos y los datos a un servidor de datos específico. Las tecnologías que emplean esta arquitectura son ASP Y PHP. Figura 8



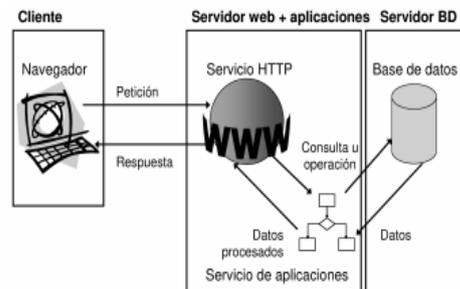
**Figura 8.Arquitectura: Servidor de datos separados**

3. **Todo en un servidor con servicio de aplicaciones:** es la arquitectura número 1 se separa lógica de negocio del servidor HTTP y se incluye el servicio de aplicaciones para gestionar los procesos que implementa la lógica de negocio. La tecnología que emplea esta arquitectura es JSP. Figura 9



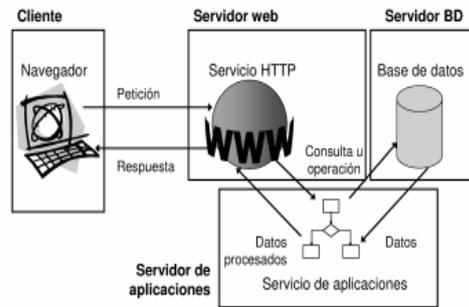
**Figura 9.Arquitectura: Todo en un servidor con servicio de aplicaciones**

4. **Servidor de datos separados con servicio de aplicaciones:** a partir de la arquitectura anterior, se separa la lógica de datos y los datos a un servidor de bases de datos específicos. La tecnología que se emplea es JSP. Figura 10



**Figura 10.Arquitectura: Servidor de datos con servicio de aplicaciones**

5. **Todo separado:** las tres funcionalidades básicas del servidor web se separan en tres servidores específicos: la tecnología que se emplea es JSP .El objetivo de separas las distintas funcionalidades (servicio http, lógica de negocio y lógica de datos) en distintos servidores es aumentar la estabilidad del sistema a un mayor rendimientos. Figura 11



**Figura 11.Arquitectura: Todo separado**

### 2.3. Protocolo HTTP

El protocolo denominado Hypertext Transfer Protocol, es el método más usual con el cual se intercambia información en internet, transfiriendo las paginas o servicios Web que provienen de un servidor y se transfieren hacia un cliente [14].Este protocolo trabaja a nivel de aplicación para sistemas de información multimedia lo que hace es trasladar ficheros de tipo HTML(Lenguaje de mercado para elaborar páginas web) entre dispositivos. HTML es un lenguaje que trabaja en lado del cliente caracterizado por emplear etiquetas para identificar a los diferentes elementos que lo conforman.

En la siguiente Figura 12 se muestra el papel que desempeña este protocolo.



**Figura 12.Función del protocolo HTTP**

Entre las propiedades del protocolo HTTP se pueden destacar las siguientes:

- Su esquema de direccionamiento es comprensible, utiliza el URL para localizar los sitios Web sobre los que hay que aplicar algún método.

- Implementa la arquitectura cliente-servidor, HTTP se basa en el paradigma solicitud. Respuesta cuya comunicación se asienta sobre los protocolos TCP-IP. Por defecto el número de puerto empleado por HTTP es el 80.
- Es un protocolo que trabaja sin conexión y sin estado, luego de que el servidor ha respondido una petición del cliente, la conexión se elimina entre ambos.

## **2.4. Modelo Cliente/Servidor**

Los clientes (o programas que representan entidades que necesitan servicios) y los servidores (o programas que proporcionan servicios) son objetos separados desde un punto de vista lógico y que se comunican a través de una red de comunicaciones para realizar una o varias tareas de forma conjunta. Un cliente hace una petición de un servicio y recibe la respuesta a dicha petición, un servidor recibe y procesa la petición y devuelve la respuesta solicitada [17].

Cuando se utiliza un servicio en Internet como consultar una base de datos, transferir un fichero o participar en un foro de discusión, se establece un proceso en el que entran en juego dos partes. Por un lado el usuario quien ejecuta una aplicación en el ordenador local el denominado programa cliente. Este programa cliente se encarga de ponerse en contacto con el ordenador remoto para solicitar el servicio deseado. El ordenador remoto por su parte responderá a lo solicitado mediante un programa que está ejecutando. Este último se denomina programa servidor. Los términos cliente y servidor se utilizan tanto para referirse a los programas que cumplen estas funciones como a los ordenadores donde son ejecutados esos programas.

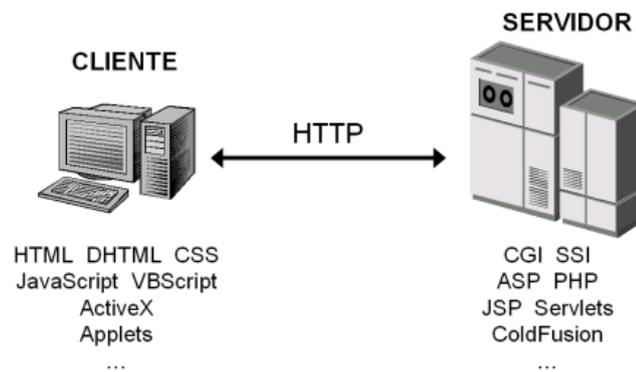
El programa o los programas cliente que el usuario utiliza para acceder a los servicios de Internet realizan dos funciones distintas. Por una parte se encargan de gestionar la comunicación con el ordenador servidor de solicitar un servicio concreto y de recibir los datos enviados por éste, y por otra es la herramienta que presenta al usuario los datos en pantalla y que le ofrece los comandos necesarios para utilizar las prestaciones que ofrece el servidor.

### 2.4.1. Lenguajes de programación del lado del cliente

Un lenguaje del lado cliente es totalmente independiente del servidor, lo cual permite que la página pueda ser albergada en cualquier sitio.

La gran ventaja de este tipo de lenguaje es que evita la recarga de trabajo en la parte del servidor de la aplicación, generando así una mayor agilidad en el desarrollo de un proceso.

En la Figura 13 muestra algunos lenguajes que se pueden utilizar tanto para cliente como servidor [18].



**Figura 13. Lenguajes tanto cliente-servidor**

#### 2.4.1.1. HTML (Hyper Text Markup Language)

Desde el surgimiento de internet se han publicado sitios web gracias al lenguaje HTML. Es un lenguaje estático para el desarrollo de sitios web (acrónimo en inglés de HyperText Markup Language, en español Lenguaje de Marcas Hipertextuales) [16]. Desarrollado por el World Wide Web Consortium (W3C). Fue creado en 1986 por el físico nuclear Tim Berners-Lee; el cual tomó dos herramientas preexistentes: El concepto de Hipertexto (Conocido también como link o ancla) el cual permite conectar dos elementos entre sí y el SGML (Lenguaje Estándar de Marcación General) el cual sirve para colocar etiquetas o marcas en un texto que indique como debe verse. (Ravioli, 2009).

Es sencillo permite describir hipertexto, el texto es presentado de forma estructurada y agradable. No necesita de grandes conocimientos cuando se cuenta con un editor de

páginas web. Consta de archivos pequeños un rápido despliegue, es de fácil aprendizaje y lo admiten todos los exploradores. (Valdez, 2009)

HTML constituye uno de los pilares sobre los que se asienta la web, es un lenguaje extensible al que se le pueden añadir características y funcionalidades mediante las CCS (Hojas de Estilo) y JavaScript obteniéndose como resultado páginas web rápidas y sencillas.

#### **2.4.2. Lenguajes de programación del lado del servidor.**

Los lenguajes de lado servidor son aquellos lenguajes que son reconocidos, ejecutados e interpretados por el propio servidor y que se envían al cliente en un formato comprensible para él [19]. Por otro lado los lenguajes de lado cliente son aquellos que pueden ser directamente "digeridos" por el navegador y no necesitan un pre-tratamiento.

Los lenguajes de programación a lado del servidor consiste en el procesamiento de una petición que el usuario realizó anteriormente a través del navegador.

Esta petición se interpreta mediante un script que se encuentra en el servidor de la aplicación, con el objetivo de generar páginas HTML dinámicamente con la respuesta a la petición realizada.

Hoy en día existen lenguajes del lado del servidor como Java, Python, Ruby, PHP, entre otros. Java está liberado en el mercado actual, debido a su portabilidad y constantes mejoras. Se profundizará el lenguaje Java ya que es la herramienta que se emplea en el presente trabajo.

## Discusión de Lenguajes

Aquí se discuten los lenguajes a ser utilizados en nuestra aplicación de tal manera ver cuál es el mejor para el desarrollo del mismo.

**TABLA I LENGUAJES DE PROGRAMACIÓN**

LENGUAJE	FRAMEWORKS APLICABLES	VENTAJAS	DESVENTAJAS
PYTHON	-DJANGO -TURBO GEARS	<ul style="list-style-type: none"> <li>• Velocidad de desarrollo.</li> <li>• Comunidad muy extendida.</li> <li>• Permite el uso de librerías de inteligencia artificial.</li> <li>• Gran poder de diseño.</li> <li>• Python es un lenguaje de programación de alto nivel cuya filosofía hace hincapié en una sintaxis muy limpia y que favorezca un código legible.</li> <li>• Se trata de un lenguaje de programación multiparadigma ya que soporta orientación a objetos, programación imperativa.</li> </ul>	Incompatibilidad con otros lenguajes
RUBY	RUBY ON RAILS	<ul style="list-style-type: none"> <li>• Arquitectura basada MVC</li> <li>• Ruby es un lenguaje de programación interpretado, reflexivo y orientado a objetos,</li> </ul>	No presenta proyectos aplicables y de renombre en inteligencia artificial
PHP	LARABEL	<ul style="list-style-type: none"> <li>• Arquitectura basada MVC</li> <li>• Es un lenguaje multiplataforma.</li> <li>• Completamente orientado al desarrollo de aplicaciones web dinámicas con acceso a información almacenada en una Base de Datos.</li> <li>• Programación segura y confiable.</li> <li>• Capacidad de conexión con la mayoría de los motores de base de datos.</li> </ul>	No presenta proyectos aplicables y de renombre en inteligencia artificial
ASP.NET	ASP MYSQL SERVER	<ul style="list-style-type: none"> <li>• Arquitectura basada MVC</li> </ul>	No presenta proyectos aplicables y

		<ul style="list-style-type: none"> <li>• ASP es una tecnología dinámica funcionando del lado del servidor.</li> <li>• La ventaja principal de las tecnologías dependientes del servidor radica en la seguridad que tiene el programador sobre su código.</li> </ul>	de renombre en inteligencia artificial
JAVA	<p>JSP</p> <p><b>Capa Web:</b> - Java Enterprise Edition</p> <p><b>Capa Presentación:</b> -Java Server Faces Facelets</p> <p><b>Capa Persistencia:</b> Java Persistence api Eclipse Link Hibernate</p>	<ul style="list-style-type: none"> <li>• Arquitectura basada MVC.</li> <li>• Comunidad muy extendida.</li> <li>• Permite el uso de proyectos desarrollados en java para redes neuronales.</li> <li>• Compatibilidad entre frameworks en todo nivel de arquitectura.</li> <li>• Compatibilidad para desarrollo con librerías nativas.</li> <li>• Compatibilidad con otros lenguajes matemáticos como Matlab y Lenguaje R</li> </ul>	Curva de aprendizaje lenta. Desarrollo medianamente rápido. Costos de implementación elevados.

Del análisis de la tabla I en comparación de los lenguajes, podemos decir que entre los distintos lenguajes para desarrollo web java es uno de los mejores candidatos ya que nos provee su plataforma Java Enterprise Edition, la cual nos proporciona los recursos necesario para el desarrollo de aplicación web MVC además de poseer servidores de aplicaciones dedicados para el desarrollo de aplicaciones empresariales (como Jboss). Existe una gran cantidad de proyectos de inteligencia artificial desarrollados en Java a más de la compatibilidad para acceso a librerías nativas JNI (JAVA NATIVE INTERFACE) y compatibilidad con lenguaje Weka.

#### 2.4.2.1. JAVA

Este es un lenguaje de programación bastante robusto basado en C Y C++, orientado a objetos basado en clases concurrentes y de propósito general, se trata de un lenguaje compilado es decir que se requiere ser traducido a partir de su código fuente por medio de un compilador en un archivo ejecutable para una plataforma determinada, una vez

compilado se puede ejecutar varias veces sin la necesidad de compilarlo en cada momento, en este sentido Java es un lenguaje multiplataforma con la particularidad de que requiere un máquina virtual denominada JVM [16,20].

### **Características**

- Orientado a objetos
- Distribuido
- Seguro
- Robusto
- Portable etc.

La novedad que aporta Java dentro de las nuevas generaciones de navegadores es la capacidad de desplazar el control de la interactividad de los servicios hacia las computadoras de los usuarios

## **2.5. FRAMEWORK**

Un Framework [21] brinda una estructura conceptual y tecnológica que ayuda a la parte gráfica de un sistema informático, lo que hace comúnmente con artefactos y módulos de software concretos, que se implementarán en una aplicación web esto con motivos de agilidad en la aplicación y sobre todo la funcionalidad.

A continuación se citan varios de los frameworks disponibles en el mercado para lenguaje Java.

- Ext-Gwt
- Vaadin
- ZK
- SmartGWT
- Java Server Faces
- IceFaces
- RichFaces

En el presente proyecto se ha decidido trabajar JSF conjuntamente con RichFaces primeramente porque permite crear fácilmente componentes para la interfaz gráfica ya que ofrece amplios componentes, además por su compatibilidad con JSP y experiencia en el desarrollo.

### **2.5.1. JSF**

JSF es una tecnología y framework para aplicaciones Java basadas en web que simplifica el desarrollo de interfaces de usuario.

JSF una de las características que tiene es que introduce una serie de etapas en el procesamiento de la petición, como por ejemplo la de validación, reconstrucción de la vista, recuperación de los valores de los elementos. Es extensible pudiendo crear nuevos elementos de la interfaz o de los que ya existen.

JSF incluye un conjunto de API para representar componentes de la interfaz de usuario y administrar su estado, manejar eventos, validar entradas, definir un esquema de navegación de las páginas y dar un soporte para internacionalización y accesibilidad.

JSF permite desarrollar rápidamente aplicaciones de negocio dinámicas en las que toda la lógica de negocio se implementa en Java o es llamado desde Java, creando páginas para la vista muy sencilla.

Es un framework que ejecuta muchas peticiones al servidor. Para optimizar dicho tráfico están empezando a aparecer implementaciones de JSF que incorpora Ajax en sus etiquetas. Esto permite actualizar los componentes en el navegador del usuario de manera selectiva, sin necesidad de recargar la página completa. La combinación de JSF Ajax brinda las páginas gran dinamismo sin complicar el desarrollo, evitando el uso de Java Script

Existen diversas implementaciones de JSF basadas en Ajax tales son RichFaces, IceFaces entre otras. Estas implementaciones cobran mayor importancia debido a su divulgación entre la comunidad web.

Para el desarrollo de nuestra interfaz se desarrolló RichFaces a continuación se detalla:

### **2.5.2. RichFaces**

RichFaces es un framework de que brinda capacidad Ajax dentro de aplicaciones JSF sin la necesidad de recurrir a JavaScript. RichFaces incluye un ciclo de vida, validaciones, conversores y la gestión de recursos estáticos y dinámicos [22].

RichFaces es un marco muy útil de código abierto que le permite añadir capacidades de Ajax a sus aplicaciones JSF (usando los componentes estándar JSF), sin la necesidad de escribir código JavaScript y administrar la compatibilidad de JavaScript entre navegadores. Se integra con el ciclo de vida de JSF y otras características de JSF estándar como la validación, la conversión y administración de recursos.

#### **Características**

- Rich Faces está completamente integrado en el ciclo de vida de JSF.
- El framework proporciona dos librerías de componentes (Core Ajax y la interfaz de usuario).
- Rich Faces permite definir eventos en la propia página.
- La librería UI (Interfaz de usuario) que contiene componentes para agregar características de interfaz de usuario a aplicaciones JSF.

#### **Ventajas**

Al pertenecer RichFaces a un subproyecto de JBoss, su integración con Seam es perfecta, y al ser RichFaces es propiedad de Exadel, se ajusta perfectamente al IDE Red Hat Developer Studio el cual permite desarrollar aplicaciones visuales con RichFaces de forma fácil.

## 2.6. Hibernate

Hibernate es un entorno de trabajo cuyo objetivo es facilitar la persistencia de objetos, Java en bases de datos relacionales y a su vez la consulta de estas bases de datos para obtener objetos, una primera parte del proceso de Hibernate se denomina "Mapeo", significa que adapta a la base de datos por completo en la aplicación web, realiza un Mapeo Objeto-Relacional (ORM) ya que adopta los estándares de las bases de datos relacionales, ésta herramienta está principalmente dedicada al lenguaje Java, aunque está disponible también para el entorno .Net de Microsoft.

Es una herramienta de Mapeo objeto-relacional para la plataforma Java, que facilita el mapeo de atributos entre un base de datos relacional tradicional y el modelo de objetos de un aplicación, mediante archivos declarativos XML, que permiten establecer estas relaciones o mediante anotaciones.

Vale la pena mencionar que se trata de una herramienta de libre distribución, bajo los términos de la licencia GNU LGPL [23].

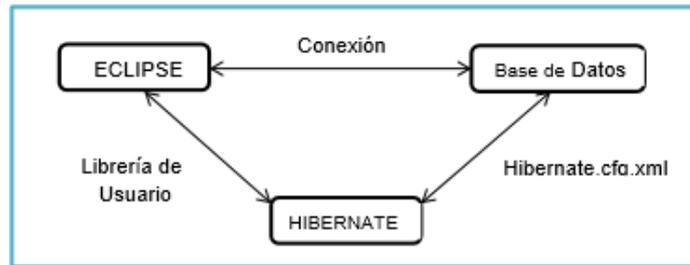
Hibernate permite a la aplicación manipular los datos de la base a través de la aplicación, actuando sobre los objetos con las características propias de la programación orientada a objetos [14].

Hibernate genera las secuencias SQL y libera al desarrollador del manejo manual de los datos que resultan de la ejecución de dichas sentencias, manteniendo la portabilidad entre todos los motores de bases de datos con un ligero incremento en el tiempo de ejecución.

### Características

- No es intrusivo.
- Es difícil testeo.
- Posee una buena documentación.
- El editor de mapeo facilita el manejo de los archivos XML.
- La consola de Hibernate permite configurar las conexiones a la base de datos.
- Ingeniería inversa, genera las clases de dominio y archivos de mapeo Hibernate.

En la figura 15 se puede apreciar el trabajo que realiza Hibernate dentro de eclipse



**Figura 14. Esquema del funcionamiento de Hibernate**

De la figura anterior eclipse se comunica con la base de datos a través de la conexión JDBC dependiendo de la base de datos que se emplea. Existen varios métodos o patrones de programación que ayudan con la organización y la arquitectura del código fuente de un sistema aplicación ,comúnmente se ven patrones como el MVC o el desarrollo por capas, pero se considera una mejor practica emplear el patrón MVC.

## 2.7. Modelo Vista-Controlador

Se trata de un patrón de arquitectura de las aplicaciones de software, su principal característica es que separa completamente la lógica de negocio de la interfaz de usuario y a su vez de la lógica de control empleada en el desarrollo. Este patrón de arquitectura, fue descrito inicialmente en el año 1979 para la compañía Smalltalk y se ha venido empleando desde entonces [24].

Muy comúnmente en aplicaciones Web se ve reflejado en el diseño el patrón MVC, se lo emplea debido a los constantes cambios que requiere la aplicación a lo largo del tiempo, es importante mantener los bloques de trabajo bien definidos e independientes, de tal manera que los cambio realizados en un bloque, se vean reflejados en otro bloque sin necesidad de grandes cambios a nivel de código fuente.

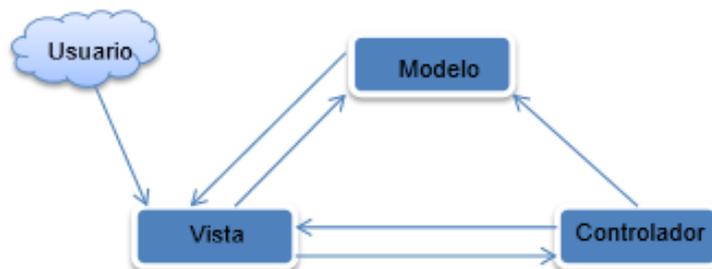
Desarrollando las partes del patrón MVC, tenemos que:

**El modelo.** Es el responsable de acceder a la capa de almacenamiento de datos. Lo ideal es que el modelo sea independiente del sistema de almacenamiento, define las reglas de negocio (la funcionalidad del sistema) [14].

**El controlador.** Se encarga de recibir los eventos de entrada (un clic, un cambio en un campo de texto, etc.).Contiene reglas de gestión de eventos, del tipo “SI Evento Z,

entonces Acción W". Estas acciones pueden suponer peticiones al modelo o a las vistas. Una de estas peticiones a las vistas puede ser una llamada al método "Actualizar ()". Una petición al modelo puede ser "Obtener\_Distributivo (\$parámetros)" [14].

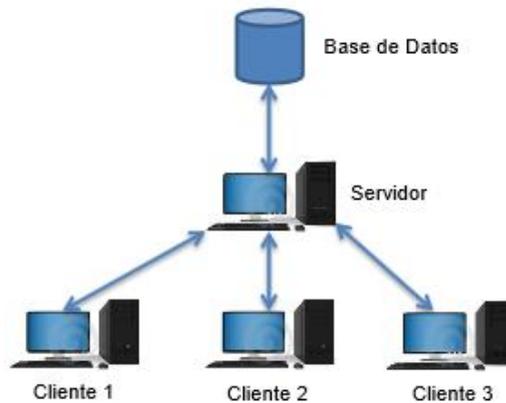
**La vista.** Responsable de recibir datos del modelo mostrándolos al usuario. Tiene un registro de su controlador asociado (normalmente porque además lo instancia). Puede dar el servicio de "Actualización ()", para que sea invocado por el controlador o por el modelo (cuando es un modelo activo que informa de los cambios en los datos producidos por otros agentes). En la figura 16 podemos apreciar la manera en la cual trabaja este patrón [14].



**Figura 15. Esquema del funcionamiento MVC**

## 2.7. Bases de Datos

Una Base de Datos no es más que un sistema de almacenamiento de información, en el cual maneja aspectos relacionados con la seguridad tratamiento, y consulta de datos, la estructura que usualmente se maneja en una aplicación web es la que se muestra a continuación en la Figura 17



**Figura 16. Funcionamiento de un BD empleando la arquitectura cliente -servidor**

Hoy en día existen varias opciones en cuanto a las bases de datos que se disponen en el mercado, tanto de software libre como software comercial, vale la pena mencionar que la mayor parte de bases de datos adaptan su estructura para aplicaciones web así como también para aplicaciones de escritorio ejemplos de estas Bases de Datos.

- Oracle
- PostgreSQL
- MySQL
- SQL Server... etc.

### **2.7.1. Herramienta para almacenar los datos.**

De entre este conjunto de base de datos en el desarrollo de la tesis se ha empleado MySQL. Considerando factores como la experiencia con la misma y además de ser una buena base de datos libre. Dispone de un modelar (ENTIDAD/RELACION) que transforma los modelos desarrollados de manera gráfica en script ejecutables.

#### **MySQL**

MySQL es un gestor de base de datos sencillo de usar e increíblemente rápido. También es uno de los motores de base de datos más usados en Internet, la principal razón de esto es que es gratis para aplicaciones no comerciales. [25]

MySQL es un sistema para la administración de bases de datos relacional (RDBMS) rápido y sólido creado por la empresa sueca MySQL AB. Está disponible desde 1996

pero su nacimiento se remonta a 1979. Ha obtenido el galardón Choice Award del Linux Journal Readers en varias ocasiones. (Gamboa & Nuez, 2009)

Es sistema gestor de base de datos relacional donde los datos están almacenados en tablas entre las cuales se establecen unas relaciones para manejar los datos de una forma eficiente y segura. Para usar y gestionar una base de datos relacional se usa el lenguaje estándar de programación SQL.

El código fuente de MySQL se puede descargar y está accesible a cualquiera, por otra parte usa la licencia GPL para aplicaciones no comerciales. [26]. MySQL usa el GPL (GNU General Public License) para definir qué puede hacer y que no puede hacer con el software en diferentes situaciones.

### **Características de SQL**

- Soporte de transacciones.
- Escalabilidad, estabilidad y seguridad.
- Soporta procedimientos almacenados.
- Permite trabajar en modo cliente-servidor, donde la información y datos se alojan en el servidor y las terminales o clientes de la red sólo acceden a la información.
- Además permite administrar información de otros servidores de datos.

## **2.8. Lenguajes de programación para Redes Neuronales**

Actualmente existen varios paquetes y complementos cada uno con sus propias características, y una diversidad de herramientas de software desarrollado bajo licencia GPL (Licencia Pública General), WEKA, RapidMiner, Tanagra, Pentaho, Kxen, Orange, SSPS Clementine, Estadística R, Toolbox entre otros a continuación detallaremos el que será utilizado para el desarrollo de nuestra aplicación es WEKA, ya que me ofrece funciones de Perceptrón multicapa y se adapta al desarrollo de aplicaciones web y para las distintas pruebas estarán hechas en el lenguaje R con la librería Rweka, por su facilidad de uso y por sus distintas funciones que ofrece. Se han evaluado pues distintos aspectos siendo especialmente destacables

sus bondades en lo que se refiere a calidad, a la cantidad de técnicas y funciones implementadas, a que es libre y a la gran comunidad científica que lo usa como estándar para el análisis de datos.

### **2.8.1. Lenguaje R**

R es un lenguaje y entorno para el tratamiento numérico de datos y la creación de gráficos. R implementa un dialecto del lenguaje S y provee una amplia variedad de técnicas estadísticas y gráficas (modelización lineal y no lineal, test estadísticos, análisis de series temporales, clasificación, ordenación, etc.) [27].

R es un lenguaje de programación y entorno de software de código abierto para computación y gráficos estadísticos, se desarrolla mediante librerías (también llamadas en R como paquetes) que lo que hacen es completar el lenguaje con nuevos desarrollos previstos para distintas áreas del análisis estadístico y gráfico de los datos.

Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásica, clasificación, agrupación en clústeres entre otros.

R es fácilmente extensible al permitir definir nuevas funciones e incluso bibliotecas (librerías), de las que existen numerosas ya disponibles. El código de R está disponible como software libre bajo las condiciones de la licencia GNU-GPL [28]

R es un lenguaje y entorno de programación libre, para análisis estadístico y gráfico [29]. Este potente lenguaje de comandos y ambiente integrado de programación para la carga, manejo, análisis y grafica de datos, es utilizado para experimentar con conceptos relacionados a la estadística. R provee una gran variedad de técnicas y gráficas para estadística.

Como entorno de programación básicamente se trata de una consola (ventana de trabajo) sobre la que se van introduciendo scripts (instrucciones más o menos complejas) que se ejecutan sobre los datos previamente cargados (conjuntos o ventanas de datos).

El entorno de R se caracteriza por su flexibilidad e incluye, entre otros:

- Un buen gestor de datos.
- Un conjunto de operadores para cálculos en arrays (vectores de gran tamaño)
- Un conjunto integrado de herramientas de análisis de datos
- Funciones gráficas para análisis y visualización de los datos.

R proporciona:

- Un conjunto coherente y extensivo de instrumentos para el análisis y el tratamiento estadístico de datos.
- Un lenguaje para expresar modelos estadísticos y herramientas para manejar modelos lineales y no lineales.
- Utilidades gráficas para el análisis de datos y la visualización en cualquier estación gráfica o impresora.
- Un eficiente lenguaje de programación orientado a objetos, que crece fácilmente merced a la comunidad de usuarios.

Se trata de un software libre, distribuido bajo licencia GPL, muy extendido en la comunidad universitaria y que está llamado a cobrar un papel cada vez más relevante en el mundo de las aplicaciones profesionales y de la empresa.

### **2.8.2. Weka (Waikato Environment for Knowledge Analysis)**

Es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.

Weka ('Waikato Environment for Knowledge Analysis') es un software empleado en aprendizaje automático y minería de datos escrito en 'Java' y desarrollado en la Universidad de Waikato. Weka es un software libre distribuido bajo licencia GNU-GPL

Weka es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Este programa escrito en Java, permite analizar datos mediante diversas técnicas de aprendizaje automático. WEKA se distribuye como software libre. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, agrupamiento, asociación y visualización [30]. La librería WEKA Group (2012) es una colección de algoritmos de aprendizaje para tareas de minería de datos. Agrupa diferentes herramientas para preprocesado de datos, agrupamiento o clustering, clasificación, regresión, generación de reglas de asociación, etc. También incluye facilidades para la visualización de los datos

## **CAPÍTULO III: REDES NEURONALES SUPERVISADAS**

### **3.1. Inteligencia Artificial**

Del latín intellegentia, es la capacidad de entender, asimilar, elaborar información y utilizarla para resolver problemas.

La IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos en base a dos de sus características primordiales: el razonamiento y la conducta. Existen distintas definiciones de IA de acuerdo a distintos enfoques algunas de estas definiciones se muestran a continuación: [31]

- La interesante tarea de lograr que las computadoras piensen máquinas con mente, en su amplio sentido literal. (Haugeland, 1985)
- La automatización de actividades que vinculamos con procesos de pensamiento humano, actividades tales como la toma de decisiones, resolución de problemas, aprendizaje. (Bellman, 1978)
- Organismo o ente capaz de tomar alguna decisión y de recordar sucesos de su entorno, con el objetivo de usarlos para la toma de decisiones.
- Organismo o ente que con el paso del tiempo aprende de su entorno, acumulando experiencia y que es usada en nuevas situaciones que le presenta el entorno.

- El estudio de las facultades mentales mediante el uso de modelos computacionales.
- El estudio de los cálculos que permiten, razonar y actuar.” (Winston, 1992).
- El arte de crear máquinas con capacidad de realizar funciones que realizadas por personas requieren de inteligencia.” (Kurzweil, 1990).
- El estudio de cómo lograr que las computadoras realicen tareas que, por el momento, los humanos hacen mejor.” (Rich y Knight, 1991).
- Un campo de estudio que se enfoca a la explicación y emulación de la conducta inteligente en función de procesos computacionales.” (Schalkoff, 1990).
- La rama de la ciencia de la computación que se ocupa de la automatización de la conducta inteligente.” (Luger y Stubblefield, 1993).

### **3.2. Redes Neuronales**

El cerebro es un procesador de información con unas características muy notables es capaz de procesar a gran velocidad grandes cantidades de información procedentes de los sentidos, combinarla o compararla con la información almacenada y dar respuestas adecuadas incluso en situaciones nuevas [32]. Logra discernir un susurro en una sala ruidosa, distinguir una cara en una calle mal iluminada o leer entre líneas en una declaración política; pero lo más impresionante de todo es su capacidad de aprender a representar la información necesaria para desarrollar tales habilidades sin instrucciones explícitas para ello. Aunque todavía se ignora mucho sobre la forma en que el cerebro aprende a procesar la información, se han desarrollado modelos que tratan de mimetizar tales habilidades denominados redes neuronales artificiales o modelos de computación conexionista.

Con esta amplia Introducción a las redes neuronales, se pretende dar a conocer los elementos básicos de lo que comúnmente se denomina Inteligencia Artificial, para así comprender de qué modo pueden llegar a «pensar» y «aprender» las máquinas.

Así la inteligencia artificial [33] es un intento por descubrir y describir aspectos de la inteligencia humana que pueden ser simulados mediante máquinas. Esta disciplina se ha desarrollado fuertemente en los últimos años teniendo aplicación en algunos campos como visión artificial, demostración de teoremas, procesamiento de información expresada mediante lenguajes humanos... etc. Las redes neuronales son otra forma de emular otra de las características propias de los humanos, la capacidad de memorizar y asociar hechos. Si examinamos con atención aquellos problemas que no pueden expresarse a través de un algoritmo nos daremos cuenta de que todos ellos tienen una característica común: la experiencia. El hombre es capaz de resolver estas situaciones acudiendo a la experiencia acumulada. Así parece claro que una forma de aproximarse al problema consista en la construcción de sistemas que sean capaces de reproducir esta característica humana. En definitiva las redes neuronales no son más que un modelo artificial y simplificado del cerebro humano, que es el ejemplo más perfecto del que disponemos de sistema que es capaz de adquirir conocimiento a través de la experiencia. Una red neuronal es un nuevo sistema para el tratamiento de la información cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano, la neurona [34].

### **3.2.1. Historia de las Redes Neuronales**

Fue en 1943 cuando Warren McCulloch y Walter Pitts propusieron el clásico modelo de neurona en el que se basan las redes neuronales actuales. Seis años después, en 1949, en su libro *The Organization of Behavior*, Donald Hebb presentaba su conocida regla de aprendizaje. En 1956 se organizó en Dartmouth la primera conferencia sobre IA. Aquí se discutió el uso potencial de las computadoras para simular “todos los aspectos del aprendizaje o cualquier otra característica de la inteligencia” y se presentó la primera simulación de una red neuronal, aunque todavía no se sabían interpretar los datos resultantes.

En 1957, Frank Rosenblat presentó el Perceptrón, una red neuronal con aprendizaje supervisado cuya regla de aprendizaje era una modificación de la propuesta por Hebb. El Perceptrón trabaja con patrones de entrada binarios, y su funcionamiento por tratarse de una red supervisada se realiza en dos fases: una primera en la que se presentan las entradas y las salidas deseadas, en esta fase la red aprende la salida que debe dar para

cada entrada. La principal aportación del Perceptrón es que la adaptación de los pesos se realiza teniendo en cuenta el error entre la salida que da la red y la salida que se desea. En la fase siguiente de operación la red «es capaz» de responder adecuadamente cuando se le vuelven a presentar los patrones de entrada.

En 1959, Widrow publica una teoría sobre la adaptación neuronal y unos modelos inspirados en esa teoría, el Adaline (Adaptative Linear Neuron) y el Madaline (Multiple Adaline). Estos modelos fueron usados en numerosas aplicaciones y permitieron usar por primera vez una red neuronal en un problema importante del mundo real filtros adaptativos para eliminar ecos en las líneas telefónicas.

En los años 60 se propusieron otros dos modelos, también supervisados basados en el Perceptrón de Rosenblat denominados Adaline y Madaline. En estos la adaptación de los pesos se realiza teniendo en cuenta el error, calculado como la diferencia entre la salida deseada y la dada por la red, al igual que en el Perceptrón. Sin embargo la regla de aprendizaje empleada es distinta. Se define una función error para cada neurona que da cuenta del error cometido para cada valor posible de los pesos cuando se presenta una entrada a la neurona. Así, la regla de aprendizaje hace que la variación de los pesos se produzca en la dirección y sentido contrario del vector gradiente del error. A esta regla de aprendizaje se la denomina Delta.

### **3.2.2. Definiciones de la red neuronal**

Existen numerosas formas de definir a las redes neuronales desde las definiciones cortas y genéricas hasta las que intentan explicar más detalladamente qué son las redes neuronales. Por ejemplo:

- Una nueva forma de computación inspirada en modelos biológicos.
- Un modelo matemático compuesto por un gran número de elementos procesales organizados en niveles [35].
- Un sistema de computación compuesto por un gran número de elementos simples, elementos de procesos muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas [35].
- Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización

jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico. [35].

Definiciones según algunos autores:

- Haykin, S.: "Una red neuronal es un procesamiento distribuido masivamente paralelo que tiene una tendencia natural para almacenar conocimiento empírico y hacerlo disponible para el uso [36]. Recuerda al cerebro en dos aspectos:
  1. Conocimiento se adquiere por la red a través de un proceso de aprendizaje.
  2. Las conexiones interneurónicas se conocen como pesos sinápticos y se usan para almacenar el conocimiento."
- Zurada, J.M.: "Sistemas de redes neuronales artificiales, o redes neuronales son sistemas celulares físicos que puedan adquirir, almacenar y usar conocimiento empírico [36]."
- Las redes neuronales son una forma de un sistema computarizado multi-proceso con:
  - ✓ Elementos de procesamiento sencillos.
  - ✓ Alto grado de interconexión.
  - ✓ Mensajes simples escalares.
  - ✓ Interacción adaptable entre elementos.
- El concepto de Red Neuronal Artificial está inspirado en las Redes Neuronales Biológicas. Una Red Neuronal Biológica es un dispositivo no lineal altamente paralelo, caracterizado por su robustez y su tolerancia a fallos. Sus principales características son las siguientes:
  - ✓ Aprendizaje mediante adaptación de sus pesos sinápticos a los cambios en el entorno.
  - ✓ Manejo de imprecisión, ruido e información probabilística.
  - ✓ Generalización a partir de ejemplos.

### 3.2.3. Ventajas de las Redes Neuronales

Debido a su constitución y a sus fundamentos las RNA presentan un gran número de características semejantes a las del cerebro. Por ejemplo son capaces de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas.

Estas ventajas incluyen:

En el proceso de aprendizaje, los enlaces ponderados de las neuronas se ajustan de manera que se obtengan unos resultados específicos. Una RNA no necesita un algoritmo para resolver un problema, ya que ella puede generar su propia distribución de los pesos de los enlaces mediante el aprendizaje. También existen redes que continúan aprendiendo a lo largo de su vida, después de completado su periodo inicial de entrenamiento.

- ✓ *Aprendizaje Adaptativo*: Es una de las características más atractivas de las redes neuronales, es la capacidad de aprender a realizar tareas basadas en un entrenamiento o una experiencia inicial
- ✓ *Auto organización*: Las redes neuronales usan su capacidad de aprendizaje adaptativo para organizar la información que reciben durante el aprendizaje y/o la operación. Una RNA puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje. Este auto organización provoca la facultad de las redes neuronales de responder apropiadamente cuando se les presentan datos o situaciones a los que no habían sido expuestas anteriormente.
- ✓ *Tolerancia a Fallos*: Comparados con los sistemas computacionales tradicionales, los cuales pierden su funcionalidad en cuanto sufren un pequeño error de memoria, en las redes neuronales, si se produce un fallo en un pequeño número de neuronas, aunque el comportamiento del sistema se ve influenciado, sin embargo no sufre una caída repentina.

La razón por la que las redes neuronales son tolerantes a fallos es que tienen su información distribuida en las conexiones entre neuronas, existiendo cierto grado de redundancia en ese tipo de almacenamiento, a diferencia de la mayoría de los ordenadores algorítmicos y sistemas de recuperación de

datos que almacenan cada pieza de información en un estado único, localizado y direccionable.

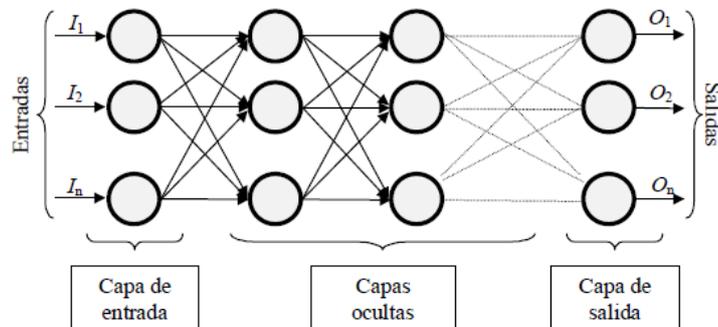
- ✓ *Operación en Tiempo Real:* Los computadores neuronales pueden ser realizados en paralelo, y se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.
- ✓ *Fácil inserción dentro de la tecnología existente:* Debido a que una red puede ser rápidamente entrenada, comprobada, verificada y trasladada a una implementación hardware de bajo costo, es fácil insertar RNA para aplicaciones específicas dentro de sistemas existentes (chips, por ejemplo). De esta manera, las redes neuronales se pueden utilizar para mejorar sistemas de forma incremental, y cada paso puede ser evaluado antes de acometer un desarrollo más amplio.

#### **3.2.4. Desventajas de las Redes Neuronales**

- ✓ Complejidad de aprendizaje para grandes tareas, cuanto más cosas se necesiten que aprenda una red, más complicado será enseñarle.
- ✓ Tiempo de aprendizaje elevado. Esto depende de dos factores: primero si se incrementa la cantidad de patrones a identificar o clasificar y segundo si se requiere mayor flexibilidad o capacidad de adaptación de la red neuronal para reconocer patrones que sean sumamente parecidos, se deberá invertir más tiempo en lograr que la red converja a valores de pesos que representen lo que se quiera enseñar
- ✓ No permite interpretar lo que se ha aprendido, la red por si sola proporciona una salida, un número que no puede ser interpretado por ella misma, sino que se requiere de la intervención del programador y de la aplicación en si para encontrarle un significado a la salida proporcionada.
- ✓ Elevada cantidad de datos para el entrenamiento, cuanto más flexible se requiera que sea la red neuronal, más información tendrá que enseñarle para que realice de forma adecuada la identificación. (Hilera 1995, Freeman 1993).

### 3.2.5. Elementos básicos de una red neuronal

A continuación se puede ver en la figura 18, un esquema de una red neuronal:



**Figura 17. Ejemplo de una red totalmente conectada**

La misma está constituida por neuronas interconectadas y arregladas en tres capas (esto último puede variar). Los datos ingresan por medio de la “capa de entrada”, pasan a través de la “capa oculta” y salen por la “capa de salida”. Cabe mencionar que la capa oculta puede estar constituida por varias capas.

La distribución de neuronas dentro de la red se realiza formando niveles o capas, con un número determinado de dichas neuronas en cada una de ellas. A partir de su situación dentro de la red se pueden distinguir tres tipos de capas:

- *De entrada*: es la capa que recibe directamente la información proveniente de las fuentes externas de la red.
- *Ocultas*: son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de redes neuronales, están encargadas de realizar el trabajo de la red
- *De salidas*: transfieren información de la red hacia el exterior.

### 3.3. Redes neuronales monocapa con aprendizaje supervisado

Las redes monocapa se utilizan típicamente en tareas relacionadas en lo que se conoce como auto asociación por ejemplo para regenerar informaciones de entrada que se presentan a la red incompleta o distorsionada.

### **3.3.1. El Perceptrón**

La primera red neuronal conocida, fue desarrollada en 1943 por Warren McCulloch y Walter Pitts; la cual consistía en una suma de las señales de entrada, multiplicadas por unos valores de pesos escogidos aleatoriamente. La entrada es comparada con un patrón preestablecido para determinar la salida de la red. Si en la comparación la suma de las entradas multiplicadas por los pesos es mayor o igual que el patrón preestablecido la salida de la red es uno (1), en caso contrario la salida es cero (0). Al inicio del desarrollo de los sistemas de inteligencia artificial, se encontró gran similitud entre su comportamiento y el de los sistemas biológicos y en principio se creyó que este modelo podía computar cualquier función aritmética o lógica [37].

La red tipo Perceptrón fue inventada por el psicólogo Frank Rosenblatt en el año 1957. Su intención era ilustrar algunas propiedades fundamentales de los sistemas inteligentes en general, sin entrar en mayores detalles con respecto a condiciones específicas y desconocidas para organismos biológicos concretos. Rosenblatt creía que la conectividad existente en las redes biológicas tiene un elevado porcentaje de aleatoriedad, por lo que se oponía al análisis de McCulloch Pitts en el cual se empleaba lógica simbólica para analizar estructuras bastante idealizadas. Rosenblatt opinaba que la herramienta de análisis más apropiada era la teoría de probabilidades, y esto lo llevó a una teoría de separabilidad estadística que utilizaba para caracterizar las propiedades más visibles de estas redes de interconexión ligeramente

### **Características**

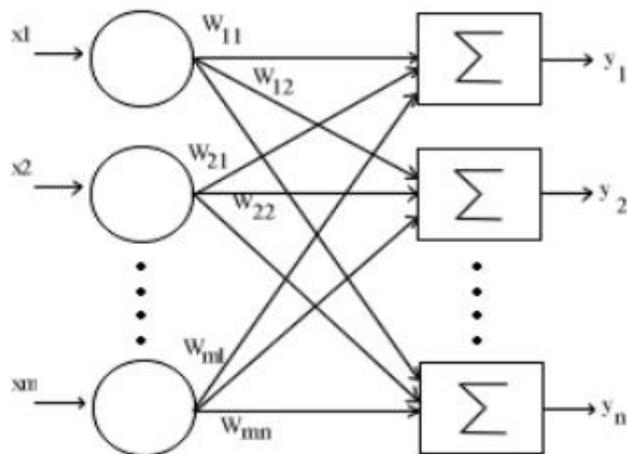
El Perceptrón simple es un modelo neuronal sin capa oculta, el cual maneja información binaria a su entrada y a su salida y su regla de aprendizaje por corrección de error es de tipo supervisado, realizando un entrenamiento offline tiene una gran aplicación para reconocimiento de patrones sencillos de tipo linealmente separables, por ejemplo clasificar compuertas lógicas como AND y OR o aplicaciones más específicas como el reconocimiento de caracteres impresos alfa-numéricos.

### **3.3.2. Arquitectura**

A pesar de que una sola neurona puede realizar modelos simples de funciones su mayor productividad viene dada cuando se organizan en redes. La red más simple es la

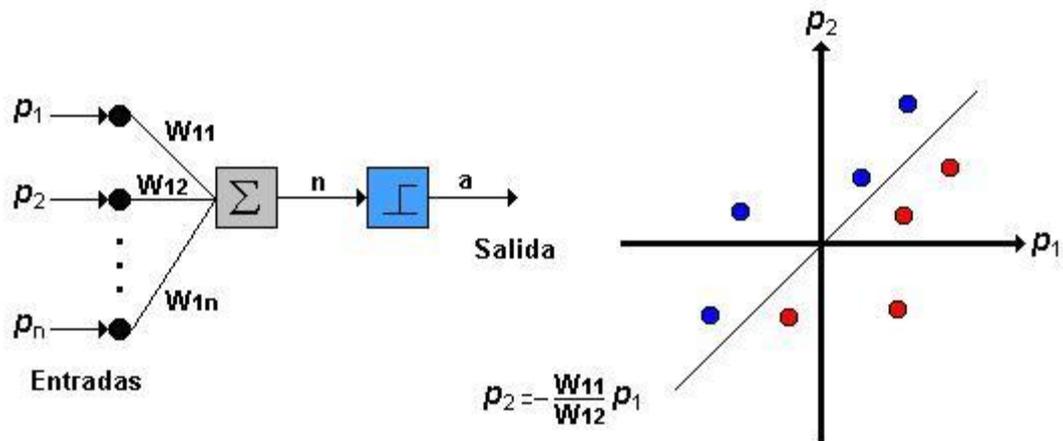
formada por un conjunto de perceptrones a los que entra un patrón de entradas y proporcionan la salida correspondiente. Por cada Perceptrón que tengamos en la red vamos a tener una salida, que se hallará como se hacía con un Perceptrón solo haciendo el sumatorio de todas las entradas multiplicadas por los pesos. Al representar añade una "capa" inicial que no es contabilizada a efectos de computación, solamente sirve para distribuir las entradas entre los perceptrones. La denominaremos la capa 0.

De esta manera, la representación gráfica de una red de capa simple sería la siguiente  
 Figura 19



**Figura 18. Arquitectura de una Red Perceptrón**

La red Perceptrón puede tener únicamente una sola capa, debido a que la regla de aprendizaje del Perceptrón es capaz de entrenar solamente una capa. Esta restricción coloca limitaciones en cuanto a lo que un Perceptrón puede realizar computacionalmente.



**Figura 19. Perceptrón Simple**

La única neurona de salida del Perceptrón realiza la suma ponderada de las entradas, resta el umbral y pasa el resultado a una función de transferencia de tipo escalón. La regla de decisión es responder +1 si el patrón presentado pertenece a la clase A, o -1 si el patrón pertenece a la clase B ver Figura 20 la salida depende de la entrada neta ( $n$  = suma de la entradas  $x_i$  ponderadas. [38].

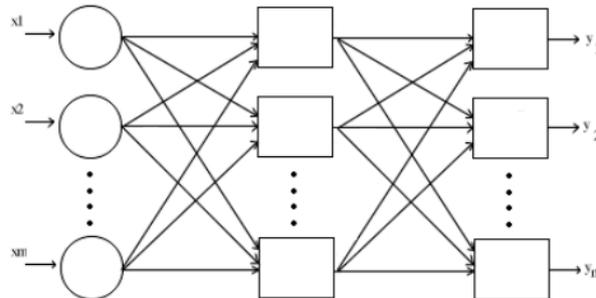
### 3.4. Redes neuronales multicapa con aprendizaje supervisado.

Las redes multicapa se forman por un conjunto de redes de capa simple en cascada unidas por pesos, donde la salida de una capa es la entrada de la siguiente capa. Generalmente son capaces de aprender funciones que una red de capa simple no puede aprender, por lo que ofrecen mejores capacidades computacionales. Para que este incremento en poder computacional sea tal tiene que existir una función de activación no lineal entre las capas, por lo que generalmente se utilizará una función de activación sigmoidea en detrimento de la lineal o umbral [39].

El Perceptrón Multicapa es un tipo de Red que está compuesta por varios perceptrones y que permite clasificar adecuadamente más de dos clases. Estas neuronas están organizadas mediante capas, las cuales transmitan la información de capa en capa. La característica de este tipo de Red es que sus conexiones están hechas de atrás hacia adelante y que además las neuronas de la misma capa no se relacionan entre sí. Figura 21.

Las redes multicapas son aquellas que disponen de un conjunto de neuronas agrupadas en varios (2, 3, etc.) niveles o capas

Para calcular la salida de una red multicapa se debe hacer de la misma manera que en las redes de capa simple, teniendo en cuenta que las salidas de una capa son las entradas de la siguiente capa.



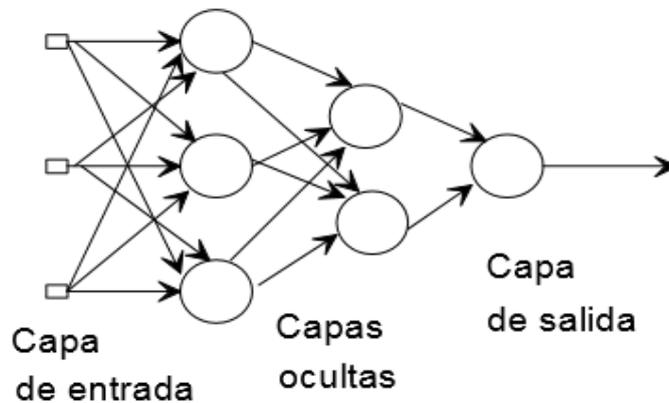
**Figura 20. Arquitectura del Perceptrón multicapa**

### 3.4.1. Arquitectura

La arquitectura de este tipo de red se caracteriza porque tiene todas sus neuronas agrupadas en distintos niveles llamados capas. El primer nivel corresponde a la capa de entrada que se encarga únicamente de propagar por el resto de la red las entradas recibidas y el último nivel es el de la capa de salida. Se encarga de proporcionar los valores de salida de la red. En las capas intermedias denominadas capas ocultas se realiza un procesamiento no lineal de los patrones recibidos.

Las conexiones del Perceptrón multicapa son hacia adelante. Generalmente todas las neuronas de un nivel se conectan con todas las neuronas de la capa inmediatamente posterior. A veces dependiendo de la red se encuentran conexiones de neuronas que no están en niveles consecutivos o alguna de las conexiones entre dos neuronas de niveles consecutivos no existe, es decir el peso asociado a dicha conexión es constante e igual a cero.

El Perceptrón multicapa es una red formada por una capa de entrada al menos una capa oculta y una de salida su estructura se muestra en la Figura 22



**Figura 21. Estructura de un Perceptrón multicapa**

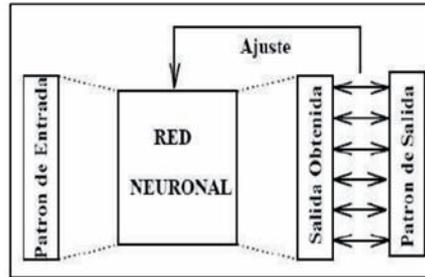
Este modelo se compone de la siguiente manera:

- Capa de entrada: sólo se encarga de recibir las señales de entrada y propagarla a la siguiente capa.
- Capa de salida: proporciona al exterior la respuesta de la red para cada patrón de entrada.
- Capas ocultas: realizan un procesamiento no lineal de los datos de entrada.

### **3.5. Tipos de Aprendizaje**

#### **3.5.1. Aprendizaje Supervisado**

El aprendizaje supervisado se caracteriza porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. [40] El supervisor controla la salida de la red y en caso de que ésta no coincida con la deseada se procederá a modificar los pesos de las conexiones con el fin de conseguir que la salida obtenida se aproxime a la deseada. Figura 23



**Figura 22. Estructura de un aprendizaje supervisado**

El entrenamiento consiste en presentar un vector de entrada a la red para calcular la salida de la red, compararla con la salida deseada y el error o diferencia resultante se utiliza para realimentar la red y cambiar los pesos de acuerdo con un algoritmo que tiende a minimizar el error. El objetivo del algoritmo de aprendizaje es ajustar los pesos de la red de manera tal que la salida generada por la ANN sea lo más cercanamente posible a la verdadera salida dada una cierta entrada.

En este tipo de aprendizaje se suelen considerar a su vez tres formas de llevarlo a cabo que dan lugar a los siguientes aprendizajes supervisados:

- 1) Aprendizaje por corrección de error.
- 2) Aprendizaje por refuerzo.
- 3) Aprendizaje estocástico

### **1) Aprendizaje por corrección de error.**

Consiste en ajustar los pesos de las conexiones de la red en función de la diferencia entre los valores deseados y los obtenidos a la salida de la red, es decir, en función del error cometido en la salida.

### **2) Aprendizaje por refuerzo.**

Se trata de un aprendizaje supervisado más lento que el anterior, que se basa en la idea de no disponer de un ejemplo completo del comportamiento deseado, es decir, de no indicar durante el entrenamiento exactamente la salida que se desea que proporcione la red ante una determinada entrada.

En el aprendizaje por refuerzo la función del supervisor se reduce a indicar mediante una señal de refuerzo si la salida obtenida en la red se ajusta a la deseada (éxito = +1 o fracaso = -1), y en función de ello se ajustan los pesos basándose en un mecanismo de probabilidades. Se podría decir que en este tipo de aprendizaje la función del supervisor se asemeja más a la de un crítico (que opina sobre la respuesta de la red) que a la de un maestro (que indica a la red la respuesta concreta que debe generar), como ocurría en el caso de supervisión por corrección del error.

### **3) Aprendizaje estocástico.**

Consiste básicamente en realizar cambios aleatorios en los valores de los pesos de las conexiones de la red y evaluar su efecto a partir del objetivo deseado y de distribuciones de probabilidad.

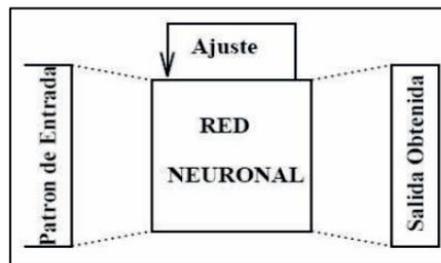
En el aprendizaje estocástico se suele hacer una analogía en términos termodinámicos, asociando a la red neuronal con un sólido físico que tiene cierto estado energético. En el caso de la red la energía de la misma representaría el grado de estabilidad de la red, de tal forma que el estado de mínima energía correspondería a una situación en la que los pesos de las conexiones consiguen que su funcionamiento sea el que más se ajusta al objetivo deseado.

Según lo anterior el aprendizaje consistiría en realizar un cambio aleatorio de los valores de los pesos y determinar la energía de la red (habitualmente la función energía es una función de Liapunov). Si la energía es menor después del cambio es decir si el comportamiento de la red se acerca al deseado se acepta el cambio, si por el contrario la energía no es menor se aceptaría el cambio en función de una determinada y preestablecida distribución de probabilidades.

#### **3.5.2. Aprendizaje no Supervisado**

No requieren de influencia externa para ajustar los pesos de las conexiones entre sus neuronas. Figura 24 La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta, son capaces de autoorganizarse [47]. .

En este tipo de aprendizaje se presenta a la red los valores de entrada, pero no se conoce a la salida los valores que debería generar. La RNA no recibe ninguna información del entorno que le permite determinar si la salida con respecto a los valores de entradas es correcta o no.



**Figura 23. Estructura de un aprendizaje no supervisado**

Estos sistemas de aprendizaje no supervisado no requieren de un vector de salidas deseadas y por tanto no se realizan comparaciones entre las salidas reales y salidas esperadas. El algoritmo de entrenamiento modifica los pesos de la red de forma que produzca vectores de salida consistentes, este proceso de entrenamiento extrae las propiedades estadísticas del conjunto de vectores de entrenamiento y agrupa en clases los vectores similares [41]. Estas redes deben encontrar las características, regularidades correlaciones o categorías que se pueden establecer entre los datos de la entrada. Pero ¿qué genera la red en la salida? Existen varias posibilidades en cuanto a interpretación:

- La salida representa el grado de familiaridad o similitud entre la información de entrada y las informaciones mostradas con anterioridad.
- Clusterización o establecimiento de categorías, indicando la red a la salida a qué categoría pertenece la información de entrada, siendo la propia red la que debe establecer las correlaciones oportunas.
- Codificación de los datos de entrada, generando a la salida una versión codificada con menos bits, pero manteniendo la información relevante de los datos.
- Mapeo de características, obteniéndose una disposición geométrica que representa un mapa topográfico de las características de los datos de entrada.

En cuanto a los algoritmos de aprendizaje no supervisado, en general se suelen considerar dos tipos, que dan lugar a los siguientes aprendizajes:

- 1) Aprendizaje hebbiano.
- 2) Aprendizaje competitivo y comparativo.

### **1) Aprendizaje hebbiano.**

Esta regla de aprendizaje es la base de muchas otras la cual pretende medir la familiaridad o extraer características de los datos de entrada. El fundamento es una suposición bastante simple: si dos neuronas  $N_i$  y  $N_j$  toman el mismo estado simultáneamente (ambas activas o ambas inactivas), el peso de la conexión entre ambas se incrementa.

Las entradas y salidas permitidas a la neurona son:  $\{-1, 1\}$  o  $\{0, 1\}$  (neuronas binarias). Esto puede explicarse porque la regla de aprendizaje de Hebb se originó a partir de la neurona biológica clásica, que solamente puede tener dos estados: activa o inactiva.

### **2) Aprendizaje competitivo y comparativo.**

Se orienta a la clusterización o clasificación de los datos de entrada. Como característica principal del aprendizaje competitivo se puede decir que si un patrón nuevo se determina que pertenece a una clase reconocida previamente, entonces la inclusión de este nuevo patrón a esta clase matizará la representación de la misma. Si el patrón de entrada se determinó que no pertenece a ninguna de las clases reconocidas anteriormente entonces la estructura y los pesos de la red neuronal serán ajustados para reconocer la nueva clase.

## **3.6. Aplicaciones de la Red Neuronal**

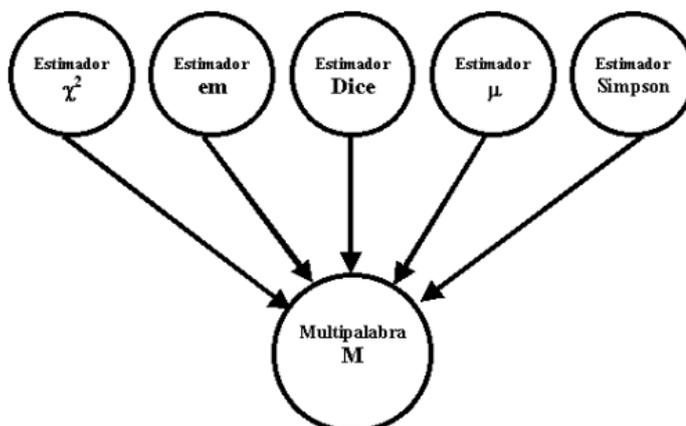
Las redes neuronales en general han sido propuestas en numerosas ocasiones como instrumentos útiles para la Recuperación de Información y también para la clasificación automática. Coincidiendo con el auge general en todos los campos de las redes neuronales especialmente en los primeros años 90, encontramos varias aplicaciones en el campo que nos interesa [41,42].

➤ **Aplicación de redes neuronales y redes bayesianas en la detección de multipalabras para tareas IR**

En esta aplicación se compara el uso de dos métodos distintos para detectar si una pareja de términos son o no multipalabras. Por un lado se usa una red neuronal para clasificar dichos bigramas, y por otro una red bayesiana para obtener la confianza en que los bigramas sean multipalabras. La clasificación está basada en diferentes estimadora, actualmente disponible en la literatura usada como entradas a las dos redes. El resultado obtenido en esta clasificación ha sido usado en tareas de recuperación de información. [43].

Los experimentos demuestran que los dos métodos mejoran la precisión alcanzada por un sistema IR y entre ellos es la red bayesiana la que mejores resultados ofrece.

En la Figura 25 puede apreciarse el aspecto de la sencilla red bayesiana que hemos utilizado en nuestros experimentos. Existirá una de estas redes bayesianas para cada candidata a multipalabras que consideremos.



**Figura 24. Red bayesiana para el reconocimiento de multipalabras**

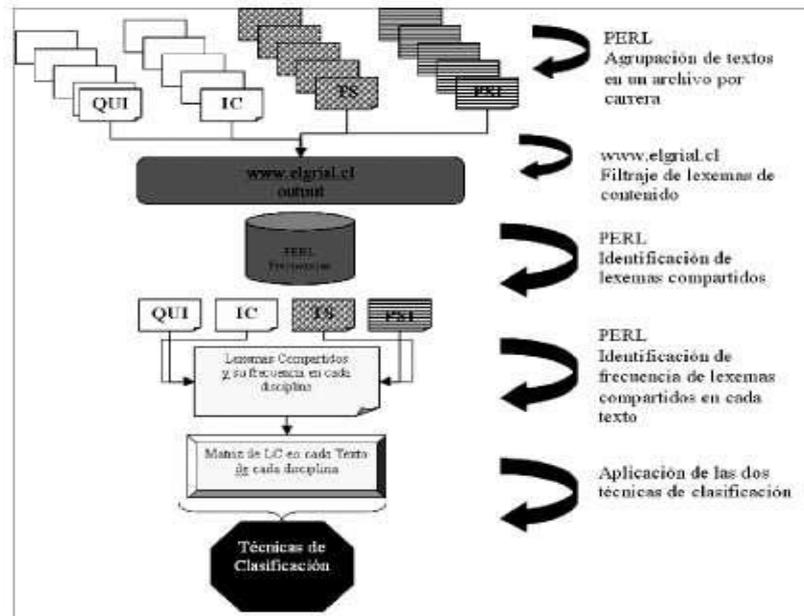
Para poder entrenar y evaluar las redes (neuronal y bayesiana), se construyeron dos listas de ejemplos. Una primera formada por multipalabras y otra por bigramas escogidos aleatoriamente. Cada ejemplo corresponde a una pareja de términos junto con los valores alcanzados por tal pareja para cada estimador.

Para obtener la lista de multipalabras hemos recurrido a WordNet (Miller, 1995), una base de datos léxica rica en multipalabras, tanto endógenas como exógenas. Sin embargo, tan sólo las exógenas son objeto de nuestro estudio.

➤ **Clasificación de textos académicos en función de su contenido léxico-semántico**

Esta investigación es clasificar utilizando y comparando dos métodos de categorización automática, los textos académicos incluidos en el Corpus PUCV-2006 perteneciente al trabajo realizado en el proyecto Fondecyt 1060440. Estos métodos están basados en los lexemas de contenido semántico compartidos en el corpus de textos académicos usados en cuatro carreras profesionales de la Pontificia Universidad Católica de Valparaíso, Chile. El corpus PUCV-2006 actualmente está conformado por 652 textos, los que en cantidad total de palabras alcanza a 96.288.874. Para los propósitos de esta investigación, utilizamos una muestra de 216 textos (30.886.081 palabras) divididos en cuatro áreas disciplinares: 26 usados en Ingeniería en Construcción, 31 en Química, 64 en Trabajo Social y 95 en Psicología. Los métodos de clasificación a comparar en esta investigación son Bayes Ingenuo y Máquina de Soporte de Vectores, ambos métodos permiten identificar un pequeño grupo de lexemas compartidos, que una vez pesados estadísticamente sirven para clasificar un nuevo texto en alguna de las cuatro áreas disciplinares. Los resultados nos permiten establecer que la Máquina de Soporte de Vectores clasifica más eficientemente los textos académicos, con altos valores de precisión y exhaustividad. Con este método podemos identificar automáticamente el dominio disciplinar de un nuevo texto académico en consulta con un alto porcentaje de exactitud (93,9%). Proyectamos usar este método como parte de un análisis multidimensional más acabado del Corpus PUCV-2006. [44].

Los procedimientos metodológicos realizados en esta investigación pueden ser agrupados en dos grandes etapas: preprocesamiento de los textos y aplicación de cada técnica de clasificación Figura 26.



**Figura 25. Preprocesamiento de los textos de la muestra para aplicación de las técnicas de clasificación.**

➤ **Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática**

En este trabajo se propone el modelo de Redes Neuronales Artificiales con aprendizaje supervisado: Perceptrón Multicapa, utilizando como criterio de clasificación el área disciplinar y la caracterización de los textos basada en distribución de frecuencias de las categorías morfo-sintácticas. [45].

Redes Neuronales Artificiales: El Perceptrón Multicapa

*El modelo*

En esta aplicación se consideraron 60 textos de cada una de las disciplinas consideradas. Cada una de estas muestras fue dividida aleatoriamente en dos submuestras de igual tamaño de modo de utilizar una de ellas en la fase de entrenamiento de la red y la otra en la etapa de validación

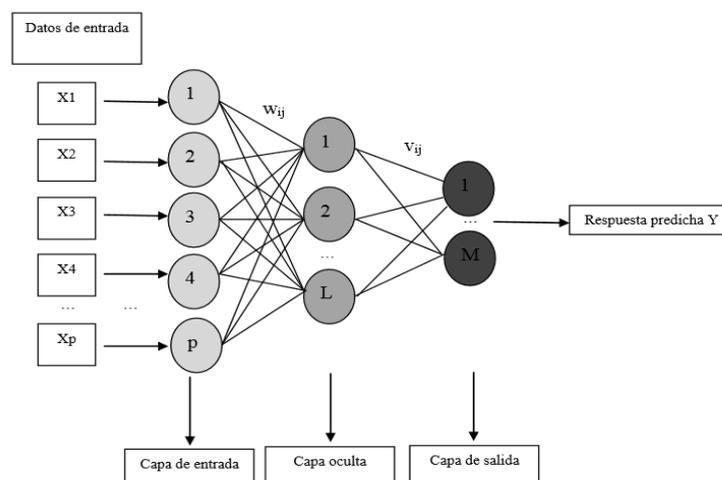
*Arquitectura*

Un Perceptrón multicapa está compuesto por una capa de entrada una capa de salida y una o más capas ocultas, aunque se ha demostrado que para la mayoría de problemas

bastará con una sola capa oculta. En la figura 27 podemos observar un Perceptrón típico formado por una capa de entrada con P neuronas, una capa oculta con L neuronas y una de salida con M neuronas. En este tipo de arquitectura las conexiones entre neuronas son siempre hacia delante es decir las conexiones van desde las neuronas de una determinada capa hacia las neuronas de la siguiente capa no hay conexiones laterales ni conexiones hacia atrás.

En dicho diagrama  $w_{ij}$  representa el peso de conexión entre la neurona de entrada  $i$  y la neurona oculta  $j$ , y  $v_{kj}$  es el peso de conexión entre la neurona oculta  $j$  y la neurona de salida  $k$ .

En esta aplicación las P neuronas de la capa de entrada corresponden a las proporciones de las P categorías morfológicas consideradas y la capa de salida estará constituida por las 3 neuronas que corresponden a las áreas disciplinares.



**Figura 26. Perceptrón multicapa**

*Análisis preliminar.*

La primera comparación que se realiza como ya se mencionó al describir la muestra es la del número de palabras por texto. La misma se lleva a cabo mediante el test no paramétrico de Kruskal Wallis, arrojando una probabilidad asociada  $p=0.16$ , evidenciando que no existen diferencias significativas entre los corpus respecto al tamaño de los textos.

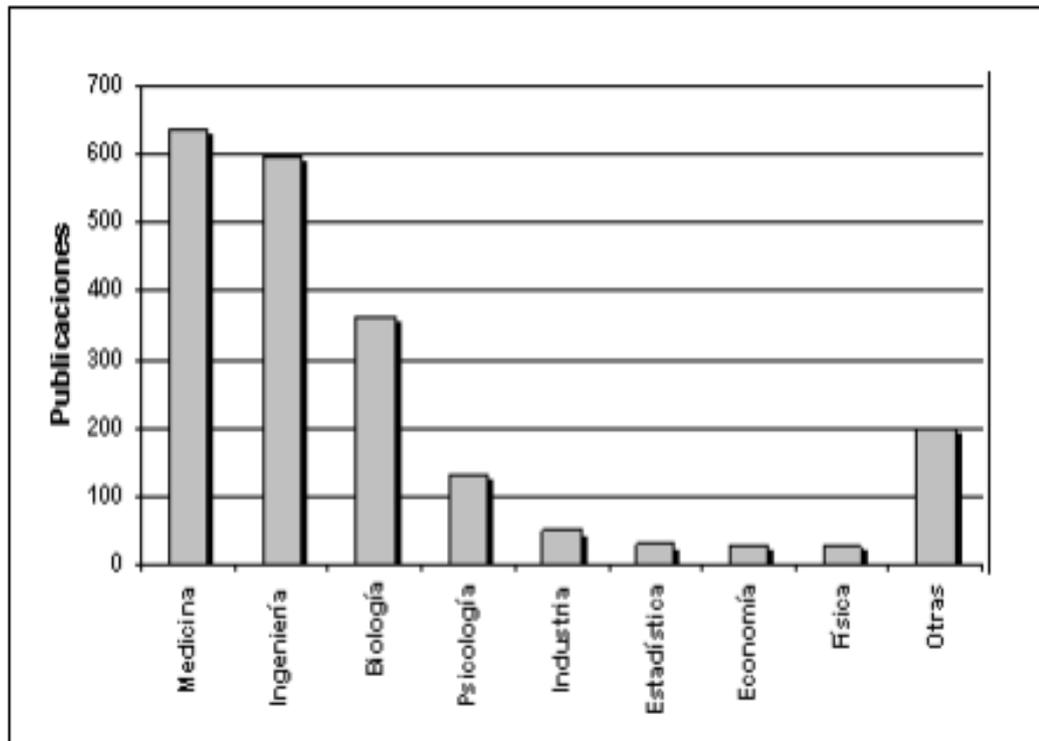
Comparaciones similares entre los corpus se llevan a cabo para las restantes variables hallando diferencias significativas ( $p < 0.05$ ) para el número de clíticos y de adverbios en los documentos analizados Figura 28. El número de clíticos es mayor en los textos de biometría y el número de adverbios es superior en los textos de filosofía.

Número promedio de:	BIOMETRIA	FILOSOFIA	LINGÜÍSTICA INFORMÁTICA	Valor de p
adjetivos	17,9	21,3	11,1	0.0031
adverbios	2,9	5,9	2,33	0.0007
Clíticos	4,1	2,7	2,44	0.0072
copulativos	4,7	6,0	4,0	0.0122
determinantes	26,8	32,4	20,9	0.0031
Nombres	44,6	45,0	30,2	0.0010
preposición	30,0	29,7	21,5	0.0077
Verbos	16,1	18,4	24,0	0.2592
Otro	18,8	21,4	16,7	0.6324
<b>TOTAL PALABRAS</b>	165,8	182,9	155,1	0.1664

**Figura 27. Comparación mediante test de Kruskal Wallis**

En la figura 29 se muestra las disciplinas que mayormente se está utilizando en el campo de aplicación de redes neuronales las cuales las más frecuentes son: medicina (637

registros), ingeniería (597 registros), biología (362 registros) y psicología (132 registros)



**Figura 28. Áreas de aplicación más recurrentes (2001)**

Hemos podido observar en la figura 29 que el campo de aplicación mayoritario es la medicina ingeniería con importantes trabajos en el campo de la energía nuclear y la biología con la identificación de cadenas de ADN, en el campo de la Psicología en el área de Evaluación, Personalidad y Tratamiento los autores se interesan principalmente por el diagnóstico de trastornos mentales.

## e. Materiales y Métodos

A partir del estudio de diferentes bibliografías software y teniendo en cuenta las características del trabajo a desarrollar, la implementación del sistema inteligente para clasificar documentos ya digitalizados se realizó a través del lenguaje de programación Java, donde implementaremos todos los métodos necesarios capaces de realizar la clasificación de documentos utilizando redes neuronales, MySQL para gestionar desde ahí todos los datos necesarios para nuestra aplicación y teniendo en cuenta las características del trabajo a desarrollar se escogió Weka (Waikato Environment for Knowledge Analysis) para el desarrollo de nuestra aplicación e interacción con Java, ya que es un software de código abierto que posee una colección de algoritmos de aprendizaje automático para la extracción de datos y el lenguaje R para las distintas pruebas de clasificación de los documentos mostrando resultados estadísticos de asertividad y error de clasificación .

Con respecto a los métodos que permitieron llevar a cabo la investigación fueron:

**Método Científico:** en este método se trabajó con investigación científica referente a consultas de artículos, tesis doctorales, sitios especializados y oficiales de la temática a buscar lo que permitió llevar a precisar el estado de arte, la elaboración de conceptos (teorías) que permiten acercarnos al entendimiento del problemas logrando una investigación objetiva, ordenada pero sobre todo alcanzable de acuerdo a los parámetros propuestos en la realización del trabajo(*ver sección Revisión de Literatura*).

Una vez terminado todo lo referente a consultas científicas se procedió a la organización de toda la información referente a la clasificación automática de ahí parte el método inductivo-deductivo.

- **Método inductivo:** en este método permitió determinar a partir de los estudios de casos particulares llegar a la conclusión que era necesario un clasificador automático de documentos que facilite al usuario la búsqueda de documentos.
- **Método deductivo:** permitió establecer un vínculo de unión entre teoría y observación permitiendo deducir a partir de la teoría los problemas que están

ocasionando una clasificación manual de documentos, dando solución a este problema con el desarrollo de un sistema inteligente de clasificación automática.

Así mismo, es fundamental apoyarse en técnicas que permitirán extraer información, entre las utilizadas se mencionan a continuación:

**Lectura Científica**, proporcionó el conocimiento y comprensión metódica y secuencial de los temas que constituyen la sustentación del proyecto.

**Observación Directa**, permitió palpar el problema de forma personal de cómo se esa llevando un clasificación manual de documentos y cuáles son sus consecuencias.

Como parte del estudio para llegar a culminar con éxito nuestro trabajo de planteo la siguiente metodología basada en una red neuronal:

### **Metodología de investigación.**

La metodología emplea para el desarrollo de nuestro trabajo es la metodología de la red neuronal que está comprendida por varias fases:

**Fase 1. Análisis:** en esta fase se realiza una revisión de todos textos académicos, resúmenes de trabajos presentados a congresos y revistas, para llegar a tener una mejor claridad del tema a investigar y del corpus de estudio para nuestro tema. Luego del análisis de los posibles corpus de datos a trabajar se llegó a la conclusión que la muestra a estudiar va a estar comprendido por un corpus de 5000 documentos del repositorio de Red de Revistas Científicas de América Latina y el Caribe, España y Portugal de la Universidad autónoma de México, la cual va a permitir con este corpus demostrar la clasificación automática de textos.

**Fase 2. Diseño:** en esta fase se elabora el diseño de la red sus entradas y salidas de la red, una vez llegado a la conclusión del repositorio a utilizar es necesario determinar cuáles van a ser nuestras entradas de estudio para la red, estas entradas estarán representadas por el título, descripción y palabras claves que inicialmente extraeremos por un método que se diseñó para obtener los metadatos de los libros, luego cada texto de entrada serán analizadas por un lematizador extrayendo los lemas de cada palabra

para quedarnos con su raíz, tras aplicar el proceso de lematización a la colección se tienen los documentos y todos sus términos correspondientes en lexemas que corresponden a la raíz de los términos procesados, este resultado será analizado por la red a fin de determinar su categoría correspondiente, las salidas están conformadas por 14 categorías las cuales son : física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso. Además para la asignación de la respectiva categorización se lo hizo a través de un vocabulario creado a base del Sistema de Clasificación Dewey, cuantificando la relevancia de un término para describir a una categoría.

**Fase 3. Entrenamiento :** para el entrenamiento de la red existen dos tipos de entrenamiento supervisado y no supervisado, como primer paso se seleccionó el tipo de entrenamiento el cual se escogió el supervisado, para evaluar la clasificación de los 5000 documentos se extrajo dos conjuntos de entrenamiento, el primer entrenamiento se lo hizo con un grupo conformado por 1000 documentos trabajando con un vocabulario medio indicando como resultado que los documentos están asignándose correctamente a su categoría y el segundo grupo con 3000 documentos, en este grupo se trabajó con un mayor número de información lo que hace necesario incrementar nuevas palabras al vocabulario, mostrando resultados satisfactorios con respecto a la clasificación, todo este entrenamiento se lo realizó en RWEKA ya que es un software que ayuda a evaluar la calidad de clasificación, tanto el primer y segundo grupo de entrenamiento sus resultados están basados al tamaño del vocabulario indicando que mejorara en la medida que el vocabulario aumente.

**Fase 4. Pruebas:** como fase final evaluamos el desempeño del categorizador propuesto usando documentos de la biblioteca del Área de la Energías, las Industrias y los Recursos Naturales no Renovables de la Universidad Nacional de Loja, con el fin de estudiar la validez de los resultados demostrando que cada documento ingresado sea asignado a su categoría correspondiente, además de trabajar con libros en inglés.

## **f. Resultados**

El objetivo del TT consiste en la construcción de un sistema inteligente para la clasificación de documentos ya digitalizados aplicando redes neuronales. Para lograr el desarrollo de dicho sistema se ha pasado por varias etapas que son descritas dentro de los objetivos específicos a lo largo de este capítulo, donde se muestran los resultados obtenidos a partir de las diversas pruebas realizadas en base a los corpus de documentos.

### **1. Análisis**

#### **1.1. Revisión sobre Redes neuronales supervisadas e Inteligencia artificial**

Es la primera parte y la fundamental de este proyecto ya que a partir de ella se adquieren los conocimientos necesarios para obtener los mejores resultados en el trabajo. El primer paso consistió en la lectura de publicaciones científicas en el tema en cuestión para poder conocer y entender el estado del arte en los sistemas de clasificación automática de documentos.

#### **1.2. Análisis de la selección del corpus de datos a utilizar**

En las últimas décadas, las dinámicas del trabajo de categorización y clasificación de documentos han cambiado sustancialmente a causa de los avances de las tecnologías informáticas y su aporte para la disciplina.

En el presente trabajo me centro en algunas reflexiones valiosas para el trabajo de categorización de documentos digitales: primero la relevancia de un corpus con una descripción técnica del área y lo segundo que contenga los metadatos necesarios para la categorización de los mismos y así lograr el objetivo del presente trabajo.

A continuación se muestra los 4 repositorios de estudio para su selección:

### A. Repositorio de Ciencia - Science Library

En esta web realizada por un grupo de desarrolladores que incorporan libros de interés científico e investigativo, los cuales realizan foros para cualquier comentario o información sobre novedades del repositorio. [46]. Figura 30

Características:

- Es un repositorio libre descarga total o parcialmente.
- Contenido en formato PDF.
- La mayoría de documentos están el lenguaje español.
- El contenido de los metadatos de los documentos es incompleto.
- Los documentos son completos.
- Contiene 154 ejemplares.



### Repositorio de Ciencia - Science Library

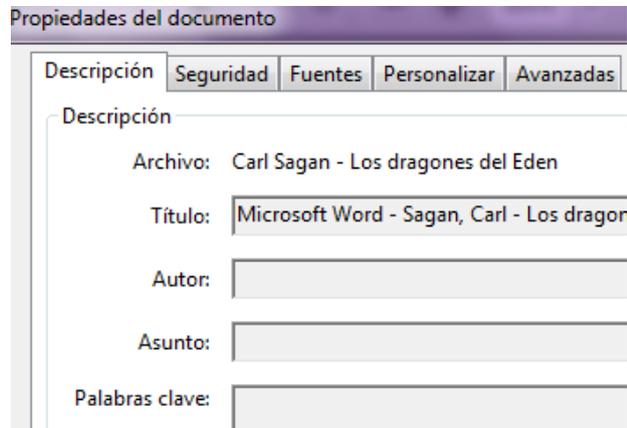
Para cualquier comentario e información sobre novedades del repositorio: [Foro-Board](#)

Para el que le interese descargarse el repositorio completo esta disponible un torrent de 330MB: [bz.otsoa.net - Biblioteca Divulgacion Cientifica v1.0.zip.torrent](#)

Inicio - Home  
06-19-2014 04:26:14

**Figura 29.Repositorio de Ciencia - Science Library**

En la Figura 31 se puede observar de un ejemplo de metadatos de un documento seleccionado del repositorio anterior.



**Figura 30. Metadatos del Repositorio de Ciencia - Science Library**

## **B. Biblioteca Virtual Universal**

La web cuenta con gran número de documentos y libros referentes a materias científicas e investigativas, este repositorio está elaborada por dos bibliotecas: Bibliotecas Rurales Argentinas y Bibliotecas Latinoamericanas, las cuales son sin fines de lucro dan el servicio de ayuda acerca de esta biblioteca, sugiere textos para digitalizar que no se encuentre disponible. [47]. Figura 32

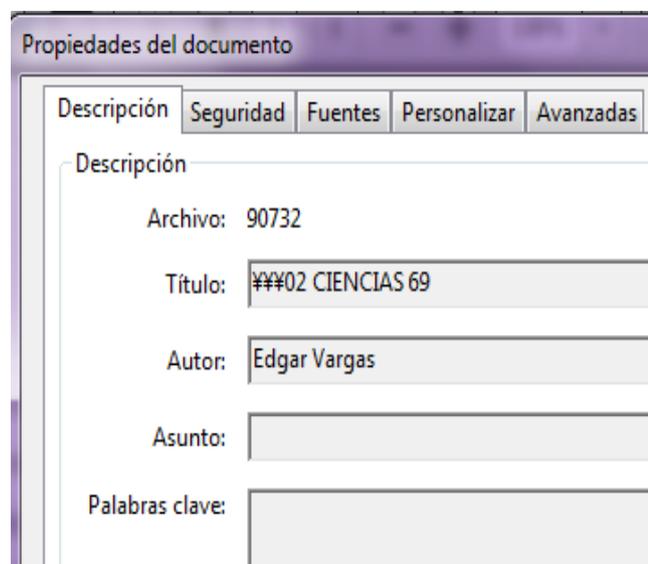
Características:

- Es un buscador de libre acceso.
- El contenido se encuentra en un índice general alfabéticamente ordenado.
- La descarga de los libros es libre de acceso.
- Cuenta con un total 30.015 obras digitalizadas
- No contiene información en la estructura de metadatos



**Figura 31. Biblioteca Virtual Universal**

En la Figura 33 se muestra un ejemplo de los metadatos que contiene este repositorio antes mencionado:



**Figura 32. Metadatos de la Biblioteca Virtual Universal**

### C. Biblioteca Utpl

Esta página web es desarrollada por la Universidad Técnica Particular de Loja, en la plataforma Dspace es de uso libre con acceso a tesis y revistas científicas [48].Figura 34.

Características:

- Es de acceso libre para las descargas de documentos.
- El contenido se encuentra ordenado y con un buscador que permite el fácil acceso de los mismos ya sea por título, autor, fecha, materia.
- El contenido se encuentra con los metadatos.

The screenshot shows the Biblioteca Utpl website interface. At the top, there is a navigation bar with language options (English and Español) and a search bar. Below the search bar, there are five icons representing search criteria: COMUNIDADES, FECHA, AUTOR, TÍTULO, and MATERIA. The main content area is titled 'Búsqueda por Título' and features a search input field with a dropdown menu for letters (0-9 A-Z) and a 'Buscar' button. Below the search bar, there are filters for 'Ordenado por: Título', 'Orden Ascendente', 'Número de registros: 20', and 'Autores Ilimitado', along with an 'Actualizar' button. The results are displayed in a table with three columns: FECHA DE PUBLICACIÓN, TÍTULO, and AUTOR(ES).

FECHA DE PUBLICACIÓN	TÍTULO	AUTOR(ES)
22-jun-2012	DF	-
16-jun-2010	Gestión del Conocimiento: Modelos y Tecnologías Caso: Universidad Técnica Particular de Loja	Rodríguez Morales, Germania; Cueva Carrión, Samanta; Naranjo, Jorge
24-jun-2011	Implementation of Social and Semantic Tools into Open Educational Resources Production	Cueva, Samanta
7-ene-2010	OER, Estándares y Tendencias	Cueva, Samanta; Rodríguez, Germania
18-oct-2010	Recursos Educativos Abiertos - Licencias y prospectiva; Caso Universidad Técnica Particular de Loja UTPPL	Cueva Carrión, Samanta; Pacheco Montoya, Patricia; Rodríguez Morales, Germania
4-may-2010	La secretaria ejecutiva del honorable consejo provincial de Pichincha	Yanchaguano, Mónica
21-abr-2009	Social Network and OER 'S	Cueva Carrión, Samanta; Rodríguez, Germania

Figura 33.Biblioteca UTPPL

En la Figura 35 se da a conocer de los metadatos del repositorio.

<b>Título:</b>	Gestión del Conocimiento: Modelo
<b>Autor:</b>	Rodríguez Morales, Germania Cueva Carrión, Samanta Naranjo, Jorge
<b>Palabras clave:</b>	Gestión del Conocimiento Modelos de GC Tecnologías de GC Metodologías de GC Herramientas Web 2.0
<b>Fecha de publicación:</b>	<b>16-jun-2010</b>
<b>Citación:</b>	Volumen II;312580 / 10
<b>Resumen:</b>	La Gestión del Conocimiento es un fundamental para las actividades d permita aprovecharlo efectivamente de las Instituciones de Educación n Conocimiento (GC) para la Universi Distancia (MaD), que inició con el ar tecnologías que los soportan; conti del Modelo estructurado en procesi tecnológica principal para su soport procesos.
<b>URI:</b>	<a href="http://hdl.handle.net/12345678">http://hdl.handle.net/12345678</a>
<b>ISBN:</b>	978-989-96247-3-3
<b>Aparece en las colecciones:</b>	OER

**Figura 34. Metadatos Biblioteca UTPL**

#### **D. Red de Revistas Científicas de América Latina y el Caribe, España y Portugal**

Este repositorio de documentos está desarrollada por la Universidad Autónoma del Estado de México con acceso libre, cuenta con un buscador que permite el acceso de los de documentos y revistas, dicho contenido está organizado por: Artículos, Autores, Revistas y Disciplinas [49]. Figura 36

Características:

- Es de acceso libre para las descargas de los documentos.
- Tiene un buscador que permite el acceso de los documentos a través de título, autor, materia entre otros.
- Contiene un repositorio de 912 Revistas Científicas, 26425 Fascículos, 336437 Artículos a texto Completo
- Los metadatos de los contenido en su mayoría son completos

**Figura 35. Red de Revistas Científicas de América Latina y el Caribe, España y Portugal**

En la figura 37 se presenta un ejemplo de metadatos de este repositorio

```

<meta name="keywords" content="Toma de Decisiones, Problemas Multicriteri
<meta name="citation_abstract_html_url" content="http://www.redalyc.org/1
<meta name="citation_author" content="JUAN CARLOS OSORIO G&Oacute;MEZ" />
<meta name="citation_author" content="JUAN PABLO OREJUELA CABRERA" />
<meta name="citation_publication_date" content="2008" />
<meta name="citation_journal_title" content="Scientia Et Technica" />
<meta name="citation_issn" content="0122-1701" />
<meta name="citation_volume" content="XIV" />
<meta name="citation_issue" content="39" />
<meta name="citation_title" content="EL PROCESO DE AN&Aacute;LISIS JER&Aa

```

**Figura 36. Metadatos Red de Revistas Científicas de América Latina y el Caribe, España y Portugal**

## DISCUSIÓN DEL CORPUS DE DATOS

En la siguiente tabla se muestra un detalle de comparación de los 4 repositorios de estudio, sus ventajas y desventajas de cada uno de ellos con la finalidad de seleccionar el más idóneo para realizar la clasificación de los documentos.

**TABLA II REPOSITORIOS DE ESTUDIO**

<b>Repositorios</b>	<b>Cantidad documentos</b>	<b>Tipos de documentos</b>	<b>Ventajas</b>	<b>Desventajas</b>
<b>Repositorio de Ciencia - Science Library [46]</b>	Poco (154 documentos)	Cuenta con libros de divulgación científica y artículos	Tiene un enlace de descarga total de documentos digitalizados	- Es un repositorio muy pequeño de documentos  -Metadatos incompletos
<b>Biblioteca Virtual Universal[47]</b>	Alto (30.015 doc.)	- Cuenta con obras clásicas y estudios científicos.  - Trabajos de divulgación científica y de investigación	Contiene un índice medio de documentos a requerir	La mayoría de los metadatos son incompletos
<b>Biblioteca Utpl[48]</b>	No presenta la cantidad de documentos existentes.	Tesis doctorales, revistas y libros	Solo trabaja con documentos de su propia biblioteca	La estructura de sus metadatos tienen muy poca información
<b>Red de Revistas Científicas de América Latina y el Caribe, España y Portugal [49]</b>	Alto(338 816 documentos)	Contiene artículos científicos a texto completos, revistas científicas y fascículos	Existe una amplia variedad de documentos digitalizados	Dos de cada 20 documentos no contienen metadatos completos

Luego de haber analizados sus ventajas y desventajas de cada uno de ellos se escogió el repositorio que más se acopla con nuestro tema de investigación llegando a la conclusión de que el mejor corpus a trabajar es el repositorio Red de Revistas Científicas de América Latina y el Caribe, España y Portugal en el que se lo encuentra en el siguiente enlace <http://www.redalyc.org/home.oa>.

### **Red de Revistas Científicas de América Latina y el Caribe, España y Portugal**

Como se ha expuesto anteriormente una de las razones para el uso de este repositorio es debido a que su conjunto de datos son amplios y sus metadatos en su mayoría tiene una estructura completa como son autor, título, palabras claves publicación, url entre otros los cuales estos serán de gran ayuda para el desarrollo de nuestro proyecto fin de carrera a diferencia de los otros repositorios sus metadatos son incompletos en su mayoría.

Dado que el repositorio tiene un alta cantidad de documentos digitalizados como son 338 816, y compuesta por una variedad de documentos como son: artículos científicos a texto completos, revistas científicas y fascículos esto nos facilitara a la red neuronal obtener un numero de 5000 documentos para hacer el entrenamiento a la red neuronal y de esta forma quede lista la aplicación para que clasifique cualquier clase de documentos.

#### **1.3. Identificación de los metadatos a utilizar**

Los metadatos son datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar. En otras palabras los metadatos son datos sobre los datos es por ello que luego del análisis de los repositorios y la identificación de sus metadatos que tienen cada uno se puedo identificar que los metadatos a ser utilizados son los siguientes como se muestra en la tabla III ya que corresponden metadatos básicos de un libro.

**TABLA III METADATOS PARA LA BASE DE DATOS**

Autores
Título
Descripción
Fecha de publicación
Palabras Clave
Lenguaje
Url

Los metadatos de título, palabras claves y descripción servirán como entrada para la red y de estas se determinara cuáles son las salidas o a que categoría pertenece el documento.

#### **1.4. Identificación de las técnicas de clasificación de documentos**

Por lo general cuando se habla de clasificación automática se distingue entre dos escenarios diferentes que obviamente requieren soluciones distintas. Estos escenarios reciben diversos nombres pero básicamente consisten en lo siguiente: de un lado una situación en la que se parte de una serie de clases o categorías conceptuales prediseñadas a priori, y en la que labor del clasificador (manual o automático) es asignar cada documento a la clase o categoría que le corresponda a eso se llama clasificación supervisada.

En el segundo escenario posible no hay categorías previas ni esquemas o cuadros de clasificación establecidos a priori. Los documentos se agrupan en función de su contenido, de alguna manera podemos decir que se auto organizan. Es lo que se conoce como clasificación (automática) no supervisada o clustering no supervisada porque se efectúa de forma totalmente automática, sin supervisión o sentencia manual.

Al conocer los dos conceptos de técnicas se llegó a concluir que nuestro trabajo va a estar basado en una clasificación supervisada, esta técnica parte de una serie de clases o categorías prediseñadas a priori, en las cuales hay que colocar a cada uno de los documentos a su categoría correspondiente.

## 2. Diseño

Para el diseño de los algoritmos de clasificación de documentos se lo realizó en la herramienta Enterprise Architect [50]. , por lo que es una herramienta completa y versátil que permite el modelado del proyecto desde sus primeras fases de análisis hasta las de pruebas.

### 2.1. Modelo de Dominio

En la figura 38. Se visualiza las relaciones que existen entre las distintas clases del sistema: libro, glosario, categoría, y administrador.

La clase **libro** esta clase tiene todos los atributos correspondiente a la descripción de un libro.

La clase **categoría** permite asignar el libro a una categoría.

La clase **glosario** contiene todos los glosarios de cada categoría.

La clase **administrador** contiene todos los usuarios autorizados para usar la aplicación además de validar el id de cada usuario

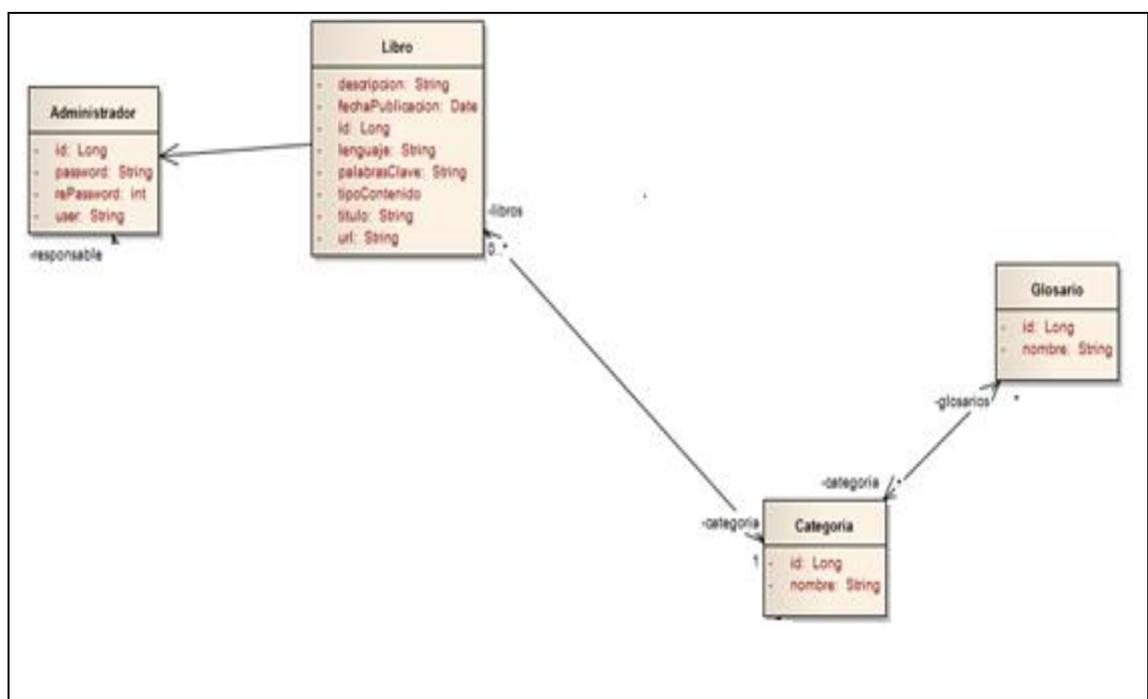


Figura 37. Modelo de Clases

## 2.2. Modelado de la base de datos

Para realizar el diseño de la base de datos, se determinaron las posibles estructuras de las tablas que requiere el sistema, en la Figura 39 muestra el diagrama relacional (modelo de datos) en este diagrama se observa 5 tablas que guardan relación entre sí indica las estructuras de los archivos de la base de datos y la relación que existen entre cada uno de ellos, además se puede apreciar la cardinalidad entre las tablas, por ejemplo la tabla libro posee un relación de muchos a uno con la tabla de administrador, ya que muchos libros puede tener varios administradores, así como la relación entre la tabla de libro con la tabla categoría en donde muchos libros pueden pertenecer a una sola categoría, otra relación que se puede observar es la relación que existen entre la tabla categoría con la tabla categoríaglosario que es de uno a muchos ya que una categoría puede tener muchos glosarios es decir en la tabla de categoríaglosario tiene la unión de glosario y categoría, y a su vez la tabla glosario igual tiene una relación con la tabla de categoríaglosario de uno a muchos es decir un glosario tiene varias categorías formando una tabla intermedia de categoríaglosario. Para el diseño de esta base de datos de implemento como manejador de base de datos My SQL 5.5.

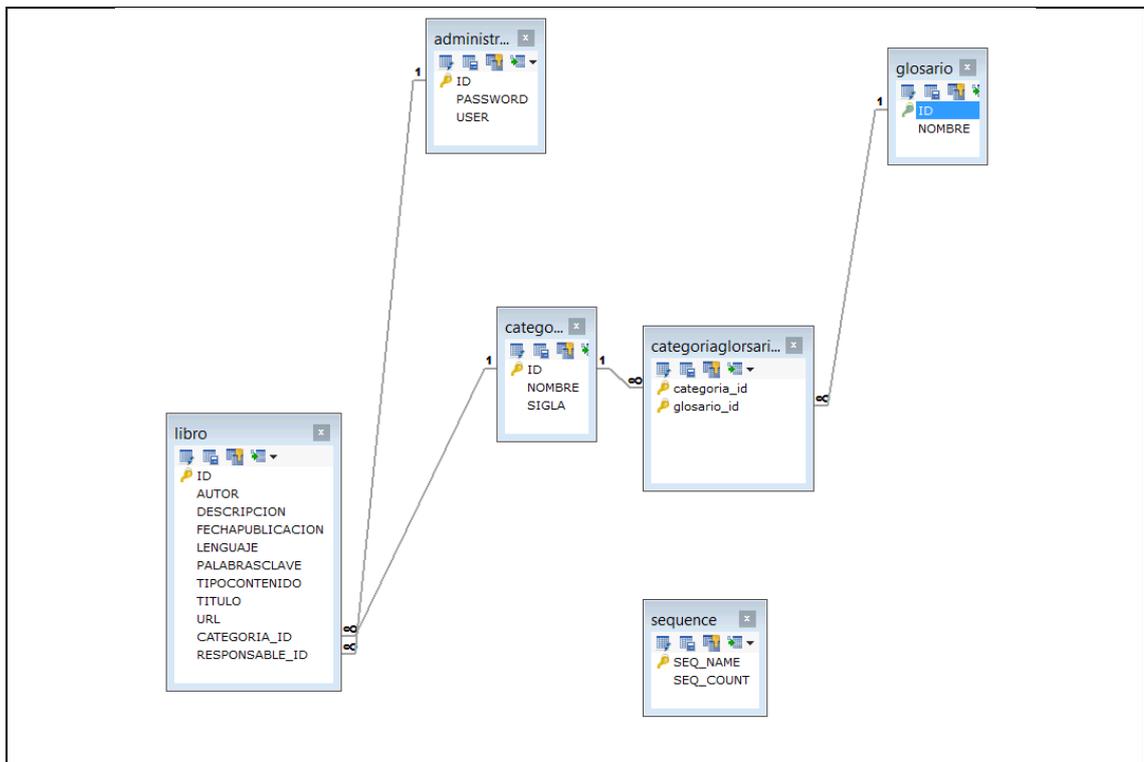


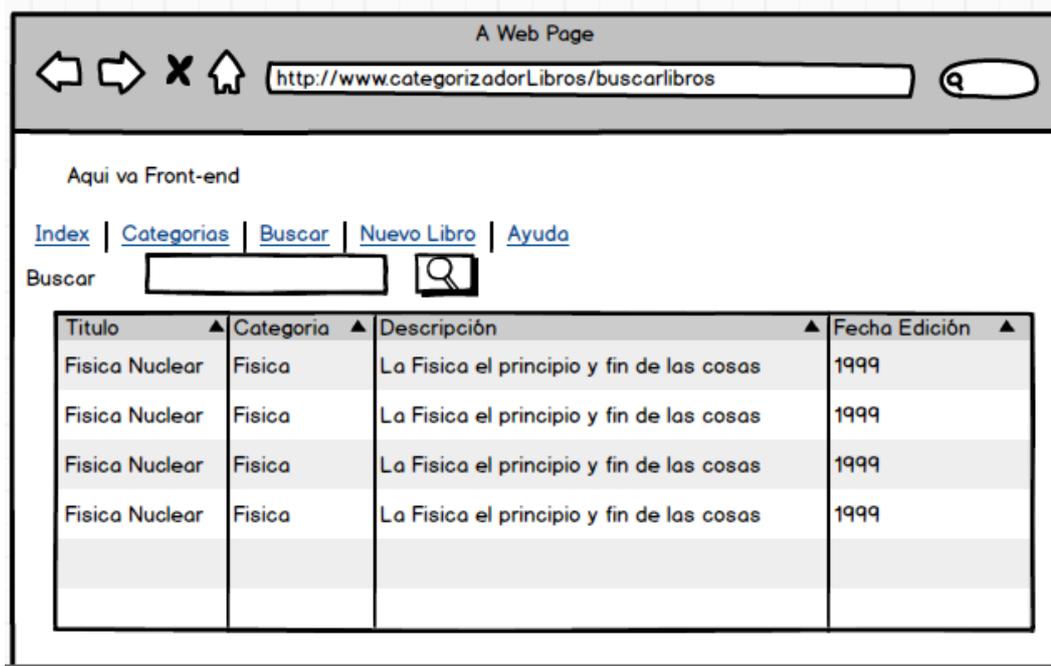
Figura 38. Modelo de la base de datos

### 2.3. Diseño de los prototipos de pantallas

En el diseño de las pantallas se elaboraron con Balsamiq Mockup ya que es una herramienta sencilla, intuitiva y útil para el desarrollo de prototipos de aplicaciones. A continuación se muestra los diseños de forma básica que van a ser utilizadas para la clasificación de los documentos y su representación.

- **Buscar Libro**

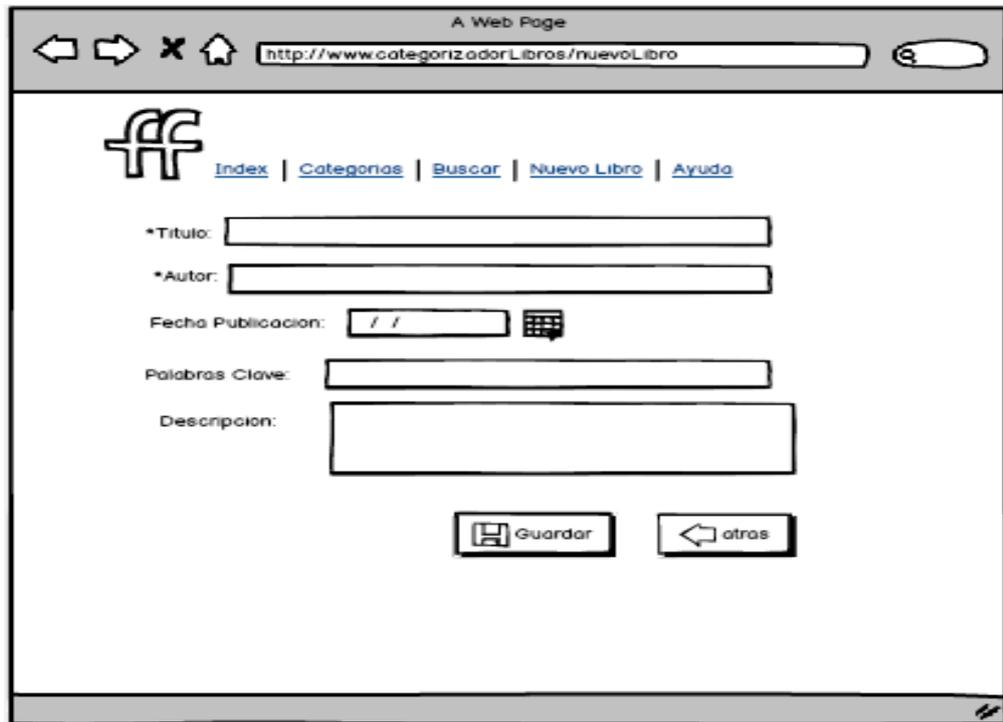
En la Figura 40 muestra un prototipo de ventana que permite buscar algún libro por ciertas características.



**Figura 39. Ventana Buscar Libros**

- **Crear libro**

En la Figura 41 muestra el tipo de ventana que permite el ingresar un libro llenando ciertos datos que se requieren.



**Figura 40. Ventana Crear Libro**

- **Mostrar libro**

En la Figura 42 en este tipo de ventana presenta las características o los datos del libro que se ha pedido mostrar.



**Figura 41. Ventana Mostrar Libro**

## **2.4. Diseño de la arquitectura de la red**

Debido a que se va a construir una red multicapa supervisada de antemano se debe tener en claro cuáles son las salidas de la red, estas salidas están conformadas por 14 categorías las cuales son: física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso, estas categorías se las menciona luego de hacer un estudio y una revisión exhaustiva de cada uno de los documentos que contiene el repositorio se dio a dictaminar las categorías que iba a tener la red, además permitirá la creación de nuevas categorías con la finalidad de que los libros estén acorde a la categoría perteneciente .

Para determinar que un documento pertenezca a esta categoría se tuvo que definir cada una de ellas un vocabulario es decir cada categoría cuenta con subcategorías a las que puede ser asignado un documento. Para ellos usamos el sistema de clasificación denominado Sistema de Clasificación Dewey [51], el cual constituye listas estructuradas de términos (conceptos) que representan de forma unívoca el contenido conceptual de los documentos, además es un sistema que cuantifica la relevancia de un término para describir a una categoría.

### **✓ Sistema de Clasificación Dewey**

El Sistema de Clasificación Decimal Dewey (también llamado CDD) es un sistema de clasificación de bibliotecas. [51]

Fue desarrollado por Melvil Dewey, bibliotecario del Amherst College en Massachusetts en 1876 y desde ese momento ha sido enormemente modificado y ampliado en el transcurso de sus veintitrés principales ediciones que han ocurrido hasta 2012. Durante este tiempo y desde 1894 también se han desarrollado 14 ediciones abreviadas, basadas en la Edición mayor desarrollada generalmente un año antes.

### **Clases Principales**

Las 10 grandes clases que lo conforman son (basadas en la Edición 14 abreviada, del año 2008

000 - Ciencia de los Computadores, Información y Obras Generales.

100 - Filosofía y Psicología.

200 - Religión, Teología.

- 300 - Ciencias Sociales.
- 400 - Lenguas.
- 500 - Ciencias Básicas.
- 600 - Tecnología y Ciencias Aplicadas
- 700 - Artes y recreación.
- 800 - Literatura.
- 900 –Historia y Geografía

Estas categorías principales se subdividen a su vez cada una de ellas en diez clases, en un modelo jerárquico decimal de diversos niveles. Ver Figura 43 El primer nivel (también llamado sumario) comprende estos diez grupos, en segundo sumario estaría formado por cien grupos, diez por cada uno de los diez anteriores. El tercer nivel abriría un abanico de mil posibilidades e incluso podríamos seguir añadiendo más si fueran necesarias ver más en Anexos.

## Segundo Sumario<sup>™</sup> Las cien divisiones

<b>000 Generalidades</b>	<b>500 Ciencias naturales y matemáticas</b>
010 Bibliografía	510 Matemáticas
020 Bibliotecología y ciencias de la información	520 Astronomía y ciencias afines
030 Obras enciclopédicas generales	530 Física
040	540 Química y ciencias afines
050 Publicaciones seriadas generales	550 Ciencias de la tierra
060 Organizaciones generales y museología	560 Paleontología Paleozoología
070 Medios noticiosos, periodismo, publicación	570 Ciencias de la vida Biología
080 Colecciones generales	580 Plantas
090 Manuscritos y libros raros	590 Animales
<b>100 Filosofía y psicología</b>	<b>600 Tecnología (Ciencias aplicadas)</b>
110 Metafísica	610 Ciencias médicas Medicina
120 Epistemología, causalidad, género humano	620 Ingeniería y operaciones afines
130 Fenómenos paranormales	630 Agricultura y tecnologías relacionadas
140 Escuelas filosóficas específicas	640 Economía doméstica y vida familiar
150 Psicología	650 Gerencia y servicios auxiliares
160 Lógica	660 Ingeniería química
170 Ética (Filosofía moral)	670 Manufactura
180 Filosofía antigua, medieval, oriental	680 Manufactura para usos específicos
190 Filosofía moderna occidental	690 Construcción
<b>200 Religión</b>	<b>700 Las artes Bellas artes y artes decorativas</b>
210 Filosofía y teoría de la religión	710 Urbanismo y arte paisajístico
220 La Biblia	720 Arquitectura
230 Cristianismo Teología cristiana	730 Artes plásticas Escultura
240 Moral cristiana y teología piadosa	740 Dibujo y artes decorativas
250 Órdenes cristianos e iglesia local	750 Pintura y pinturas
260 Teología social y eclesiástica	760 Artes gráficas Arte de grabar y grabados
270 Historia del cristianismo y de la iglesia cristiana	770 Fotografía y fotografías
280 Confesiones y sectas cristianas	780 Música
290 Religión comparada y otras religiones	790 Artes recreativas y de la actuación
<b>300 Ciencias sociales</b>	<b>800 Literatura y retórica</b>
310 Colecciones de estadística general	810 Literatura norteamericana en inglés
320 Ciencia política	820 Literaturas inglesa e inglesa antigua
330 Economía	830 Literaturas de lenguas germánicas
340 Derecho	840 Literaturas de lenguas romances
350 Administración pública y ciencia militar	850 Literaturas italiana, rumana, retorromana
360 Problemas y servicios sociales; asociaciones	860 Literaturas española y portuguesa
370 Educación	870 Literaturas itálicas Literatura latina
380 Comercio, comunicaciones, transporte	880 Literaturas helénicas Literatura griega clásica
390 Costumbres, etiqueta, folclor	890 Literaturas de otras lenguas
<b>400 Lenguas</b>	<b>900 Geografía e historia</b>
410 Lingüística	910 Geografía y viajes
420 Inglés e inglés antiguo	920 Biografía, genealogía, insignias
430 Lenguas germánicas Alemán	930 Historia del mundo antiguo hasta ca. 499
440 Lenguas romances Francés	940 Historia general de Europa
450 Italiano, rumano, retorromano	950 Historia general de Asia Lejano Oriente
460 Lenguas española y portuguesa	960 Historia general de África
470 Lenguas itálicas Latín	970 Historia general de América del Norte
480 Lenguas helénicas Griego clásico	980 Historia general de América del Sur
490 Otras lenguas	990

**Figura 42. Grupos principales de Sistema de clasificación Dewey**

Las entradas de la red estarán constituidas por los metadatos de extracción de título, descripción y palabras claves, a partir de este texto de entrada el siguiente paso es llevar a un proceso de lematización por lo tanto procederemos a la reducción a la raíz de las palabras. Esta etapa se denomina lematización o reducción a raíces léxicas (en inglés "stemming"). Almacenamos sólo las raíces de los términos de modo se pueden llegar a reducir su dimensión considerablemente.

## ✓ **STEMMING**

**Stemming** es un método para reducir una palabra a su raíz o (en inglés) a un *stem* o lema. Esta etapa se denomina lematización o reducción a raíces léxicas (en inglés "stemming").

Stemming aumenta el recall que es una medida sobre el número de documentos que se pueden encontrar con una consulta. Por ejemplo una consulta sobre "bibliotecas" también encuentra documentos en los que solo aparezca "bibliotecario" porque el stem de las dos palabras es el mismo ("bibliotec") [52].

Es una técnica de reducción que permite detectar variantes morfológicas de un mismo término por ejemplo palabras como cómputo, computadoras, computable, computación son variantes del término computar, y reemplazarlas por el término raíz o lema.

El uso del stemming o lematización posibilita: a) tener índices de menor tamaño y b) una mayor cantidad de respuestas a una consulta dada, debido a que ahora al aplicarse lematización al corpus y a la consulta se recuperan documentos que tengan todas las variantes morfológicas de los términos contenidos en la consulta. La técnica de stemming permite extraer sufijos y/o prefijos comunes, de tal forma que palabras que literalmente son diferentes, pero tienen una raíz común, pueden ser consideraras como un solo término en base a su raíz. El siguiente ejemplo muestra términos base y algunas de sus posibles variantes:

### **Término base**

Casa  
Poder  
Amable  
Computas

### **Variantes**

casas, casita, casitas  
poderes, poderíos  
amables, amabilidad, amabilidades  
computadora, computado, computable,  
computadoras

Almacenando sólo las raíces de los términos, se puede llegar a reducir su dimensión considerablemente. La reducción de los términos puede realizarse bien durante la indización o bien en la propia búsqueda. La primera variante presenta la ventaja de ser más eficiente y ahorrar espacio.

En nuestro caso hemos optado por aplicar el algoritmo de Snowball Steamer[53], de esta forma tras esta fase obtendremos una lista de términos lematizados representativos de cada documento ver Figura 44. Esta lista la podemos manejar computacionalmente con la metodología propuesta.

El Snowball es un stemmer que sigue un modelo basado en reglas lexicográficas definidas para el idioma Inglés y extendidas para muchos idiomas, en nuestro caso hemos usado las reglas adaptadas al español de <http://snowball.tartarus.org/algorithms/spanish/stemmer.html>

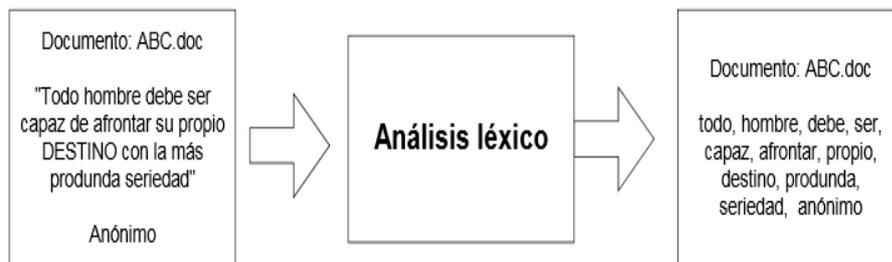
Here is a [sample](#) of [Spanish vocabulary](#), with the stemmed forms that will be [generated](#) with this algorithm.

word	stem	word	stem
che	che	torá	tor
checa	chec	tórax	torax
checar	chec	torcer	torc
checo	chec	toreado	tor
checoslovaquia	checoslovaqui	toreados	tor
chedraoui	chedraoui	toreándolo	tor
chefs	chefs	torear	tor
cheliabinsk	cheliabinsk	toreara	tor
chelo	chel	torearlo	tor
chemical	chemical	toreó	tore
chemicalweek	chemicalweek	torero	torer
chemise	chemis	toreros	torer
chepo	chep	torio	tori
cheque	chequ	tormenta	torment
chequeo	cheque	tormentas	torment
cheques	chequ	tornado	torn
cheraw	cheraw	tornados	torn
chesca	chesc	tornar	torn
chester	chest	tornen	torn
chetumal	chetumal	torneo	torne

**Figura 43. Vocabulario de Stemming**

Luego de trabajar con stemmer se experimentó con documentos los cuales sus resultados fueron aceptables, porque se redujo considerablemente el tamaño del texto de tal forma ayuda a minimizar una gran cantidad de palabras repetitivas o invalidas para nuestro trabajo. De esta forma tras aplicar esta etapa del procesamiento documental logramos abreviar muchas palabras redundantes que no iban a ser útiles para el desarrollo del sistema.

En la siguiente figura 45 muestra un ejemplo de como un analizador léxico trata un flujo de caracteres:



**Figura 44. Diagrama ejemplo de flujo de entrada y transformación realizada**

## APRENDIZAJE DE LA RED NEURONAL

Luego de conocer todos los conceptos básicos de los tipos de aprendizaje (revisión literaria del capítulo de redes neuronales) se eligió el aprendizaje por corrección de error, ya que cuando una red neuronal genera una salida esta es comparada por la que se espera, los pesos de la conexiones se ajustan de acuerdo a esta diferencia, es decir en función al error cometido en la salida y van cambiando hasta lograr que la respuesta que genero la red sea la que se desea. La fórmula para la corrección de pesos es la siguiente:

$$\text{Incr}(w_{ji}) = \beta y_i (d_j - y_j)$$

Donde:

Incr ( $w_{ji}$ ): variación en el peso de la conexión entre las neuronas  $i$  y  $j$

$y_i$ : salida de la neurona  $i$

$d_j$ : valor de la salida deseada para la neurona  $j$

$y_j$ : valor de la salida obtenida para la neurona  $j$

$\beta$ : valor de aprendizaje ( $0 < \beta \leq 1$ ) que regula la velocidad de aprendizaje

La idea principal de este aprendizaje es minimizar el error entre la salida esperada y la que se obtiene.

Este método se describe como se menciona:

- ❖ Inicializar los pesos
- ❖ Presentación del conjunto de entrenamiento
- ❖ Obtención de las salidas para el conjunto de entrenamiento
- ❖ Comparación de las salidas deseadas con las actuales
- ❖ Si se verifica el criterio de finalización ir al siguiente paso, si no ir al paso 2.
- ❖ Fin

El aprendizaje supervisado se caracteriza por conocer la respuesta que debería tener la red frente a una determinada entrada. De esta manera se compara la salida deseada con la salida de la red y si existen discrepancias se ajusta iterativamente los pesos. Así la etapa de aprendizaje tiene por objeto hacer mínimo el error entre la salida brindada por la red y la salida deseada o verdadera.

En la figura 46 muestra la estructura de la red y sus conexiones entre cada de sus entradas y salidas, al conocer que es una red supervisada se debe tener definidos sus salidas las cuales son 14 categorías (física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso). Como se puede observar el número de entradas de la red son 14 las cuales están representadas por los metadatos de los documentos como son:

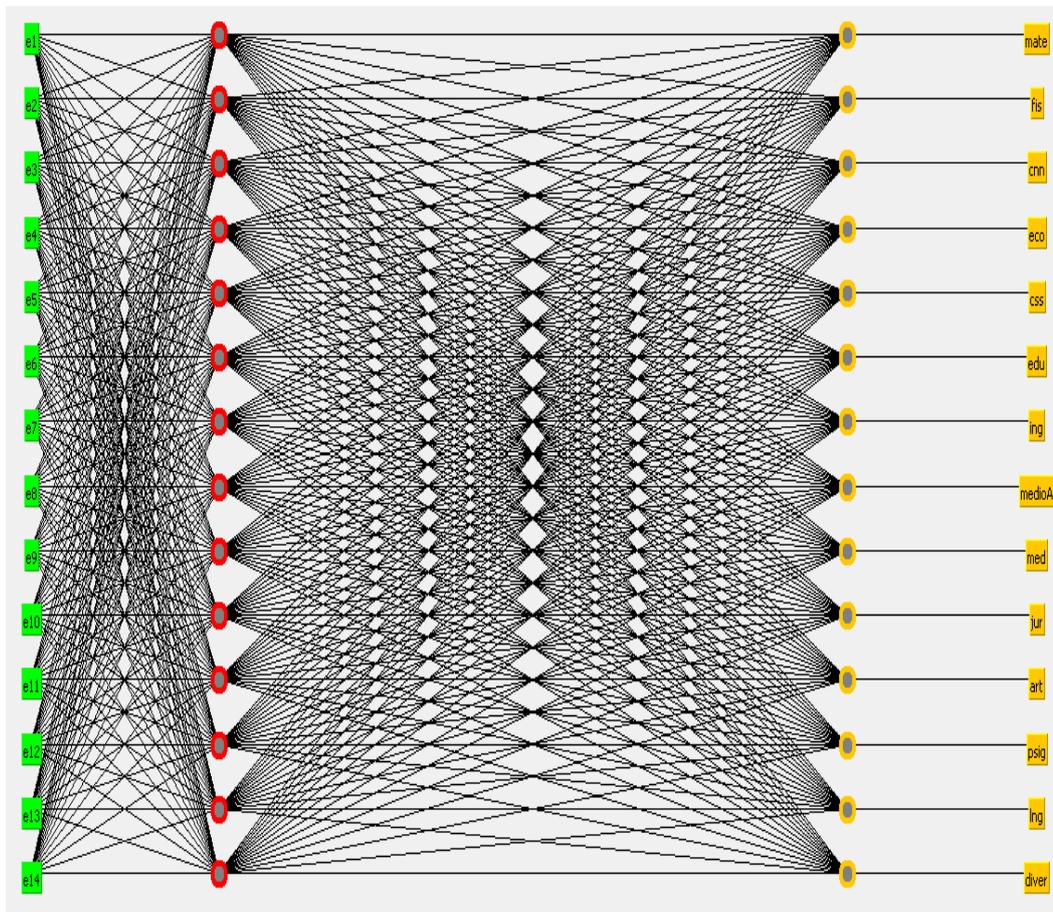
**Título:** este metadato describirá el título del documento lo cual servirá como entrada para nuestra red.

**Palabras claves:** este metadato describirá las palabras claves de un documento de igual forma servirá como entrada a la red.

**Descripción:** este metadato describe lo que está compuesto el libro una descripción pequeña del documento.

Estos 3 metadatos mencionados anteriormente se unirán para formar una sola entrada a la red la que definirá la categoría correspondiente.

El modelo final corresponde a una red con 14 entradas y 14 salidas.



*Figura 45. Resultado de la red en RWeka*

### 3. Desarrollo

#### 3.1. Algoritmo para la lectura de los metadatos.

En la figura 47, se observa el diseño del algoritmo el cual permite la lectura de los metadatos necesarios para nuestra aplicación ya que existe una gran variedad de metadatos en nuestra aplicación haremos uso de los siguientes metadatos como son: título, descripción, palabras claves, autor, url, año de publicación los cuales estos son los más comunes en la identificación de un documentos.

En la figura 47, se observa Abre conexión que representa la conexión con la URL del repositorio de estudios lo que permitirá realizar la lectura de un documento de tal forma que mientras haya conexión el algoritmo permitirá extraer los metadatos necesarios y de esta forma guarda la información del documentos en la base de datos.



**Figura 46 Diseño del algoritmo para lectura de metadatos**

A continuación se muestra el método para la lectura de metadatos el cual permite extraer los metadatos de título, autor, palabras claves, año, url, descripción, de un cierto libro del repositorio Red de Revistas Científicas de América Latina y el Caribe, España y Portugal, cada uno de estos metadatos identifican las características de un documentos y de esta forma poder tener la facilidad de acceso a la información de un documento en nuestra base de datos.

```

public boolean cargarLibroWeb(String urls){

    Document doc;
    try {
        URL url = new URL(urls);
        URLConnection conexion = url.openConnection();
        conexion.connect();
        InputStream is = conexion.getInputStream();
        BufferedReader br = new BufferedReader(new InputStreamReader(is));
        char[] buffer = new char[1000];
        int leido;
        List <String> cont = new ArrayList<String>();
    }
}
  
```

```

String aux = null;
while ((leido = br.read(buffer)) > 0) {
    aux = aux + new String(buffer, 0, leido);
}
doc = Jsoup.parse(aux);
String title = doc.title();
List<Element> autor = new ArrayList<>();
try {
    autor = doc.select("meta[name=citation_author]");
} catch (Exception e) {
}

String desc = doc.select("meta[name=description]").first().attr("content");

String keybor="";
try {
    keybor = doc.select("meta[name=keywords]").first().attr("content");
} catch (Exception e) {
    keybor = doc.select("meta[name=citation_keywords]").first().attr("content");
}

String urlPag = doc.select("meta[name=citation_pdf_url]").first().attr("content");
String autores = "";
if (!autor.isEmpty()) {
    for (Element element : autor) {
        autores = autores + " - "+element.attr("content");
    }
}
String ano = "";
try {
    ano = doc.select("meta[name=citation_publication_date]").first().attr("content");
} catch (Exception e) {
    ano = "";
}

Libro l=new Libro();
l.setTitulo(title);
l.setDescripcion(desc);
l.setFechaPublicacion(ano);
l.setAutor(autores);
l.setPalabrasClave(keybor);
l.setUrl(urlPag);
}
}

```

### 3.2. Desarrollo del método de la red de Perceptrón multicapa.

El Perceptrón Multicapa (MLP, por sus siglas en inglés “Multi-Layer Perceptrón”) tiene como objetivo la categorización o clasificación de forma supervisada. Para este trabajo se ha utilizado el algoritmo de MLP aplicado a la clasificación de textos en diferentes categorías compuestas por 14 categorías (física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso). Este algoritmo es el más complejo y es el encargado de realizar toda la clasificación de los documentos y su conexión con la base de datos, además este permite la integración de todos los métodos creados con la finalidad de lograr la clasificación de los documentos.

En la figura 48, muestra el desarrollo en si del funcionamiento de la red neuronal con Perceptrón multicapa este método permite que cualquier documento de entrada llegue a ser asignado a una categoría correspondiente.

```
public static String multilayerNeuralAlgorithm(String[] entrada) {
    String conclus = "diver";
    try {
        if (tienePalabras(entrada)) {
            String dirReporte = System.getProperty("user.dir") + "\\src\\bc\\categorizador.arff";
            ConverterUtils.DataSource archivo = new ConverterUtils.DataSource(dirReporte);
            Instances coleccion = archivo.getDataSet();
            coleccion.setClassIndex(coleccion.numAttributes() - 1);
            MultilayerPerceptron clasificador = new MultilayerPerceptron();
            clasificador.buildClassifier(coleccion);
            Instance datosdiag = new Instance(entrada.length + 1);
            datosdiag.setDataset(coleccion);
            for (int i = 0; i < entrada.length; i++) {
                datosdiag.setValue(i, Float.parseFloat(entrada[i]));
            }
            double indiceClase = clasificador.classifyInstance(datosdiag);
            conclus=coleccion.classAttribute().value((int) indiceClase);
        } else {
            conclus = "diver";
        }
    } catch (Exception e) {
        System.out.println("no tiene categoria" + e);
        return conclus;
    }
    return conclus;
}
```

**Figura 47 Método del funcionamiento de la red neuronal MLP**

## 4. Pruebas

### 4.1. Entrenamiento con el lenguaje R

#### ➤ Lenguaje R

R es un lenguaje de programación y entorno de software de código abierto para computación y gráficos estadísticos. Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásica, clasificación, agrupación en clústeres entre otros. [54].

Para la elección de R se han evaluado pues distintos aspectos, siendo especialmente destacables sus bondades en lo que se refiere a calidad, a la cantidad de técnicas y funciones implementadas, a que es libre y a la gran comunidad científica que lo usa como estándar para el análisis de datos. Dicha comunidad ha desarrollado y desarrolla herramientas integradas en paquetes en la actualidad más de 800, que dan solución a una gran variedad de problemas estadísticos.

#### ➤ Paquetes de lenguaje R para redes neuronales

Existe una diversidad de herramientas de software desarrollado bajo licencia GPL (Licencia Pública General), como las que se detallan a continuación: WEKA, RapidMiner, Tanagra, Pentaho, Kxen, Orange, SSPS Clementine, Statistica, entre otros. [55].

Paquetes que utiliza R para redes neuronales

- Paquete `nnet`: El paquete `nnet` nos permite crear redes neuronales de clasificación monocapa. [56].
- Paquete `RWeka`: para interactuar con Weka que permite leer y escribir ficheros en el formato *arff* y enriquecer R con los algoritmos de minería de datos de dicha plataforma.
- Paquete `AMORE`: permite la creación de modelos de regresión basados en modelos de MLP [56].
- Paquete `neural` (no está en los repositorios): permite la creación de modelos MLP con el algoritmo `RPROP` [56].

Los paquetes contribuidos se instalan directamente desde el ambiente de trabajo del lenguaje R, utilizando el menú Packages. Esto puede hacerse de dos maneras: Directamente de las páginas de CRAN, para lo cual hay que estar conectado al internet. A través de los archivos ejecutables previamente bajados en formato Zip, en esta opción no es necesario estar conectado a internet. Un aspecto interesante es que los paquetes que han sido instalados previamente pueden ser actualizados directamente de la página de CRAN, esta facilidad solo se puede utilizar si está conectado a internet

➤ **Utilización del paquete RWEKA**

- ❖ Para utilizar el lenguaje R se crea un script
- ❖ Instalamos paquetes con la opción `install.packages('RWeka')`: permite instalar RWEKA desde internet.
- ❖ `options(java.home="C:\\Program Files\\Java\\jre7\\")`: permite configurar el lenguaje R con la máquina virtual de Java
- ❖ `library(RJava)`: permite ejecutar librerías del lenguaje Java en R
- ❖ `library("RWeka")`: permite la utilización de la librería Weka en el lenguaje R utilizando todos sus atributos

➤ **Análisis de los resultados con el Lenguaje R**

El paquete RWEKA como ya se mencionó anteriormente permite realizar las diferentes pruebas de la red neuronal, verificar los resultados de entrenamiento que muestran en base a cada uno de los corpus de pruebas. Realizamos las pruebas con los siguientes comandos:

**TABLA IV PASOS PARA UNA CLASIFICACION MLP**

Lee el archivo arff y lo asigna a la variable data.	<code>data = read.arff("e:\\bc.arff")</code>
Clasificación Multilayer Perceptrón, el resultado se lo asigna a NN	<code>NN=make_Weka_classifier("weka/classifiers/functions/MultilayerPerceptron")</code>
Muestra información de cada variable que se puede utilizar en RWEKA	<code>WOW(NN)</code>

Es una variable que contiene le módulo de la RN con la siguiente estructura :formula, datos, control	ResultNN = NN(resultado ~. ,data=data, control=Weka_control(S=0,L=0.3,M=0.2,N=500,V=0,G=TRUE))
Contiene los datos de predicción de nuestro modelo con respecto a la BC en la variable Summary	Summary = summary(ResultNN)

### ➤ Resultados y experimentos

Para realizar los estudios necesarios se utilizaron 2 muestra de diferentes tamaños utilizadas en la fase de entrenamiento de la red, a continuación se presentan y se discuten los resultados obtenidos en RWEKA en base a cada uno de ellos.

Durante el aprendizaje o entrenamiento del sistema se evalúan las condiciones de pertenencia a cada una de las categorías. El aprendizaje supervisado se caracteriza por conocer la respuesta que debería tener la red frente a una determinada entrada. De esta manera, se compara la salida deseada con la salida de la red verificando que se esté llevando correctamente la categorización.

#### **Muestra1.**

De los 5000 documentos se seleccionó un primer conjunto de entrenamiento consta de 1000 documentos, se observa que sus resultados tienen un alto porcentaje de clasificación correcta lo cual indica que hay una buena clasificación, a continuación se muestran sus resultados en porcentajes de la versatilidad de clasificación y margen de error.

```
=== Evaluation on training set ===
=== Summary ===
```

```
Correctly Classified Instances      1152          97.4619 %
Incorrectly Classified Instances     30           2.5381 %
Kappa statistic                     0.9717
Mean absolute error                  0.011
Root mean squared error              0.0629
Relative absolute error               8.5963 %
Root relative squared error          24.8401 %
```

Al analizar la primera muestra se pudo apreciar que las instancias de clasificación correcta muestran un 97% de documentos bien clasificados y un 3% una mínima parte de margen de error indicando un mayor rango de clasificación correcta de documentos como se muestra en la Figura 49.



**Figura 48 Resultado de la muestra 1**

La matriz de confusión muestra el tipo de las predicciones correctas e incorrectas cuando se aplica el modelo sobre el conjunto de prueba. La misma permite comprender en qué sentido se equivoca la red al intentar clasificar los nuevos textos. En el gráfico de esta matriz, las predicciones correctas están representadas sobre la diagonal, mientras que el resto indican el tipo de error cometido (qué valor ha predicho el modelo y cuales el valor verdadero).

```

=== Confusion Matrix ===
  a  b  c  d  e  f  g  h  i  j  k  l  m  n  <-- classified as
164  0  0  0  0  0  0  0  0  0  0  0  0  0 | a = mate
  0 198  0  0  0  0  0  0  0  0  0  0  0  0 | b = fis
  0  0  98  0  0  0  0  0  0  0  0  0  0  0 | c = cnn
  0  0  0 120  0  0  0  0  0  0  0  0  0  1 | d = eco
  0  0  0  1 141  0  0  0  0  0  0  0  0  0 | e = css
  0  1  0  0  0 109  0  0  0  0  0  0  0  0 | f = edu
  0  2  1  0  0  0 13  0  0  0  0  1  0  0 | g = ing
  1  0  1  0  0  0  0 69  0  0  0  0  1  1 | h = medioA
  0  1  1  0  0  0  0  0 24  0  2  0  0  0 | i = med
  0  0  0  0  1  1  0  0  0 51  0  0  0  2 | j = jur
  0  0  1  0  2  0  0  0  0  0 20  0  0  0 | k = art
  2  2  2  0  0  0  0  0  1  0  0 62  0  0 | l = psig
  0  0  0  0  0  0  0  0  0  0  0  0 21  0 | m = lng
  1  0  0  0  0  0  0  0  0  0  0  0  0 62 | n = diver

```

## Muestra2.

El segundo conjunto de datos de entrenamiento se llevó a cabo con 3000 documentos al incrementar la cantidad de documentos se creó conveniente agregar nuevas palabras al vocabulario, con la finalidad de que el margen de asertividad no decaiga.

```
=== Evaluation on training set ===  
=== Summary ===  
  
Correctly Classified Instances      2223      89.3489 %  
Incorrectly Classified Instances    265      10.6511 %  
Kappa statistic                    0.8819  
Mean absolute error                0.023  
Root mean squared error            0.1109  
Relative absolute error             17.8743 %  
Root relative squared error        43.7063 %
```

El resultado de este nuevo grupo de entrenamiento muestra un valor del 90% de instancias correctas y con un margen de error del 10% ver figura 50, a diferencia del primer grupo se redujo un mínimo porcentaje el margen de asertividad debido a la gran cantidad de documentos, es decir que a medida que el aumento el número de documentos a ser analizados se deberá incrementar el vocabulario.



**Figura 49. Resultados de la muestra 2**

La siguiente matriz de confusión muestra las predicciones correctas e incorrectas de cada una de las categorías.

```

== Confusion Matrix ==

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  <-- classified as
153 0  4  1  2  4  0  0  0  0  1  0  0  0 | a = mate
  1 198 2  4  0  3  6  1  0  1  5  0  2  0 | b = fis
  2  0 121 3  2  5  1  8  0  0  2  1  1  0 | c = cnn
  1  0  1 113 2  4  0  0  1  2  2  1  0  0 | d = eco
  3  0  0  2 146 8  1  2  0  3  7  1  1  0 | e = css
  0  0  1  1  1 357 2  1  1  3  4  0  2  0 | f = edu
  4  4  1  1  2  9 354 3  1  2  7  1  5  0 | g = ing
  2  2  3  2  3  4  4 267 3  0  2  0  6  1 | h = medioA
  6  0  3  1  0  3  2  2 165 0  1  0  0  0 | i = med
  4  0  0  2  2  4  1  0  0 104 4  1  1  0 | j = jur
  3  1  0  0  0  2  1  2  0  0 101 0  2  0 | k = art
  5  1  3  0  0  2  1  1  0  2  1 22  1  0 | l = psig
  0  0  1  1  0  1  0  0  0  0  0  0 120 0 | m = lng
  1  0  1  1  0  1  0  0  0  0  1  0  0  2 | n = diver

```

Finalmente, estos resultados se van a ir mejorando con el ingreso de nuevos documentos, al ser un sistema supervisado esto logrará a llegar a obtener un grado mayor de asertividad de clasificación. Además el sistema clasificara los 5000 documentos descargados del repositorio elegido.

## 5. Implementación

Para implementar el sistema es necesario tener instalado un gestor de base de datos, Servidor web y un lenguaje interpretado. MySQL es utilizado para la gestión con los datos, Java es el que permite el desarrollo del sistema y el servidor para analizar las peticiones de los clientes.

Una vez entrenada la red y realizadas las pruebas con los dos grupos de entrenamiento correspondientes a los 5000 documentos del repositorio de Red de Revistas Científicas de América Latina y el Caribe, España y Portugal se utilizó como fase final la implementación de 100 documentos de la biblioteca del AEIRNNR de Universidad Nacional de Loja de tal manera que con este nuevo conjunto de documentos se logró demostrar el correcto funcionamiento del sistema bajo la supervisión de la bibliotecaria (ver anexo 7) dando fe de que el sistema es capaz de categorizar correctamente cualquier documentos que sea ingresado y determinar la categoría a la que pertenece .

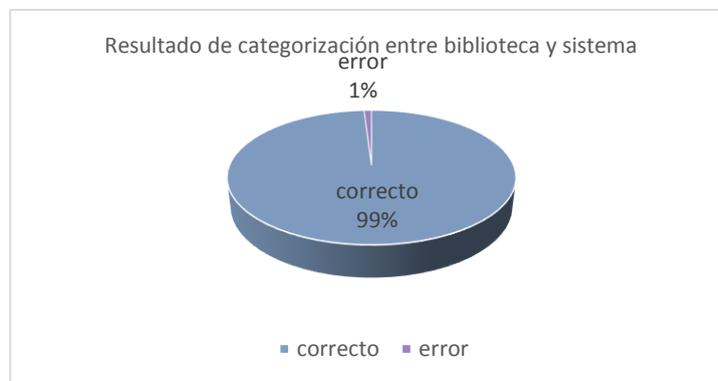
En la siguiente tabla se muestra algunos de los libros de la biblioteca del Área que fueron utilizados para la categorización, así mismo se detalla la categoría asignada por la biblioteca y la categoría que el sistema asigna al documento.

<b>Tema:</b>	<b>Categoría asignada por la biblioteca del AEIRNNR</b>	<b>Categoría asignada por el sistema</b>
instrumentación industrial	Instrumentos para pruebas, mediciones, sensores	Ingeniería
Circuitos eléctricos.	Redes	Física
El montador electro mecánico.	Ingeniería eléctrica	Física
Máquinas eléctricas: operación en estado estacionario.	Ingeniería eléctrica	Física
Curso práctico de electricidad.	Ingeniería eléctrica	Física
Circuitos eléctricos.	Ingeniería eléctrica	Física
Generadores de vapor.	Generación y transmisión de vapor	Física
Formulario de mecánica.	Mecánica y materiales de ingeniería	Matemáticas

Manual de recipientes a presión: diseño y cálculo.	Ingeniería y operaciones afines	Matemáticas
Principios de eco toxicología: diagnóstico, tratamiento y gestión del medio ambiente.	Toxicología industrial	Medio ambiente
Seguridad industrial y salud.	Salud industrial y ocupacional	Medicina
Dibujo técnico.	Dibujo técnico	Arte
Enciclopedia juvenil: ciencia y técnica.	Diccionarios , enciclopedias	Física
Libro rojo de los mamíferos del Ecuador.	Mammalia	Ciencias sociales
Química medioambiental.	Ecología	Medio ambiente
Ecología: el vínculo entre las ciencias naturales y las sociales.	Ecología	Medio ambiente
Biología general.	Biología	Ciencias naturales
Paleontología: vertebrados, peces, anfibios, reptiles y aves.	Paleontología	Ciencias naturales
Paleontología: parte general e invertebrados.	Paleontología	Ciencias naturales
Geología física básica.	Geología ,hidrología ,meteorología	Medio ambiente
Geología.	Geología	Medio ambiente
Minerales y rocas: una guía de identificación con 400 fotografías en color 600 dibujos y una introducción a la mineralogía y la petrografía.	Minerales en rocas	Arte
Mineralogía aplicada; para ingenieros, técnicos y estudiantes.	Mineralogía	Medio ambiente
Nociones de minería.	Minería y operaciones relacionales	Medio ambiente
Atlas de Mineralogía.	Mineralogía	Medio ambiente

Riesgos naturales: procesos de la tierra como riesgos, desastres y catástrofes.	Geología .meteorología	Medio ambiente
Ciencias de la tierra: una introducción a la geología física.	Ciencias de la tierra	Medio ambiente
Problemas de campos electromagnéticos.	Electricidad y electrónica	Física
Electromagnetismo y circuitos eléctricos.	Electricidad y electrónica	Física
Curso de electricidad general.	Electricidad	Física
Termodinámica.	Termodinámica	Física
Teorías termológicas: aplicación a la arquitectura y a las ingenierías.	Temperatura	Física
Transferencia de calor y masa: fundamentos y aplicaciones	Física	Física
Física universitaria.	Física	Física
Fundamentos de física conceptual.	Física	Física
Introducción a la física II: acústica-óptica electromagnetismo.	Física	Física
Levantamiento aerofoto gráfico.	Fotogrametría	Medio ambiente
Inferencia estadística y análisis de datos.	Inferencias estadística	Matemáticas
Geometría analítica.	Geometría analítica	Matemáticas
Teoría y problemas de geometría analítica: plana y del espacio.	Geometría analítica	Matemáticas
Geometría y trigonometría.	Trigonometría	Matemáticas
Cálculo.	Análisis vectorial	Matemáticas
Introducción a los métodos numéricos.	Ecuaciones mixtas	Matemáticas
Matemáticas avanzadas para ingeniería: ecuaciones diferenciales.	Ecuaciones diferenciales	Matemáticas

Luego de revisar los datos de la tabla se pudo notar que la clasificación referente a la categoría que tiene la biblioteca del área con respecto a la categoría que asigna el Sistema es similar, ya que en nuestro caso las categorías que tenemos son categorías generalizadas lo cual abarca muchos contenidos, demostrando que la clasificación se está llevando a un 99% de exactitud en la clasificación y un 1% de error como se muestra en la figura 51, pero debido a que es una clasificación con aprendizaje supervisado al no categorizar correctamente algún tipo de documento el supervisor podrá indicar la categoría que a su criterio sea la correcta .



**Figura 50. Resultado entre la clasificación del sistema y de la biblioteca**

## **g. Discusión**

### **1. Desarrollo de la propuesta alternativa**

El desarrollo de la propuesta alternativa se basa en la realización y cumplimiento de los objetivos planteados.

#### ***1. Objetivo específico 1: Analizar el corpus de datos que van a ser tratados por los algoritmos de lectura y mediante metadatos permita el aprendizaje para la clasificación de documentos.***

Para alcanzar el presente objetivo haremos uso de la colección de Red de Revistas Científicas de América Latina y el Caribe, España y Portugal de la Universidad autónoma de México (ver sección Resultados fase 1), ya que de acuerdo al análisis realizado en el transcurso del trabajo sobre este repositorio se puede notar que es el más idóneo para este trabajo. Lo cual este repositorio permitirá demostrar la clasificación automática de los documentos para 14 categorías (física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso) estas categorías fueron seleccionadas después de hacer un revisión exhaustiva de los documentos contenidos en el repositorio.

Luego para obtener información de los documentos se realizó un método para extraer los metadatos este tiene la función de obtener solo los metadatos que nos van a servir para nuestro trabajo.

Para la entrada de la red vamos a utilizar los metadatos de título, descripción, y palabras claves, como primer paso es reducir todas las palabras a su raíz (en inglés “stemming”). Por ejemplo, las palabras “medicina”, “médico” y “medicinal” se reducen a la forma “medic”, todos estos metadatos serán sometidos a este proceso esta etapa se denomina lematización o reducción a raíces léxicas (en inglés “stemming”). Almacenando sólo las raíces de los términos, se puede llegar a reducir sudimensión considerablemente presentando la ventaja de ser más eficiente y ahorrar espacio. En nuestro caso hemos optado por aplicar el algoritmo de Snowall Steamer de esta forma obtendremos una lista de términos lematizados representativos de cada documento.

Tras aplicar el proceso de stemming al corpus de estudio se tienen los documentos y todos sus términos correspondientes en lexemas que corresponden a la raíz de los términos procesados, los resultados fueron aceptables para los fines experimentales, porque se redujo considerablemente el tamaño de los mismos.

**2. *Objetivo específico 2: Diseñar e implementar la red neuronal supervisada para la clasificación de los documentos.***

Dentro de este objetivo para realizar la clasificación de documentos se ha utilizado redes neuronales con Perceptrón multicapa en conjunto con el lenguaje R con el paquete RWEKA para dichos entrenamiento de la red.

Esta herramienta permite realizar los entrenamientos correspondientes de los documentos indicando que el resultado de entrenamiento se esté dando correctamente, de tal manera se pueda determinar si se está llevando correctamente la clasificación de los documentos indicándonos margen de asertividad y error en la clasificación y su matriz de confusión. Para dicho entrenamiento se han tomado muestras de documentos de estas colecciones para analizar el comportamiento del algoritmo a fin de maximizar el número de aciertos, y ver el promedio de aciertos entre las diferentes muestras y finalmente hacer uso del sistema.

**3. *Objetivo específico 3: Demostrar el funcionamiento del proceso de clasificación de los documentos a través de la red neuronal supervisada.***

Para el desarrollo y culminación del objetivo, se procedió a demostrar y probar el funcionamiento del sistema, ingresando manualmente 100 documentos de la Universidad Nacional de Loja de la biblioteca del AEIRNNR el mismo que permite observar que es capaz de clasificar cualquier documento a su categoría correspondiente, además de demostrar que es aplicable a cualquier tipo de biblioteca que se quiera implementar ver anexo 7.

## 2. Valoración técnica económica ambiental

El presente TT, titulado “Desarrollo de un sistema inteligente para la clasificación de documentos ya digitalizados aplicando red neuronal supervisadas” se considera desde el punto de vista técnico como un aporte viable para futuras investigaciones, como es en el caso de mejorar la eficiencia y reducir el tiempo de clasificación manual de documentos .

En el aspecto económico el proyecto como desarrollo es apto para trabajar en muchos sistemas que hagan uso de clasificación de documentos, contribuirá como apoyo a futuras investigaciones para personas interesadas que quieran mejorar o tomarlo como guía y adquirir sin necesidad de un coste en la compra del software, ni pagos por licencia, permitiendo con ello obtener ahorros significativos. Destacando que el trabajo en estudio no tiene ningún impacto negativo en el ecosistema ya que no existe peligro alguno para el medio ambiente al momento de ser implementado por el ahorro en la utilización de papel en la impresión de libros y en la copia ya que estos serán digitalizados en algún sistema.

A continuación se detalla el talento humano, material, técnico, tecnológico y servicios utilizado en el trabajo de titulación:

El talento humano que participó en el trabajo de Titulación, está conformado principalmente por el investigador quien fue el encargado de llevar a cabo el desarrollo del presente trabajo, se tuvo la ayuda de un asesor de proyectos quien fue guía para esquematizar de mejor forma el anteproyecto. Sin duda durante el transcurso del desarrollo, se contó con las tutorías del director de tesis. En la Tabla V se detalla los recursos humanos.

**TABLA V TALENTO HUMANO**

DESCRIPCIÓN	Tiempo(Horas)	(\$)Precio/Horas	(\$)Valor Total
Investigador	400	8.00	3200.00
Director de Tesis	200	0.00	0.00
<b>SUBTOTAL</b>			3200.00

En la Tabla VI se hace una descripción detallada de los recursos materiales que fueron necesarios para presentar los borradores y el informe final del trabajo de titulación.

**TABLA VI RECURSOS MATERIALES**

DESCRIPCIÓN	CANT.	(\$)Valor Unitario	(\$)Valor Total
Resma de papel	2	3.80	7.60
Cartuchos de tinta	5 h	20.00	100.00
Flash Memory	1	15.00	15.00
<b>SUBTOTAL</b>			122.60

En la Tabla VII y VIII se aprecia la suma parcial de los recursos técnicos y tecnológicos usados para su posterior inclusión en la suma total de recursos usados.

**TABLA VII RECURSOS TÉCNICOS**

DESCRIPCIÓN	Tiempo de utilización(horas)	(\$)Valor Unitario	(\$)Valor Total
Computadoras	400	1.00	400.00
Impresora	----	85.00	85.00
<b>SUBTOTAL</b>			485.00

**TABLA VIII RECURSOS TECNOLÓGICOS**

DESCRIPCIÓN	Valor Unitario	Desarrollador	Costo Total
JSE 1.7	Gratuito	Sun Microsystems	----
Enterprise Architect 3.6	Demo 30 días	SpaxSystems	----

<b>Open Office V.3.0</b>	Gratuito	Sun Microsystems	----
<b>Grand Project</b>	Gratuito	-----	----
<b>WEKA</b>	Licencia GNU- GPL.	Universidad de Waikato.	-----
<b>Lenguaje R</b>	Licencia GNU- GPL		
<b>SUBTOTAL</b>			00.00

La tabla IX. Describe los servicios que fueron necesarios durante el desarrollo del trabajo.

**TABLA IX SERVICIOS**

<b>Servicio</b>	<b>Descripción</b>	<b>(\$)</b> Valor <b>Unitario</b>	<b>(\$)</b> Valor <b>Total</b>
<b>Internet</b>	5 meses	20.00	100.00
<b>Transporte</b>	100 días	0.50	50.00
<b>SUBTOTAL</b>			150.00

En la Tabla X. Presenta la suma total del talento humano, material, técnico, tecnológico y servicios utilizado en el trabajo de titulación, que da una aproximación del costo real.

**TABLA X COSTO TOTAL**

<b>DESCRIPCIÓN</b>	<b>(\$)</b> Total
<b>Talento Humano</b>	3200.00
<b>Materiales</b>	122.60
<b>Técnicos</b>	485.00
<b>Tecnológicos</b>	00.00
<b>Servicios</b>	150.00
<b>Imprevistos 10%</b>	409.76
<b>TOTAL (\$)</b>	4367.36

## **h. Conclusiones**

- La utilización de lenguaje R provee diversos paquetes como RWeka que permite la ejecución de pruebas de entrenamiento de la red con la finalidad de observar resultados sobre la pertenencia de un texto.
- Para garantizar una buena clasificación de documentos se creyó necesario en la muestra 2 se incrementa un mayor número de palabras al vocabulario con la finalidad de dar una mejor exactitud en la clasificación de los documentos.
- Definir correctamente el vocabulario de cada una de las categorías hace que la clasificación tenga un buen porcentaje de aciertos logrando que el documento sea asignado correctamente a su categoría, de tal medida que el desempeño de la exactitud de la red mejora de acuerdo al tamaño del vocabulario.
- La utilización de metadatos fueron de gran ayuda para lograr mejores resultados a la hora de la representación, localización y recuperación de recursos electrónicos.

### **❖ Trabajos futuros**

A raíz de la investigación realizada, han quedado abiertas diferentes líneas de trabajo para un desarrollo futuro:

- Extensión de las técnicas de aprendizaje para agrupamiento de documentos considerando la posibilidad de devolver los resultados ordenados por relevancia en la recuperación de los documentos.
- Extensión del sistema de clasificación, exploración y búsqueda para clasificar la colección de tesis considerando el contenido.
- Mejorar la visualización de los contenidos cuando se agrupe una gran colección documental, mostrando gráficamente los distintos niveles de una jerarquía, utilizando colores significativos asociados a cada metadato importante. La

posibilidad de representar en un único mapa toda una colección formada por miles de documentos, e ir descendiendo en los diversos niveles jerárquicos que se hayan establecido, sin cambiar la forma de representación, proporcionaría una interfaz robusta, intuitiva y fácil de usar. Hay que tener en cuenta que, en la actualidad, grandes corporaciones están destinando muchos recursos en la búsqueda de sistemas de visualización para mostrar grandes colecciones de documentos, sin que hasta ahora se hayan conseguido unos resultados que conciten una satisfacción general.

- **Evaluar la arquitectura en nuevos escenarios de extracción:** La extracción de información puede ser llevada a muchos escenarios Ej. correos electrónicos, páginas Web, artículos científicos entre otros.
- **Emplear técnicas de aprendizaje supervisado:** Para implementar la arquitectura es necesario contar con un corpus de entrenamiento previamente clasificado y etiquetado para demostrar el funcionamiento del clasificador.

## **i. Recomendaciones**

- Se recomienda analizar con mucho cuidado la forma de representar los contenidos de los documentos o normalizar las entradas de tal forma que estos sean comprensibles para el sistema de clasificación .
- Se recomienda siempre tener un visor de error mientras la red se entrena, ya que a la hora de probar permite constatar el valor de asertividad y de error de la taza de aprendizaje.
- Para eliminar caracteres que no ofrecen ningún tipo de información se recomienda realizar un lematizador lo que permite una menor dimensión de textos.
- Las muestras utilizados para el entrenamiento del Perceptrón multicapa deben seleccionarse adecuadamente para que la red adquiera la capacidad de arrojar buenos resultados.

## j. Bibliografía

### Referencias Bibliográficas

[1] Mendialdua, " *Aproximaciones a SVM semisupervisado multiclase para clasificación de páginas Web*", Tesis de Master, Dept. De Lenguajes y Sistemas, Universidad Nacional de Educación a Distancia, , 2007-2008.

[2] Pérez Abelleira, M.A. y Cardoso, C.A., " *Minería de texto para la categorización automática de documentos*",2010; [En línea] Link: <http://www.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p11-alicia-articulo-cuadernos-formateado.pdf>

[3] Mendoza, M. y Ortiz, I., " *Categorización de textos en bases documentales a partir de modelos computacionales livianos* ", *Revistas Signos*, vol. 44(77), dic 2011.

[4] " *Ciclo de vida de un sistema de información*". [En línea] Link: <http://elvex.ugr.es/idbis/db/docs/design/1-process.pdf>

[5] " *Gestión Documental*", [En línea] link: [http://www.tcmugt.es/puertos/images/documentos/formacion/contenidos\\_y\\_manuales\\_formativos/manuales\\_formativos/GESTION\\_DOCUMENTAL\\_%283%29.pdf](http://www.tcmugt.es/puertos/images/documentos/formacion/contenidos_y_manuales_formativos/manuales_formativos/GESTION_DOCUMENTAL_%283%29.pdf)

[6] Castillo Sequera, J.L., " *Nueva propuesta evolutiva para el agrupamiento de documentos en sistemas de recuperación de información*", Tesis Doctoral, Dept. De Ciencias de la Computación, Universidad de Alcalá, 2010.

[7] Yolis, E., Britos, P., ' *Algoritmos Genéticos Aplicados A La Categorización Automática De Documentos* "Tesis de grado en Ingeniería Informática, Laboratorio de Sistemas Inteligentes, Universidad de Buenos Aires, abril 2003

[8] Almuna Herrera,E.M., " *Contenido para repositorios de esquema y metadatos*" Ingeniero Civil de computación, Universidad de Chile ,Facultad de Ciencias de Física y matemáticas, Dept. de ciencias de la computación, Enero 2007.

[9] Pacheco, I.R., J.M., Muñido. " *Los Metadatos y principales repositorios de objetos digitales* "; CODICE Gestión de la Información, España. <http://users.dcc.uchile.cl/~cvasquez/introehistoria.pdf>

[10] Vásquez Paulus.C., " *METADATOS: Introducción e historia*" [En línea] Link <http://users.dcc.uchile.cl/~cvasquez/introehistoria.pdf>

[11] Tesca, P., Ceriotto, P., " *Descripción de objetos digitales: Metadatos* ", Sistema Integrado de Documentación, Universidad Nacional del Cuyo.

[12] " *Información y documentación - procesos de gestión de documentos: metadatos para la gestión de documentos*", revista española de documentación científica 31,2, ,273-301, Abril-Junio 2008 ISSN 0210.0614

[13] Adell,J., Bellver, C,. " *La internet como telaraña en la w.w.w* ", métodos de información, vol 2, N 3, ENERO 1995, P.25-32

- [En línea]  
<http://www.metodosdeinformacion.es/mei/index.php/mei/article/viewFile/63/85>
- [14] Aguilar Rivera, E.G., y Dávila Garzon, D.A., "Análisis, Diseño e Implementación de la Aplicación Web para el manejo del Distributivo de la Facultad de Ingeniería", Tesis de Grado en Ingeniería en Sistemas, Facultad de Ingeniería, Universidad de Cuenca, 2013.
- [15] Elena Acosta, Andreina Rojas, Sabrina Ridente, Joel Tarazona, "internet, extranet, intranet".
- [16] Lugan Mora, S., "Programación de aplicaciones Web: Historia, principios básicos y cliente web", Editorial club universitario, ISBN 84-8454-206-8
- [17] "Modelo Cliente Servidor", Dept. Informática, Universidad de Valladolid [http://www.infor.uva.es/~fdiaz/sd/2005\\_06/doc/SD\\_TE02\\_20060305.pdf](http://www.infor.uva.es/~fdiaz/sd/2005_06/doc/SD_TE02_20060305.pdf)
- [18] Steban Sánchez Mendoza, L.I., "Antología de la programación Web", Ingeniería en sistemas computacionales, Enero 2012.
- [19] "Lenguajes a lado del servidor o cliente" [En línea] Link [http://www.adelat.org/media/docum/nuke\\_publico/lenguajes del lado servidor o cliente.html](http://www.adelat.org/media/docum/nuke_publico/lenguajes_del_lado_servidor_o_cliente.html)
- [20] "Introducción al Lenguaje Java", Universidad de Navarra, [En línea] Link <http://www.unav.es/SI/manuales/Java/indice.html>
- [21] Javier J. Gutiérrez. "Qué es un Framework web" [En línea] Link [http://www.lsi.us.es/~javierj/investigacion\\_ficheros/Framework.pdf](http://www.lsi.us.es/~javierj/investigacion_ficheros/Framework.pdf)
- [22] Hidalgo López, C.R., Quisiguiña, Guevara, F. A., "Análisis, diseño y codificación del módulo de inventarios de especies valoradas para SION de correos del Ecuador", Tesis para obtención de Ingeniero en Sistemas, Facultad de Ingeniería en Sistemas, Universidad politécnica Salesiana, Quito mayo 2010.
- [23] "JBoss. Hibernate Community Documentation", [En línea] Link <http://docs.jboss.org/hibernate/orm/3.5/reference/es-ES/html/tutorial.html>
- [24] "Espinosa, A.T., Sagredo, J.G.C., Reyes, M.M., "Automatización de la codificación del patrón modelo vista controlador (MVC) en proyectos orientados a la Web", Revista científica multidisciplinaria de Mexico, vol 19.N 3, 2012 ISBN 148.215.1.245
- [25] "Sistema de gestor de base de datos", [En línea] Link [http://www.ecured.cu/index.php?title=Sistema Gestor de Base de Datos&oldid=2125072](http://www.ecured.cu/index.php?title=Sistema_Gestor_de_Base_de_Datos&oldid=2125072)
- [26] COBO, ÁNGEL y GÓMEZ, PATRICIA. 2005. "PHP y MySQL- tecnologías para el desarrollo de aplicaciones web." S.l.: Ediciones Díaz de Santos, 2005. 8479787066.
- [27] Alcaraz Ariza, F.J., "Clasificación y ordenación con R", Universidad de Murcia, Febrero 2013.
- [28] "Descarga de lenguaje R" [En línea] Link <http://www.r-project.org>.

- [29] Arriaza Gomez,A.J., Fernández Palacin,F.,López Sanchez,M.A., Muñoz Márquez., Pérez Plaza,S., Sánchez Navas, A., “*Estadística Básica con R y R-commander*”, Servicio de Publicaciones de la Universidad de Cadiz,2008, [En línea] Link <http://knuth.uca.es/ebrcmdr>
- [30] García Gonzales, F.J., “*Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA)*”, Trabajo fin de Master, Universidad de Granada, 2013.
- [31] “*Inteligencia Artificial /SOFTWARE LIBRE EXISTENTE PARA INTELIGENCIA ARTIFICIAL*”, noviembre 2013 [En línea] Link <http://www.mathworks.com/products/neural-network/>
- [32] López Takeyas, B., “*Introducción a la inteligencia artificial*”, Instituto Tecnológico de Nuevo Laredo, Reforma Sur 2007, C.P. 88250, Nuevo Laredo, Tamps. México, [En línea] LINK <http://www.itnuevolaredo.edu.mx/takeyas/Articulos/Inteligencia%20Artificial/ARTICULO%20Introduccion%20a%20la%20Inteligencia%20Artificial.pdf>
- [33] Alvarado Rodríguez, J.A., García Núñez, Y.N., Hernández Santos, W., Pimentel Vargas, N.A., Tristán Martínez., “*Redes neuronales*”, Instituto Tecnológico de nuevo Laredo, junio 2005.
- [34] Malpica Velasco, J.A., “*Inteligencia Artificial y conciencia*”, [En línea] Link [http://www2.uah.es/benito\\_fraile/ponencias/inteligencia-artificial.pdf](http://www2.uah.es/benito_fraile/ponencias/inteligencia-artificial.pdf)
- [35] Matich, D.J., “*Redes Neuronales: Conceptos Básicos y Aplicaciones*”, Facultad Regional Rosario, Dept. Ingeniería Química, Universidad Tecnológica Nacional, marzo 2001.
- [36] ARAUJO, B., “*Redes Neuronales aprendizaje automático: conceptos*”, edición n 1, ISBN 84-8322-318-X
- [37] “*Principales tipos de redes neuronales*”, [En línea] link <http://medicinaycomplejidad.org/pdf/redes/Perceptron.pdf>
- [38] Tabares,H.,Branch,J.,Valencia,J., “*Generación dinámica de la topología de una red neuronal artificial del tipo perceptron multicapa*”, Revista Facultad de Ingeniería N.o38. pp. 146-162. Septiembre, 2006
- [En línea] Link <http://www.scielo.org.co/pdf/rfiua/n38/n38a14.pdf>
- [39] Cárdenas Almeida, R., “*Inteligencia Artificial*”, [En línea] Link [http://www2.ulpgc.es/hege/almacen/download/38/38584/practica\\_ia\\_2.pdf](http://www2.ulpgc.es/hege/almacen/download/38/38584/practica_ia_2.pdf)
- [40] Alvarado Valderrama, J.E., “*Red Perceptrón Multicapa*”, Dept. Profesional de Informática, Universidad Nacional de Trujillo, Perú
- [41] Rodríguez Robles, V., Calderón Aller, C., “*Redes neuronales*”, Escuela de Ingeniería Industrial, [En línea] Link [http://www.academia.edu/7245695/Redes\\_Neuronales\\_Artificiales](http://www.academia.edu/7245695/Redes_Neuronales_Artificiales)

- [42] Matich, J.D., "*Redes Neuronales: Conceptos Básicos y Aplicaciones*", Facultad Regional Rosario, Dept de Ingeniería Química, Grupo de Investigación Aplicada a la Ingeniería Química, Universidad Tecnológica Nacional, 2011
- [43] Martínez., Díaz, M.C., Martín, M.T., Rivas, V.M., "*Aplicación de redes neuronales y redes bayesianas en la detección de multipalabras para tareas IR*", Universidad de Jaén, España
- [44] Venegas, R., "*Clasificación de textos académicos en función de su contenido léxico-semántico*", Rev. Signos v.40 n.63, 239-271 Universidad Católica Valparaíso 2007.
- [45] Beltrán, "*Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática*". Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina, Revista de Epistemología y Ciencias Humanas
- [46] "*Repositorio de Ciencia - Science Library*" [En línea] Link <http://bz.otsoa.net/>
- [47] "*Biblioteca virtual Universal*", [En línea] Link <http://www.biblioteca.org.ar/default.asp>
- [48] "*Repositorio Institucional de Trabajos de fin de Titulación*", [En línea] Link <http://dspace.utpl.edu.ec/>
- [49] "*Red de Revistas Científicas de América Latina y el Caribe, España y Portugal*", [En línea] Link <http://www.redalyc.org/homeBasic.oa>
- [50] Sparx Systems, "*Enterprise Architect 11*", Copyright Sparx Systems 2014 [En línea] Link: <http://www.sparxsystems.com/downloads/whitepapers/EAReviewersGuide.pdf>
- [51] "*Introducción al sistema de clasificación decimal Dewey*" [En línea] Link <http://bibliotecamachala.wikispaces.com/file/view/Clasificacion+decimal+dewey.pdf>
- [52] Catillo Sequera, J.J., "*Nueva propuesta evolutiva para el agrupamiento de documentos en sistemas de recuperación de información, tesis doctoral*", Escuela técnica Superior de Ingeniería Informática, Dpte.de las ciencias de computación. Universidad de Alcalá, 2010
- [53] "*Spanish stemming algorithm*" [En línea] Link <http://snowball.tartarus.org/algorithms/spanish/stemmer.html>
- [54]. García González, F.J., "*Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA)*", Trabajo fin de master, Universidad de Granada, 2013
- [55] Ruiz.C.A.,Basualdo,M.S., "*Redes neuronales*",Catedra aplicada a la ingeniería de procesos, Universidad tecnológica Nacional, Facultad Regional Rosario , departamento de Ingeniería Química, grupo de investigación aplicada a la ingeniería química , 2001 [En línea] Link <ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf>

[56] Velásquez, J.D., Zambrano, C., "*ARRN: Un paquete para la predicción de series de tiempo usando redes neuronales autoregresivas*", Facultad de Minas", Universidad Nacional de Colombia, Julio 2001

## **K. Anexos**

### **Anexo 1. Artículo**

#### **Desarrollo de un sistema inteligente para la clasificación de documentos ya digitalizados aplicando redes neuronales supervisadas**

Doris Jiménez<sup>a</sup>, Henry Paz<sup>b</sup>

<sup>a</sup> Universidad Nacional de Loja, Ecuador, ya\_dy\_j20@hotmail.com

<sup>b</sup> Universidad Nacional de Loja, Ecuador, hpaz@unl.edu.

#### **Resumen.**

El aumento exponencial de la información disponible en formato digital durante los últimos años y las expectativas de crecimiento futuro, hacen necesaria la organización de la información con el fin de mejorar la búsqueda y acceso a la información. Por esta razón adquiere importancia la investigación e implementación de un sistema de clasificación automática de textos que permitan la organización y facilidad de categorizar documentos de acuerdo a su categoría correspondiente utilizando redes neuronales con aprendizaje supervisado, de tal manera que realice un proceso más rápido en un menor tiempo y costo. El criterio para realizar la clasificación de documentos está basada de acuerdo a categorías definidas.

**Palabras Clave:** categorización de textos, redes neuronales artificiales, Perceptrón multicapa.

### **1 Introducción**

La clasificación automática de documentos ha ganado gran interés en los últimos tiempos, pues el aumento exponencial de la información disponible en formato digital durante los últimos años y las expectativas de crecimiento futuro, hacen necesaria la organización de todo este contenido con el fin de mejorar la búsqueda y acceso a la información, lo que se ha convertido en una difícil tarea la clasificación manual de documentos [1]. Con este fin adquiere importancia el desarrollo de un sistema inteligente para la clasificación automática de textos.

El desarrollo de la ciencia y tecnología viene avanzando aceleradamente la información en cada área de conocimiento se incrementa en forma exponencial, y su tratamiento así como su almacenamiento se hace más compleja. El explosivo crecimiento de la información disponible en documentos digitales en el área de informática y sistemas, ha hecho necesario desarrollar nuevos instrumentos y herramientas que faciliten la realización de procesos de búsquedas de forma eficiente y efectiva así como la administración de estos recursos. Es frecuente que para facilitar la búsqueda de información se proceda a la categorización de los documentos en un conjunto acotado de clases o categorías. Estas clases permiten representar áreas específicas del conocimiento y son generalmente consolidadas por expertos [2].

El contexto del presente trabajo tiene como finalidad crear un sistema inteligente que permita categorizar automáticamente documentos utilizando sistemas expertos [3] de tal manera que realice un proceso más rápido en un menor tiempo y costo. Para el desarrollo de la investigación se utilizaron datos disponibles del repositorio de la Universidad Autónoma de la ciudad de México (Red de Revistas Científicas de América

Latina y el Caribe, España y Portugal) en distintos formatos (Microsoft Word, PDF, texto plano) que han sido utilizados para entrenar y posteriormente evaluar los resultados obtenidos en cada grupo de entrenamiento. Cabe indicar que la categorización de los documentos van a estar categorizados en 14 categorías las cuales son física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso.

## 2. REDES NEURONALES

### 2.1. Definición

Existen numerosas formas de definir a las redes neuronales; desde las definiciones cortas y genéricas hasta las que intentan explicar más detalladamente qué son las redes neuronales. Por ejemplo:

*Una nueva forma de computación, inspirada en modelos biológicos.*

*Un modelo matemático compuesto por un gran número de elementos procesales organizados en niveles [4].*

*Un sistema de computación compuesto por un gran número de elementos simples, elementos de procesos muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas [4].*

*Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico. [4].*

Definiciones según algunos autores:

- Haykin, S.: "Una red neuronal es un procesamiento distribuido masivamente paralelo que tiene una tendencia natural para almacenar conocimiento empírico y hacerlo disponible para el uso [5]. Recuerda al cerebro en dos aspectos:

1. Conocimiento se adquiere por la red a través de un proceso de aprendizaje.
2. Las conexiones interneurónicas se conocen como pesos sinápticos y se usan para almacenar el conocimiento."

- Zurada, J.M.: "Sistemas de redes neuronales artificiales, o redes neuronales son sistemas celulares físicos que puedan adquirir, almacenar y usar conocimiento empírico [5]."

- El concepto de Red Neuronal Artificial está inspirado en las Redes Neuronales Biológicas. Una Red Neuronal Biológica es un dispositivo no lineal altamente paralelo, caracterizado por su robustez y su tolerancia a fallos. Sus principales características son las siguientes:

- Aprendizaje mediante adaptación de sus pesos sinápticos a los cambios en el entorno.

- Manejo de imprecisión, ruido e información probabilística.
- Generalización a partir de ejemplos.

## 2.2. Elementos de una red neuronal

A continuación se puede ver en la figura 1, un esquema de una red neuronal:

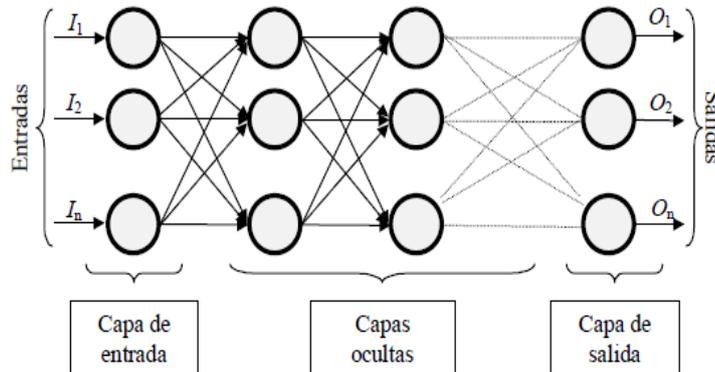


Fig. 1. Ejemplo de una red totalmente conectada

La misma está constituida por neuronas interconectadas y arregladas en tres capas (esto último puede variar). Los datos ingresan por medio de la “capa de entrada”, pasan a través de la “capa oculta” y salen por la “capa de salida”. Cabe mencionar que la capa oculta puede estar constituida por varias capas.

La distribución de neuronas dentro de la red se realiza formando niveles o capas, con un número determinado de dichas neuronas en cada una de ellas. A partir de su situación dentro de la red, se pueden distinguir tres tipos de capas:

- De entrada: es la capa que recibe directamente la información proveniente de las fuentes externas de la red.
- Ocultas: son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de redes neuronales, están encargadas de realizar el trabajo de la red.
- De salidas: transfieren información de la red hacia el exterior.

## 2.3. Redes Neuronales Multicapa con aprendizaje supervisado.

Las redes multicapa se forman por un conjunto de redes de capa simple en cascada unidas por pesos, donde la salida de una capa es la entrada de la siguiente capa. Generalmente son capaces de aprender funciones que una red de capa simple no puede aprender, por lo que ofrecen mejores capacidades computacionales. Para que este incremento en poder computacional sea tal, tiene que existir una función de activación no lineal entre las capas, por lo que generalmente se utilizará una función de activación sigmoidea en detrimento de la lineal o umbral [6].

El Perceptrón Multicapa es un tipo de Red que está compuesta por varios perceptrones y que permite clasificar adecuadamente más de dos clases. Estas neuronas están organizadas mediante capas, las cuales transmitan la información de capa en capa. La característica de este tipo de Red es que sus conexiones están hechas de atrás hacia adelante y que además las neuronas de la misma capa no se relacionan entre sí.

Las redes multicapas son aquellas que disponen de un conjunto de neuronas agrupadas en varios (2, 3, etc.) niveles o capas.

Para calcular la salida de una red multicapa se debe hacer de la misma manera que en las redes de capa simple, teniendo en cuenta que las salidas de una capa son las entradas de la siguiente capa.

### 2.3.1. Arquitectura

La arquitectura de este tipo de red se caracteriza porque tiene todas sus neuronas agrupadas en distintos niveles llamados capas. El primer nivel corresponde a la capa de entrada que se encarga únicamente de propagar por el resto de la red las entradas recibidas.

El último nivel es el de la capa de salida. Se encarga de proporcionar los valores de salida de la red. En las capas intermedias denominadas capas ocultas se realiza un procesamiento no lineal de los patrones recibidos.

Las conexiones del Perceptrón multicapa son hacia adelante. Generalmente todas las neuronas de un nivel se conectan con todas las neuronas de la capa inmediatamente posterior. A veces dependiendo de la red se encuentran conexiones de neuronas que no están en niveles consecutivos, o alguna de las conexiones entre dos neuronas de niveles consecutivos no existe, es decir el peso asociado a dicha conexión es constante e igual a cero. Además, todas las neuronas de la red tienen un valor umbral asociado.

El Perceptrón multicapa es una red formada por una capa de entrada, al menos una capa oculta y una de salida su estructura se muestra en la Figura 2

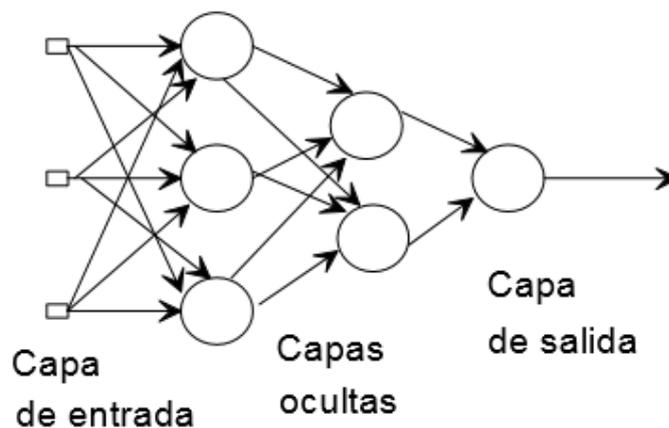


Fig. 2. Estructura de un Perceptrón multicapa

Este modelo se compone de la siguiente manera:

- o Capa de entrada: sólo se encarga de recibir las señales de entrada y propagarla a la siguiente capa.
- o Capa de salida: proporciona al exterior la respuesta de la red para cada patrón de entrada.
- o Capas ocultas: realizan un procesamiento no lineal de los datos de entrada.

### 3. RESULTADOS

#### 3.1. FASE DE SELECCIÓN O ANÁLISIS

##### 3.1.1. Selección del repositorio de estudio

Para la utilización del conjunto de documentos se ha hecho un estudio de varios repositorios como se muestra en la Tabla 1, luego de haber analizados sus ventajas y desventajas de cada uno de ellos se escogió el repositorio que más se acopló con el tema de investigación, llegando a la conclusión que el repositorio para ser utilizado es el de la Red de Revistas Científicas de América Latina y el Caribe, España y Portugal de la Universidad autónoma de México [7]. Este repositorio cuenta con una gran cantidad de documentos digitalizados como es 338 816 documentos compuestos por: artículos científicos a texto completos, revistas científicas y fascículos.

**Tabla 1.** Análisis del corpus de datos.

Repositorios	Tipos de documentos	Ventajas	Desventajas
<b>Repositorio de Ciencia - Science Library</b>	Cuenta con libros de divulgación científica y artículos	Tiene un enlace de descarga total de documentos digitalizados	Es un repositorio muy pequeño de documentos Metadatos incompletos
<b>Biblioteca Virtual Universal</b>	- Cuenta con obras clásicas y estudios científicos. - Trabajos de divulgación científica y de investigación	Contiene un índice medio de documentos a requerir	La mayoría de los metadatos son incompletos
<b>Biblioteca UtpI</b>	Tesis doctorales, revistas y libros	Solo trabaja con documentos de su propia biblioteca	La estructura de sus metadatos tienen muy poca información
<b>Repositorio Red de Revistas Científicas de América y el Caribe, España y Portugal</b>	Contiene artículos científicos a texto completos, revistas científicas y fascículos	Existe una gran numero y variedad de documentos digitalizados	Dos de cada 20 documentos no contienen metadatos completos

Como se ha expuesto anteriormente una de las razones para el uso de este repositorio es que sus metadatos tiene una estructura completa como son autor, título, palabras claves, publicación, url entre otros. Cada uno de estos datos permiten dar mayor información sobre el documento, es por ello que al seleccionar el repositorio se tuvo en cuenta los metadatos del corpus ya que posteriormente estos serán de gran ayuda para dicha clasificación.

### 3.1.2. Identificación de los metadatos a utilizar

Los metadatos son datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar. En otras palabras los metadatos son datos sobre los datos, es por ello que luego del análisis de los repositorios se extraerán los metadatos que se van a emplear para la identificación de los documentos, los cuales estos datos son los siguientes como se muestra en la tabla 2.

**Tabla 2.** Metadatos utilizados para un documento

Autores
Título
Descripción
Fecha de publicación
Palabras Clave
Lenguaje
Url

### Desarrollo del algoritmo para la lectura de los metadatos.

El siguiente código muestra el método para la lectura de metadatos el cual permite extraer los metadatos de título, autor, palabras claves, año, url, descripción, de un cierto documento del repositorio Red de Revistas Científicas de América Latina y el Caribe, España y Portugal. Los metadatos como título, palabras claves y descripción servirán como entrada para la red, a partir de la unión de estos datos de entrada se obtiene un texto lematizado [8], y el resultado determinará cuál es la salida o a que categoría pertenece el documento.

```

public boolean cargarLibroWeb(String urls){
    Document doc;
    try {
        URL url = new URL(urls);
        URLConnection conexion = url.openConnection();
        conexion.connect();
        InputStream is = conexion.getInputStream();
        BufferedReader br = new BufferedReader(new InputStreamReader(is));
        char[] buffer = new char[1000];
        int leído;
        List<String> cont = new ArrayList<String>();
        String aux = null;
        while ((leído = br.read(buffer)) > 0) {
            aux = aux + new String(buffer, 0, leído);
        }
        doc = Jsoup.parse(aux);
        String title = doc.title();
        List<Element> autor = new ArrayList<>();
        try {
            autor = doc.select("meta[name=citation_author]");
        }
    }
}

```

```

    } catch (Exception e) {
    }

    String desc = doc.select("meta[name=description]").first().attr("content");

    String keybor="";
    try {
        keybor = doc.select("meta[name=keywords]").first().attr("content");
    } catch (Exception e) {
        keybor = doc.select("meta[name=citation_keywords]").first().attr("content");
    }

    String urlPag = doc.select("meta[name=citation_pdf_url]").first().attr("content");
    String autores = "";
    if (!autor.isEmpty()) {
        for (Element element : autor) {
            autores = autores + " - "+element.attr("content");
        }
    }
    String ano = "";
    try {
        ano =
doc.select("meta[name=citation_publication_date]").first().attr("content");
    } catch (Exception e) {
        ano = "";
    }
    Libro l=new Libro();
    l.setTitulo(title);
    l.setDescripcion(desc);
    l.setFechaPublicacion(ano);
    l.setAutor(autores);
    l.setPalabrasClave(keybor);
    l.setUrl(urlPag);
}
}

```

### 3.2. FASE DE DISEÑO

#### 3.2.1. Diseño de la arquitectura de la red

Debido a que se va a construir una red multicapa supervisada de antemano se debe tener en claro cuáles son las salidas de la red, esta red está aplicada a la clasificación de documentos en 14 categorías las cuales son: física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso, estas categorías se las menciona luego de hacer un estudio y una revisión exhaustiva de cada uno de los documentos del repositorio .

Para determinar que un documento pertenezca a una categoría se tuvo que asignar a cada una de ellas un vocabulario, es decir cada categoría cuenta con subcategorías que describen de la mejor manera a la categoría. Para ellos usamos el sistema de clasificación denominado Sistema de Clasificación Dewey [9], el cual constituye listas estructuradas de términos (conceptos) que representan de forma univoca el contenido conceptual de los documentos, además es un sistema que cuantifica la relevancia de un término para describir a una categoría.

Para llegar a concluir la categoría se usa la relación que existen categoría y glosario, es decir la entrada de la red será comparada con cada glosario o vocabulario de cada categoría, de esta forma se irán asignado pesos a cada categoría llegando a determinar la categoría a la que corresponde.

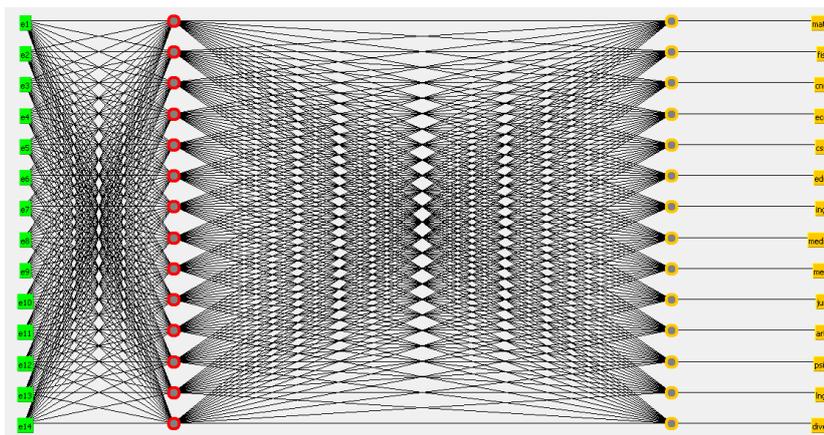
```

public String[] neuralNetworkInput(List<String> words){
    String[] entrada = {"0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0",
"0"};

    for (String word : words) {
        Glosario gosa =glosarioList.obtenerGlosarioPorPalabra(word);
        if (gosa!=null) {
            List<Categoria> cats = gosa.getCategorias();
            for (Categoria categoria : cats) {
                int indice = Integer.parseInt(""+categoria.getId()-1);
                float aux = Float.parseFloat(entrada[indice]);
                if (aux!=0) {
                    aux = (float) (aux + 0.1 - 0.01);
                }else{
                    aux = (float) (aux + 0.1);
                }
                entrada[indice]= ""+aux;
            }
        }
    }
    return entrada;
}

```

En la figura 3 muestra la estructura de la red y sus conexiones, la capa de entradas estarán representadas por los datos de título, palabras claves y descripción y la capa de salida estará constituida por las 14 categorías. El modelo final corresponde a una red con 14 entradas y 14 salidas.



**Fig. 3.** .Resultado de la red en Weka

### 3.2.2. Aprendizaje de la red

El aprendizaje supervisado se caracteriza por conocer la respuesta que debería tener la red frente a una determinada entrada. De esta manera se compara la salida deseada con la salida de la red y si existen discrepancias se ajusta iterativamente los pesos. Así la etapa de aprendizaje tiene por objeto hacer mínimo el error entre la salida brindada por la red y la salida deseada o verdadera

### 3.3. FASE DE DESARROLLO

#### 3.3.1. FASE DE ENTRENAMIENTO

Estos resultados están basados en los dos conjuntos de entrenamiento de 1000 y 3000 documentos, a continuación se presentan y se discuten los resultados de cada uno de ellos. Para desarrollar las pruebas correspondientes de entrenamiento de clasificación de los textos según las categorías se empleó Weka(Waikato Environment for Knowledge Analysiscorpu), lo cual permite la verificación y constancia de los resultados de la correcta y mala clasificación que se pueda estar dando.

#### Muestra 1

El primer conjunto de datos consta de 1000 documentos, se observa que sus resultados tienen un alto porcentaje de clasificación correcta, lo cual indica que hay una buena clasificación, a continuación se muestran sus resultados en porcentajes de la versatilidad de clasificación y margen de error.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances      1152           97.4619 %
Incorrectly Classified Instances     30            2.5381 %
Kappa statistic                     0.9717
Mean absolute error                  0.011
Root mean squared error              0.0629
Relative absolute error              8.5963 %
Root relative squared error          24.8401 %
```

Al analizar la muestra 1 las instancias de clasificación correcta muestran un 97% de documentos bien clasificados y un 3% una mínima parte de margen de error, indicando un mayor rango de clasificación correcta de documentos.



La matriz de confusión muestra el tipo de las predicciones correctas e incorrectas sobre el conjunto de documentos. La misma permite comprender en qué sentido se equivoca la red al intentar clasificar el nuevo conjunto de documentos. En el gráfico de esta matriz las predicciones correctas están representadas sobre la diagonal.

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	<-- classified as
a = mate	164	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b = fis	0	198	0	0	0	0	0	0	0	0	0	0	0	0	0
c = cnn	0	0	98	0	0	0	0	0	0	0	0	0	0	0	0
d = eco	0	0	0	120	0	0	0	0	0	0	0	0	0	0	1
e = css	0	0	0	1	141	0	0	0	0	0	0	0	0	0	0
f = edu	0	1	0	0	0	109	0	0	0	0	0	0	0	0	0
g = ing	0	2	1	0	0	0	13	0	0	0	0	1	0	0	0
h = medioA	1	0	1	0	0	0	0	69	0	0	0	0	1	1	0
i = med	0	1	1	0	0	0	0	0	24	0	2	0	0	0	0
j = jur	0	0	0	0	1	1	0	0	0	51	0	0	0	2	0
k = art	0	0	1	0	2	0	0	0	0	0	20	0	0	0	0
l = psig	2	2	2	0	0	0	0	0	1	0	0	62	0	0	0
m = lng	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0
n = diver	1	0	0	0	0	0	0	0	0	0	0	0	0	62	0

## Muestra 2

El segundo conjunto de datos se llevó a cabo con 3000 documentos, al incrementar la cantidad de documentos se creó conveniente agregar nuevas palabras al vocabulario con la finalidad de que el margen de asertividad no decaiga.

```
=== Evaluation on training set ===  
=== Summary ===  
  
Correctly Classified Instances      2223      89.3489 %  
Incorrectly Classified Instances    265       10.6511 %  
Kappa statistic                    0.8819  
Mean absolute error                 0.023  
Root mean squared error             0.1109  
Relative absolute error              17.8743 %  
Root relative squared error         43.7063 %
```

El resultado de este nuevo grupo de entrenamiento muestra un valor del 90% de instancias correctas y con un margen de error del 10%, a diferencia del primer grupo se redujo un mínimo porcentaje el margen de asertividad debido a la gran cantidad de documentos a ser analizadas, es decir que a medida que el aumento el número de documentos a ser analizados se deberá incrementar el vocabulario.



La matriz de confusión muestra el tipo de las predicciones correctas e incorrectas sobre el conjunto de datos. La misma permite comprender en qué sentido se equivoca la red al intentar clasificar los nuevos textos.

```

== Confusion Matrix ==

  a  b  c  d  e  f  g  h  i  j  k  l  m  n  <-- classified as
153 0  4  1  2  4  0  0  0  0  1  0  0  0 | a = mate
  1 198 2  4  0  3  6  1  0  1  5  0  2  0 | b = fis
  2  0 121 3  2  5  1  8  0  0  2  1  1  0 | c = cnn
  1  0  1 113 2  4  0  0  1  2  2  1  0  0 | d = eco
  3  0  0  2 146 8  1  2  0  3  7  1  1  0 | e = css
  0  0  1  1  1 357 2  1  1  3  4  0  2  0 | f = edu
  4  4  1  1  2  9 354 3  1  2  7  1  5  0 | g = ing
  2  2  3  2  3  4  4 267 3  0  2  0  6  1 | h = medioA
  6  0  3  1  0  3  2  2 165 0  1  0  0  0 | i = med
  4  0  0  2  2  4  1  0  0 104 4  1  1  0 | j = jur
  3  1  0  0  0  2  1  2  0  0 101 0  2  0 | k = art
  5  1  3  0  0  2  1  1  0  2  1 22  1  0 | l = psig
  0  0  1  1  0  1  0  0  0  0  0  0 120 0 | m = lng
  1  0  1  1  0  1  0  0  0  0  1  0  0  2 | n = diver

```

### 3.4. CONCLUSIONES

La utilización de la herramienta Weka permite la ejecución de pruebas de entrenamiento de Redes Neuronales, con la finalidad de predecir el área de pertenencia de un texto.

Para garantizar una buena clasificación de documentos se creyó necesario en la muestra 2 se incrementa un mayor número de palabras al vocabulario con la finalidad de dar una mejor exactitud en la clasificación de los documentos.

Definir correctamente el vocabulario de cada una de las categorías hace que la clasificación tenga un buen porcentaje de aciertos logrando que el documento sea asignado correctamente a su categoría, de tal medida que el desempeño de la exactitud de la red mejora de acuerdo al tamaño del vocabulario.

La utilización de metadatos fue de gran ayuda para lograr mejores resultados a la hora de la representación, localización y recuperación de recursos electrónicos.

### Referencias

- [1] Cortez Vasquez,A y Rojas Lazo,O y Calmet Agnelli,R ., “Categorización de Textos mediante Máquinas de Soporte Vectorial,” Revistas Signos, pp. 1-24, 2011
- [2] Mendoza, M y Ortiz, I y Rojas, V., “Categorización de texto en bases documentales a partir de modelos computacionales liviano,” Revista de investigación de Sistemas e Informática,vol 10,N. 1,pp. 2-12,Enero-Junio 2013.
- [3] Martínez Pérez, P. y Colarte, J. (feb. 2007) Multimedia para discapacitados. Presentada en: Congreso y Feria Internacional Informática 2007 [en línea]. Disponible en: <http://www.informaticabana.cu/eventovirtual/educacion/discapacitados.pdf>
- [4] Match, D.J., “Redes Neuronales: Conceptos Básicos y Aplicaciones”, Facultad Regional Rosario, Dept. Ingeniería Química, Universidad Tecnológica Nacional, marzo 2001, [En línea] Link [http://www.fro.utn.edu.ar/repositorio/catedras/quimica/5\\_anio/orientadora1/monogriais/match-redesneuronales.pdf](http://www.fro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monogriais/match-redesneuronales.pdf)
- [5] “Redes Neuronales”, [En línea] LINK <http://www.usmp.edu.pe/publicaciones/boletin/fia/info32/pag4.htm>

- [6] Cárdenas Almeida, R., "Inteligencia Artificial", [En línea] Link  
[http://www2.ulpgc.es/hege/almacen/download/38/38584/practica\\_ia\\_2.pdf](http://www2.ulpgc.es/hege/almacen/download/38/38584/practica_ia_2.pdf)
- [7] Red de Revistas Científicas de América Latina y el Caribe, España y Portugal,2013. [En línea].  
 Disponible en:  
<http://www.redalyc.org>
- [8] "Snowball"; Link: <http://snowball.tartarus.org/>
- [9] "Introducción al sistema de clasificación decimal Dewey" [En línea] Link  
<http://bibliotecamachala.wikispaces.com/file/view/Clasificacion+decimal+dewey.pdf>

## Anexo 2: Sumarios de la clasificación de Dewey

- Primer Sumario Las diez principales clases

<b>000</b>	<b>Generalidades</b>
<b>100</b>	<b>Filosofía y psicología</b>
<b>200</b>	<b>Religión</b>
<b>300</b>	<b>Ciencias sociales</b>
<b>400</b>	<b>Lenguas</b>
<b>500</b>	<b>Ciencias naturales y matemáticas</b>
<b>600</b>	<b>Tecnología (Ciencias aplicadas)</b>
<b>700</b>	<b>Las artes Bellas artes y artes decorativas</b>
<b>800</b>	<b>Literatura y retórica</b>
<b>900</b>	<b>Geografía e historia</b>

- Segundo Sumario las cien divisiones

## Segundo Sumario\*

### Las cien divisiones

<b>000 Generalidades</b>	<b>500 Ciencias naturales y matemáticas</b>
010 Bibliografía	510 Matemáticas
020 Biblioteología y ciencias de la información	520 Astronomía y ciencias afines
030 Obras enciclopédicas generales	530 Física
040	540 Química y ciencias afines
050 Publicaciones seriadas generales	550 Ciencias de la tierra
060 Organizaciones generales y museología	560 Paleontología Paleozoología
070 Medios noticiosos, periodismo, publicación	570 Ciencias de la vida Biología
080 Colecciones generales	580 Plantas
090 Manuscritos y libros raros	590 Animales
<b>100 Filosofía y psicología</b>	<b>600 Tecnología (Ciencias aplicadas)</b>
110 Metafísica	610 Ciencias médicas Medicina
120 Epistemología, causalidad, género humano	620 Ingeniería y operaciones afines
130 Fenómenos paranormales	630 Agricultura y tecnologías relacionadas
140 Escuelas filosóficas específicas	640 Economía doméstica y vida familiar
150 Psicología	650 Gerencia y servicios auxiliares
160 Lógica	660 Ingeniería química
170 Ética (Filosofía moral)	670 Manufactura
180 Filosofía antigua, medieval, oriental	680 Manufactura para usos específicos
190 Filosofía moderna occidental	690 Construcción
<b>200 Religión</b>	<b>700 Las artes Bellas artes y artes decorativas</b>
210 Filosofía y teoría de la religión	710 Urbanismo y arte paisajístico
220 La Biblia	720 Arquitectura
230 Cristianismo Teología cristiana	730 Artes plásticas Escultura
240 Moral cristiana y teología piadosa	740 Dibujo y artes decorativas
250 Órdenes cristianos e iglesia local	750 Pintura y pinturas
260 Teología social y eclesiástica	760 Artes gráficas Arte de grabar y grabados
270 Historia del cristianismo y de la iglesia cristiana	770 Fotografía y fotografías
280 Confesiones y sectas cristianas	780 Música
290 Religión comparada y otras religiones	790 Artes recreativas y de la actuación
<b>300 Ciencias sociales</b>	<b>800 Literatura y retórica</b>
310 Colecciones de estadística general	810 Literatura norteamericana en inglés
320 Ciencia política	820 Literaturas inglesa e inglesa antigua
330 Economía	830 Literaturas de lenguas germánicas
340 Derecho	840 Literaturas de lenguas romances
350 Administración pública y ciencia militar	850 Literaturas italiana, rumana, retorromana
360 Problemas y servicios sociales; asociaciones	860 Literaturas española y portuguesa
370 Educación	870 Literaturas itálicas Literatura latina
380 Comercio, comunicaciones, transporte	880 Literaturas helénicas Literatura griega clásica
390 Costumbres, etiqueta, folclor	890 Literaturas de otras lenguas
<b>400 Lenguas</b>	<b>900 Geografía e historia</b>
410 Lingüística	910 Geografía y viajes
420 Inglés e inglés antiguo	920 Biografía, genealogía, insignias
430 Lenguas germánicas Alemán	930 Historia del mundo antiguo hasta ca. 499
440 Lenguas romances Francés	940 Historia general de Europa
450 Italiano, rumano, retorromano	950 Historia general de Asia Lejano Oriente
460 Lenguas española y portuguesa	960 Historia general de África
470 Lenguas itálicas Latín	970 Historia general de América del Norte
480 Lenguas helénicas Griego clásico	980 Historia general de América del Sur
490 Otras lenguas	990 Historia general de otras áreas

### Anexo 3: Vocabulario utilizado para cada categoría

Categorías existentes en el sistema: física, matemáticas, ciencias sociales, ciencias naturales, arte, economía, educación, ingeniería, medio ambiente, medicina, jurídica, psicología, lenguaje y diverso.

Esta tabla consta de los vocabularios que se han utilizado para cada una de las categorías.

Inducción	Mate	económico	MedioA
Deducción	Mate	ingeniería	MedioA
Cantidad	Mate	tratado	MedioA
Teoría	Mate	comercio	MedioA
Cuantitativo	Mate	recursos	MedioA
Matemáticas	Mate	social	MedioA
Suma	Mate	huertos	MedioA
Lógica	Mate	organizaciones	MedioA
Miscelánea	Mate	alumnos	MedioA
Resta	Mate	campo	MedioA
Multiplicación	Mate	naturaleza	MedioA
División	Mate	cultivo	MedioA
Operación	Mate	calor	MedioA
Ciencias	Mate	agricultura	MedioA
Análisis	Mate	flores	MedioA
Resultado	Mate	estudios	MedioA
Aritmética	Mate	investigación	Med
Análisis	Mate	archivos	Med
Algebra	Mate	química	Med
probabilidades	Mate	equipos	Med
números	Mate	exámenes	Med
Geometría	Mate	epidemias	Med
mecánica	Fis	animales	Med
Análisis	Fis	veterinario	Med
Operación	Fis	farmacia	Med
Solidos	Fis	doctores	Med
Fluidos	Fis	odontología	Med
Física	Fis	cirugías	Med
Náutica	Fis	paciente	Med
Estructura	Fis	historial	Med
Movimiento	Fis	operaciones	Med
Átomos	Fis	tratamiento	Med
Moderna	Fis	tecnología	Med
Gases	Fis	medicina	Med
Sonido	Fis	médicos	Med

Números	Fis	anatomía	Med
Resultado	Fis	citología	Med
Vibraciones	Fis	medio	Med
Teoría	Fis	tecnología	Med
Luz	Fis	histología	Med
Parabólico	Fis	fisiología	Med
Calor	Fis	humana	Med
Electricidad	Fis	salud	Med
Electrónica	Fis	prevención	Med
Magnetismo	Fis	farmacología	Med
Espacio	Fis	terapeuta	Med
Tiempo	Fis	enfermedades	Med
Fuerza	Fis	cirugía	Med
Fenómeno	Fis	ginecología	Med
Calor	Fis	experimental	Med
Luz	Fis	personas	Med
Electrónica	Fis	auxiliar	Med
Movimiento	Fis	anatomía	Med
Náutica	Fis	clínica	Med
Números	Fis	hospital	Med
Energía	Fis	salud	Med
Teoría	Cnn	ministerio	Med
tierra	Cnn	especialidad	Med
Agricultura	Cnn	pediatría	Med
Vegetales	Cnn	niños	Med
Cultivos	Cnn	sistema	med
Campo	Cnn	nutrición	med
Huertos	Cnn	rayos	med
Producción	Cnn	terapia	med
Biología	Cnn	enfermería	med
tierra	Cnn	jurídico	jur
Ciencias	Cnn	jurisprudencia	jur
Energía	Cnn	leyes	jur
Fenómenos	Cnn	derecho	jur
Vida	Cnn	constitucional	jur
Animales	Cnn	tributario	jur
Genética	Cnn	laboral	jur
Historia	Cnn	social	jur
Ecología	Cnn	economía	jur
Organismos	Cnn	personas	jur
Plantas	Cnn	penal	jur
Aves	Cnn	privado	jur
Anatomía	Cnn	civiles	jur

Humano	Cnn	leyes	jur
Salud	cnn	reglamentos	jur
Biología	cnn	corte	jur
Bioquímica	cnn	tribunal	jur
Genética	cnn	ciencias	jur
Organismos	cnn	universidad	jur
Evolución	cnn	sentencia	jur
Orgánica	cnn	política	jur
Mineralogía	cnn	facultad	jur
Naturaleza	cnn	democracia	jur
Vertebrados	cnn	humano	jur
Laboral	eco	justicia	jur
Financiera	eco	gobiernos	jur
Economía	eco	estado	jur
Socialismo	eco	finanzas	jur
Finanzas	eco	mercantil	jur
Producción	eco	industrial	jur
Macroeconomía	eco	estatutos	jur
Derecho	eco	educativo	jur
Constitucional	eco	estatuto	jur
Comercio	eco	procedimientos	jur
Penal	eco	política	jur
inversión	eco	estado	jur
Investigación	eco	ciencias	jur
Dinero	eco	derechos	jur
Social	eco	legislativo	jur
Números	eco	arte	art
Leyes	eco	decoraciones	art
Finanzas	eco	poesía	art
Transporte	eco	plásticas	art
Vivienda	eco	baile	art
Equipos	eco	ciencias	art
Alimentos	eco	obras	art
Servicios	eco	templos	art
Dinero	eco	museos	art
Gastos	eco	estatuas	art
Ministerio	eco	estado	art
administración	eco	deporte	art
Sociología	css	cristianismo	art
antropología	css	artículos	art
Socialismo	css	eclesiásticos	art
Social	css	esculturas	art
comunidades	css	música	art

Cultura	css	dibujo	art
Política	css	cerámicas	art
Sociedad	css	graficas	art
Gobiernos	css	impresión	art
Geografía	css	carpintería	art
Derechos	css	educación	art
Civil	css	arquitectura	art
Migración	css	decoraciones	art
Legislativo	css	pinturas	art
Economía	css	diccionario	art
Derechos	css	galerías	art
Comercio	css	museos	art
costumbres	css	exposiciones	art
Estado	css	personas	art
Derecho	css	geográfico	art
Leyes	css	paisaje	art
investigación	css	espacio	art
Ciencias	css	humano	art
organismos	css	formas	art
instituciones	css	color	art
Escuelas	edu	leyenda	art
Educación	edu	mitología	art
bibliotecas	edu	religión	art
Primaria	edu	urbanismo	art
secundaria	edu	fotografía	art
Leyes	edu	pigmentación	art
tecnología	edu	escultura	art
Gobiernos	edu	formas	art
Política	edu	tallado	art
publica	edu	dibujo	art
Privada	edu	música	art
Transporte	edu	vidrio	art
Materias	edu	muebles	art
Noticias	edu	escénicas	art
publicaciones	edu	psicología	art
Ministerio	edu	teoría	art
autoridades	edu	percepción	art
Docentes	edu	emociones	psig
estudiantes	edu	movimientos	psig
educadores	edu	impulsos	psig
Escolar	edu	mente	psig
Adultos	edu	mentales	psig
Alumnos	edu	inteligencia	psig

Currículos	edu	subconsciente	psig
Superior	edu	humanos	psig
actividades	edu	personas	psig
Lenguas	edu	estados	psig
Jóvenes	edu	aplicada	psig
Oferta	edu	sueños	psig
Consejo	edu	pensamientos	psig
Ciencias	edu	personas	psig
universidad	edu	tratamiento	psig
Carreras	edu	grupos	psig
Equipos	edu	lógica	psig
computadoras	edu	deducción	psig
investigación	edu	persuasión	psig
Tecnología	edu	epistemología	psig
Religiosa	edu	casualidad	psig
programación	ing	inconsciente	psig
Redes	ing	ética	psig
Sistemas	ing	moral	psig
contabilidad	ing	sensaciones	psig
Usuario	ing	lenguaje	psig
Manual	ing	literatura	psig
Bases	ing	castellano	psig
Modelos	ing	escritura	Ing
investigación	ing	gramática	Ing
información	ing	dialecto	Ing
Maquinas	ing	lingüística	Ing
Lenguaje	ing	lengua	Ing
agrónomos	ing	habla	Ing
Técnico	ing	fonemas	Ing
Análisis	ing	fonología	Ing
Datos	ing	español	Ing
Industrias	ing	abecedario	Ing
estudiantes	ing	ingles	Ing
computación	ing	diccionario	Ing
Cursos	ing	fonética	Ing
Programas	ing	enciclopedias	Ing
aplicaciones	ing	educación	Ing
Proyectos	ing	investigación	Ing
investigación	ing	literatura	Ing
Internet	ing	teatro	Ing
Eléctrica	ing	comunicación	Ing
implementación	ing	castellano	Ing
universidad	ing	escritos	Ing

mantenimiento	ing	cartas	Ing
Mecánica	ing	etimología	Ing
computadoras	ing	ensayos	Ing
Química	ing	poesía	Ing
tecnología	ing	tecnología	Ing
Industrias	ing	discursos	Ing
Náutica	ing	novelas	Ing
Militar	ing	publicaciones	Ing
Civil	ing	bibliografía	Ing
ingenieros	ing	bibliotecas	Ing
Carreteras	ing	organizaciones	Ing
Hidráulica	ing	medios	Ing
Municipal	ing	noticias	Ing
Física	ing	periodismo	Ing
Cantidad	ing	lectura	Ing
Humanos	ing	obras	Ing
Equipos	ing	colecciones	Ing
combustibles	ing	libros	Ing
Cerámica	ing	impresos	Ing
construcción	ing	fenómenos	diver
Población	medioA	escuelas	diver
Estudios	medioA	religión	diver
Cosecha	medioA	biblia	diver
habitantes	medioA	cristianismo	diver
protección	medioA	sectas	diver
Ecología	medioA	comercio	diver
Vida	medioA	transporte	diver
sistemas	medioA	arquitectura	diver
nanotecnología	medioA	manuscritos	diver
Flores	medioA	revistas	mate
Industrias	medioA	revistas	fis
Ambiente	medioA	revistas	cnn
Fenómeno	medioA	revistas	eco
tierra	medioA	revistas	css
Geología	medioA	revistas	edu
Aire	medioA	revistas	ing
Medio	medioA	revistas	Lng
Vida	medioA	revistas	med
Biología	medioA	revistas	Jur
Sociedad	medioA	revistas	Art
Animales	medioA	revistas	Psi
Plagas	medioA	revistas	Diver
Vegetales	medioA	noticias	Diver

conservación	medioA	biblia	Diver
Frutas	medioA	manufactura	Diver
Pesca	medioA	áreas	Diver
Personas	medioA		
Caza	medioA		
Insectos	medioA		
medicina	medioA		
Ciencias	medioA		
Desarrollo	medioA		
Impacto	medioA		

## **Anexo 4: Declaración de Confidencialidad**

**Doris Yadira Jiménez Ochoa**, con CI: **1104605710**; residente a la fecha **Junio del 2015** en la ciudad de **Loja**.

**DECLARA** lo siguiente:

### **PRIMERO: Antecedentes**

Los declarantes han desarrollado el proyecto de fin de carrera titulado: “**DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA CLASIFICACIÓN DE DOCUMENTOS YA DIGITALIZADOS APLICANDO REDES NEURONALES SUPERVISADAS**”, teniendo como Director de Tesis al , Ing. Henry Patricio Paz Arias Mg. Sc.,

### **SEGUNDO: Información Confidencial**

La información referida en pruebas, configuraciones y experimentaciones debe ser utilizada con fines académicos y no con otros propósitos; por lo tanto se considerará siempre como Información Confidencial y/o Sensible en el caso de un uso ilegal.

### **TERCERO: Excepciones**

No será considerada como Información Confidencial:

- La información que era de dominio público o pase a serlo, con posterioridad, por haberse publicado sin intervención ni negligencia del declarante.
- La información que sea accesible por pasantes u otros tesisistas.
- La información revelada por una tercera persona con derecho a divulgarla.

### **CUARTO: Secretos de la Información Confidencial**

El declarante se comprometen a:

- Mantener de forma confidencial y a no revelar a personas ajenas al Ing. Henry Patricio Paz Arias Mg. Sc., toda la información y material de carácter sensible a la que se acceda en el desarrollo del TT, tanto teórico como práctico, salvo con las excepciones antes mencionadas.

- Mantener en absoluta reserva la información o documentos de carácter sensible, a los que tenga acceso como consecuencia de su formación profesional.

**QUINTO: Duración**

La obligación de los declarantes respecto al compromiso de mantener en secreto la Información Confidencial, tendrá una duración indefinida a partir de la fecha de entrega de éste documento.

En Loja, Junio 2015

.....

Doris Yadira Jiménez Ochoa

CI: 1104605710

## Anexo 5: Certificado de traducción del Resumen



Lic. María Isabel Vivanco  
**PROFESORA DEL INSTITUTO  
"FINE-TUNED ENGLISH"**

### **CERTIFICA:**

Que el documento aquí compuesto es fiel traducción del idioma español al idioma inglés del resumen de la tesis titulada **"DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA CLASIFICACION DE DOCUMENTOS YA DIGITALIZADOS APLICANDO REDES NEURONALES SUPERVISADAS"**, de la Señorita DORIS YADIRA JIMENEZ OCHOA, egresada de la carrera de Ingeniería en Sistemas del Área de la Energía, las Industrias y los Recursos Naturales No Renovables de la Universidad Nacional de Loja.

Lo certifica en honor a la verdad y autoriza a la interesada hacer uso del presente en lo que a sus intereses convenga.

Loja, 18 de febrero de 2015

María Isabel Vivanco  
PROFESORA DE F.T.E.  
Registro SENESCYT 1031-07-785801



## Anexo 6: Certificado conferido por la Bibliotecaria del funcionamiento de la Aplicación

	<b>UNIVERSIDAD NACIONAL DE LOJA</b>	
ÁREA DE LA ENERGÍA, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES BIBLIOTECA		
<hr/>		
<b>Téc. Sandra Lucía Castillo Orellana, RESPONSABLE DE LA BIBLIOTECA DEL ÁREA DE LA ENERGÍA, LAS INDUSTRIAS Y LOS RECURSOS NATURALES NO RENOVABLES DE LA UNIVERSIDAD NACIONAL DE LOJA</b>		
<b>CERTIFICA:</b>		
<p>Haber recibido la capacitación del: <b>“DESARROLLO DE UN SISTEMA INTELIGENTE PARA LA CLASIFICACIÓN DE DOCUMENTOS YA DIGITALIZADOS APLICANDO REDES NEURONALES SUPERVISADAS”</b> por parte de la tesista <b>DORIS YADIRA JIMÉNEZ OCHOA</b>, la misma que se llevó a cabo el día jueves 19 de febrero de 2015 en las instalaciones de la Biblioteca del Área de la Energía, las Industrias y los Recursos Naturales no Renovables, así como también se revisó, realizó las pruebas y prácticas del funcionamiento del sistema sin presentar ningún inconveniente. Además este programa puede ser aplicado en las otras bibliotecas de la Universidad.</p>		
<p>Es todo cuanto puedo certificar en honor a la verdad.</p>		
<p>Loja, 5 de marzo de 2015</p>		
		
<b>Téc. Sandra Lucía Castillo Orellana</b> <b>RESPONSABLE DE LA B-AEIRNNR-UNL</b>		
		
<hr/>		
<p>Ciudad Universitaria “Guillermo Falconí Espinosa”, La Argelia Teléfono: 2545310; 2545689; Telefax: 2545691 Loja-Ecuador</p>		

## Anexo 7: Licencia Creative Commons del Normativo



Trabajo de Titulación por Carrera de Ingeniería en sistemas se distribuye bajo una Licencia Creative Commons Atribución-NoComercial 4.0 Internacional.