VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSTGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: "UN PRIMER ENFOQUE PARA EL RECONOCIMIENTO DE LENGUAJE DE SEÑAS BASADO EN UN GUANTE INTELIGENTE QUE UTILIZA TÉCNICAS DE MACHINE LEARNING"

AUTOR: GODOY TRUJILLO, PAMELA ESTEFANÍA

DIRECTOR: PhD. DELGADO RODRÍGUEZ, RAMIRO NANAC

SANGOLQUÍ

2017

DEPARTAMENTO DE CIENCAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN

E INTELIGENCIA DE NEGOCIOS

CERTIFICADO DEL DIRECTOR

Certifico que el trabajo de titulación, "***UN PRIMER ENFOQUE PARA EL RECONOCIMIENTO DE LENGUAJE DE SEÑAS BASADO EN UN GUANTE INTELIGENTE QUE UTILIZA TÉCNICAS DE MACHINE LEARNING***" realizado por la Ing. **GODOY TRUJILLO PAMELA ESTEFANÍA**, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar a la Ing. **PAMELA ESTEFANÍA GODOY TRUJILLO** para que lo sustente públicamente.

Sangolquí, 27 de noviembre del 2017

Ing. Ramiro Delgado, PhD

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE

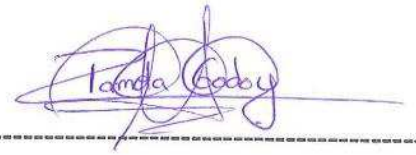
INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

OFICIO DE AUTORÍA DE RESPONSABILIDAD

Yo, ***PAMELA ESTEFANÍA GODOY TRUJILLO***, con cédula de identidad N°1003294566, declaro que este trabajo de titulación "***UN PRIMER ENFOQUE PARA EL RECONOCIMIENTO DE LENGUAJE DE SEÑAS BASADO EN UN GUANTE INTELIGENTE QUE UTILIZA TÉCNICAS DE MACHINE LEARNING***" ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 27 de noviembre del 2017



Ing. PAMELA ESTEFANÍA GODOY TRUJILLO

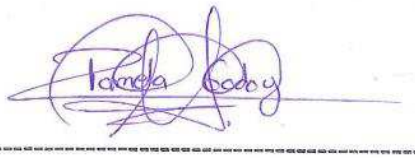
C.C.1003294566

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

OFICIO DE AUTORIZACIÓN

Yo, ***PAMELA ESTAFANÍA GODOY TRUJILLO***, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación ***"UN PRIMER ENFOQUE PARA EL RECONOCIMIENTO DE LENGUAJE DE SEÑAS BASADO EN UN GUANTE INTELIGENTE QUE UTILIZA TÉCNICAS DE MACHINE LEARNING"*** cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 27 de noviembre del 2017



Ing. PAMELA ESTEFANÍA GODOY TRUJILLO
C.C.1003294566

DEDICATORIA

El presente trabajo de grado es fruto de un arduo y constante sacrificio y dedicación, el cual está dedicado a mi familia que siempre ha estado presente en los momentos más difíciles y duros de mi vida y carrera de postgrado; en especial a mis padres y abuelo: Nomberto Fabián Godoy Rosas, Martha Lucía Trujillo Bravo y Victor Manuel Godoy Méndez que me han brindado su apoyo incondicional, además a mis hermanos Darío, Ricardo y a una persona muy especial que siempre ha estado a mi lado apoyándome. Finalmente, a Dios, ya que sin él nada sería posible.

A mis amigos y profesores que fueron parte del ciclo académico, quienes me han impulsado con sus concejos y recomendaciones para la exitosa culminación de este trabajo, con especial énfasis a Ramiro Delgado, que fue el soporte académico en este proceso y encaminó correctamente la investigación.

Gracias a todos ustedes he alcanzado un objetivo más en mi vida.

AGRADECIMIENTO

Aunque no es suficiente, agradecer es muy importante; debido a todas las facilidades y herramientas prestadas, a la ayuda brindada, y el integrado conjunto de conocimientos entregados por parte de la Universidad de la Fuerzas Armadas ESPE para el desarrollo a favor del progreso del país, al ser educada con la visión, misión y valores éticos y morales para ser una futura profesional de calidad.

De igual forma, como parte fundamental del presente trabajo, es el agradecimiento por la labor a las autoridades, docentes y miembros en las diferentes áreas de la maestría en Gestión en Sistemas de Información e Inteligencia de Negocios.

A mi director de tesis, Ramiro Delgado por su colaboración y ayuda durante el desarrollo de esta investigación, factor importante para culminar con éxito mi propósito.

A Mauricio Campaña y Paúl Rosero que en lo particular y personal influyeron y guiaron en el trabajo de investigación, tanto en el anteproyecto como en el desarrollo de la tesis con sabiduría y paciencia.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El presente trabajo de investigación trata acerca del análisis de una base de datos por medio de algoritmos de aprendizaje de máquina y la selección de prototipos de los datos obtenidos de un guante electrónico traductor de señas básicas enfocado especialmente para las personas que presentan una capacidad especial, en este caso personas con discapacidad auditiva y de lenguaje; se utilizará un sistema electrónico inteligente con la capacidad de detectar un número de signos del idioma. En el sistema electrónico se usará un sensor flexible en cada dedo, mismos que se utilizan para recolectar datos (se coloca en la mano derecha de la persona para obtener información de los números del 0 al 9 en el lenguaje de señas). Los datos obtenidos se analizan a través de un esquema que involucra las siguientes etapas: Balanceo de datos con Kennard-Stone (KS), selección de prototipos con algoritmo evolutivo (CHC) y procedimiento de optimización para reducción decremental (DROP3). Consecutivamente, el algoritmo K-Nearest Neighbours (KNN) se utiliza para la clasificación numérica. Este trabajo presenta un análisis de los considerados mejores algoritmos de clasificación de prototipos (DROP3 y CHC) y así poder determinar la adecuada para nuestro conjunto de datos. El principal objetivo del trabajo es usar las nuevas tecnologías del ambiente del Big Data para desarrollar y probar un algoritmo que consiga reducir un dataset de clasificación compuesto por muchos ejemplos a pocos prototipos que los representen sin perder calidad para ayudar a los sistemas de clasificación a enfrentarse a datasets grandes.

PALABRAS CLAVES:

- **SELECCIÓN DE PROTOTIPOS**
- **LENGUAJE DE SEÑAS**
- **KNN**
- **DROP3**
- **CHC**

ABSTRACT

The present investigation deals with the analysis of a database by means of machine learning algorithms and the selection of prototypes of the data obtained from an electronic translator glove basic sign language focused especially for people who have a special ability, in this case people with hearing disabilities and language; it will be used an intelligent electronic system with the ability to detect a number of language signs. In the electronic system will be used a flexible sensor on each finger, which is used to collect data (placed on the right hand of the person to obtain information from numbers 0 to 9 in sign language). The data obtained are analyzed through a scheme that involves the following stages: Data balancing with the Kennard-Stone (KS), selection of prototypes with evolutionary algorithm (CHC) and decremental reduction optimization procedure (DROP3). Consecutively, the algorithm K-Nearest Neighbors (KNN) is used for the numerical rating. This paper presents an analysis of the considered to be the best classification algorithms of prototypes (DROP3 and CHC) and thus be able to determine the appropriate for our data set. The main objective of this work is to use new technologies of the Big Data environment to develop and test an algorithm that get reduce a dataset classification composed of many examples to a few prototypes that represent them without losing quality to help the classification systems to face large datasets.

KEYWORDS:

- **SELECTION OF PROTOTYPES**
- **SIGN LANGUAGE**
- **KNN**
- **DROP3**
- **CHC**

CAPÍTULO I

INTRODUCCIÓN AL PROBLEMA DE INVESTIGACIÓN

Es este capítulo se describe la situación del problema, formulación del problema, objetivos, justificación e importancia, localización geográfica del proyecto, hipótesis, tipo de investigación, población y muestra, métodos de investigación, presupuesto y cronograma de actividades, de una forma más detallada.

1.1. INTRODUCCIÓN A LA SITUACIÓN PROBLÉMICA

A pesar de la gran variedad de lenguajes dactilológicos (comunicación mediante el uso de los dedos de la mano) existentes, y de la gran cantidad de signos que estos poseen, una limitación de gran importancia que continúa presente en el desarrollo de personas que tienen algún tipo de discapacidad auditiva y del habla es la comunicación verbal (López, Rodríguez, Zamora, y Sosa, 2008, p. 20). Este acto tiene una afectación demasiado grave en varios ambientes, especialmente virtuales porque ahí predomina la comunicación verbal. Por todo lo antes mencionado, este problema tiene un efecto directo en la inserción de dichas personas en varios ámbitos (familiar, social y laboral).

Las personas con capacidades especiales en nuestro país actualmente continúan sufriendo algún tipo de discriminación y prejuicio por varios sectores de la sociedad, dando origen a la división de grupos sociales (Flora Davis, 2016, p. 51); esto se podría de alguna manera contrarrestar desarrollando un prototipo de comunicación electrónico que se pueda adquirir a bajo costo en cualquier tienda comercial o centro médico especializado. La inexistencia de sistemas electrónicos de comunicación en el mercado es uno de los problemas detectados para estos casos y también en el desconocimiento de los diferentes procesos de aprendizaje para

ampliar la comunicación de estas personas dentro de un contexto familiar y social (Nooritawati, 2012).

(Planificación, 2013) Indica que:

> "El Plan Nacional del Buen Vivir conjuntamente con el Plan Nacional de Ciencia y Tecnología tienen como objetivo fomentar el cambio de la matriz productiva con el desarrollo de propias tecnologías que den por satisfecho las necesidades de los sectores productivos del país, y de esta manera aprovechar para que los procesos de investigación e innovación se dinamicen en busca de la mejora de productos, servicios y del control de los mismos, matrices de tensiones del país enfatizan las problemáticas evidentes por zonas sobre tecnología y sus escasas aplicaciones en la salud, educación, agricultura, entre otras".

Los datos obtenidos de sistemas que anteriormente no podían conectarse a redes de información hoy tienen una generación muy creciente, y por ello es necesario generar un control de la misma, de cierta forma que el tráfico sea el más eficiente posible, permitiendo generar toma de decisiones adecuadas a base de datos reales. La reducción de dimensionalidad (RD) permite percibir los datos de una forma simple y compacta, ya que al representar un conjunto de datos capturados en el tiempo y espacio de alta dimensionalidad aumenta la complejidad de entendimiento del usuario (Lee, 2007).

Los MRD y clasificación de los datos son capaces de simplificar la descripción del conjunto de datos que puedan representar grandes volúmenes de información en tiempos óptimos, evidenciando las mismas propiedades de los datos complejos. Los resultados favorecen a la compresión, eliminación de redundancia y mejora los procesos e implementación de algoritmos de machine learning a un costo computacional menor, para obtener un mejor análisis con un reconocimiento efectivo de patrones, considerando un número inferior de dimensiones (Dashun, 2015).

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo se puede elegir el o los mejores grupos representativos de datos para que pueda reaccionar de mejor forma el método clasificador de aprendizaje de máquina?

1.3. OBJETIVOS

1.3.1. Objetivo general

Optimizar el procesamiento de datos del lenguaje de señas procedentes de un guante inteligente mediante la aplicación de un algoritmo de aprendizaje de máquina para mejorar el rendimiento del sistema electrónico.

1.3.2. Objetivos específicos

1) Realizar una revisión bibliográfica que permita obtener información sobre los temas de investigación.
2) Analizar los elementos electrónicos necesarios para sensar los datos del usuario con el fin de digitalizar las señales y obtener la base de datos.
3) Establecer un análisis de los algoritmos de aprendizaje de máquina utilizados para determinar el conjunto de datos adecuados y mediante reducción de dimensionalidad obtener los que mejor representen la relación de los datos.
4) Realizar pruebas de verificación mediante la validación cruzada (en software) y número de éxito del signo realizado por una persona al utilizar el guante desarrollado para conocer el rendimiento del clasificador.

1.4. JUSTIFICACIÓN E IMPORTANCIA

Con el adelanto y desarrollo de la tecnología existe una gran cantidad de ramas o variantes que cada día cobran fuerza a medida que se investiga sobre ellas; tal es el caso de la aplicación de la electrónica. El prototipo del Guante Inteligente de Señas Básicas para personas con discapacidad auditiva y de lenguaje; permitirá ayudar a mejorar su calidad de vida y sobre todo brindarles nuevas oportunidades de inclusión en el ámbito social, educativo y sobre todo en su entorno familiar; este prototipo permitiría mejorar la capacidad de comunicación dentro del entorno que los rodea.

Según el Plan Nacional del Buen Vivir y el Plan Nacional de Ciencia y Tecnología uno de los objetivos que se plantea el estado es reducir la brecha tecnológica para que los estudiantes se vinculen con aspectos de mejora social, y que según el planteamiento de mejorar la calidad de vida de las personas con capacidades especiales en el objetivo 1 se define: Auspiciar la igualdad, la cohesión y la integración social y territorial en la diversidad. (Vivir, 2016)

Según el Art.87 de la Ley Orgánica de Discapacidades determina al MIES como autoridad nacional encargada de la inclusión económica y social para las personas con capacidades especiales, teniendo el objetivo de capacitar a las familias que tienen bajo su cuidado a estas personas, en el buen trato y atención que deben prestarles. En el Ecuador existen aproximadamente 16’221.610 de personas, las cuales el 5,6% de la población ecuatoriana presenta algún tipo de capacidad especial, es decir, alrededor de 908.320 personas, las cuales el 46,6% son hombres y el 53,4% son mujeres (Ciudadano, 2014).

Actualmente el código laboral protege y da la oportunidad para que personas con capacidades diferentes puedan insertarse en el mundo productivo, esto conduce a que muchos jóvenes y adultos mejoren sus condiciones de vida, puedan de alguna manera ayudar o mantener a su familia y apoyar al desarrollo del país a través de un trabajo productivo.

Autores como (Alwakeel, Alhalabi, Aggoune, & Alwakeel, 2015), (Celada-Funes, Román, Asensio, & Beferull, 2014) explican que las redes de sensores son nuevas tecnologías para el beneficio de personas orientadas al Internet de las Cosas, estudios como (Zafer, Turgay Tugay, & Cho, 2010) son desarrollados con el objetivo de lograr sistemas electrónicos que no saturen el canal en su transmisión de datos y permita integrar algoritmos de clasificación de información. Todos los autores coinciden en que las redes de sensores son una gran solución para la adquisición de datos en lugares de difícil acceso y de cambios rápidos de estados, todavía se encuentra abierto el uso de los diferentes algoritmos de machine learning para optimizar su funcionamiento y alargar su tiempo de vida.

El avance de tecnología se evidencia en la integración de los procesos cotidianos del ser humano, que mediante sensores, aplicaciones móviles, páginas web y sistemas integrados en empresas permiten la recopilación de datos de los usuarios con el propósito de encontrar información muy valiosa que pueda convertirse en conocimiento hacia las personas y puedan tomar decisiones acertadas. Por lo tanto, cada vez se genera mayor volumen de datos, donde los sistemas computacionales deben ser más robustos, integrando algoritmos de aprendizaje de máquina que permita esta búsqueda de conocimiento de manera óptima y poder evitar redundancia y ruido en las bases de información.

1.5. LOCALIZACIÓN GEOGRÁFICA DEL PROYECTO

Imbabura, Zona 1 del Ecuador

El proyecto abarcará datos adquiridos de un número de personas utilizando el guante y haciendo el gesto de lenguaje de señas para cada número durante un minuto. Se pretende que el área de influencia sea cualquier parte del Ecuador.

1.6. HIPÓTESIS

Ho: La aplicación de un algoritmo de aprendizaje de máquina en los datos procedentes de un guante inteligente, optimizará el rendimiento y procesamiento del sistema electrónico dependiendo de la elección del grupo que mejor represente la relación de los datos.

Hi: La aplicación de un algoritmo de aprendizaje de máquina en los datos procedentes de un guante inteligente, no optimizará el rendimiento y procesamiento del sistema electrónico.

1.7. TIPO DE INVESTIGACIÓN

Para el desarrollo del presente trabajo la investigación se determina de tipo descriptiva, analítica y correlacional; aplicada de carácter explicativa. Por lo cual, la orientación de la misma será de tipo mixta.

1.8. POBLACIÓN Y MUESTRA

Se estudia una muestra de 20 personas de diferente sexo y edad con capacidades especiales, las cuales utilizaron el guante inteligente y realizaron el gesto de lenguaje de señas para cada número (del 0 al 9) durante un período de tiempo de 1 minuto cada uno.

Para la recolección de información se registraron las acciones pertinentes de las 20 personas, teniendo presente 5 indicadores de los movimientos (pulgar, índice, medio, anular y meñique). Los datos adquiridos se almacenan en una matriz T, de orden m × n, donde m es el número de muestras y n es el número de atributos que representan cada uno de los datos y una etiqueta de posición. De esta manera, se acumuló un dataset de m=5.000 muestras y n=5 atributos.

1.9. MÉTODOS DE LA INVESTIGACIÓN

El método a utilizar es la experimentación controlada para manipular variables que no han sido usadas con anterioridad, al no contar con históricos de los datos es necesario realizar análisis y reconocer los parámetros más característicos con fines de reducir la dimensionalidad de los datos y aplicar algoritmos de aprendizaje de máquina no supervisados. La metodología propuesta es dividir los datos en un conjunto de entrenamiento y otro de prueba, equilibrio de datos, comparación de la selección de prototipo y clasificador dentro del sistema electrónico. La técnica por usar es la simulación para observar el comportamiento de los datos en relación a eventos definidos.

1.10. RECOLECCIÓN DE LA INFORMACIÓN

En la recolección de la información estadística se emplearán los siguientes actores:

a) Un investigador del presente trabajo es quien se encargará de realizar el estudio, análisis y pruebas de la base de datos.
b) Veinte personas son quienes harán uso del guante inteligente para realizar los gestos del lenguaje de señas.

1.11. TRATAMIENTO Y ANÁLISIS ESTADÍSTICO DE LOS DATOS

Se utilizará el programa Microsoft Excel 2013 para la tabulación de los datos recolectados mediante la metodología experimental dispuesta al efecto. Por otra parte, se utilizará el software Arduino para la programación en el microcontrolador y el software R-studio para procesar las variables correlacionales disponibles.

1.12. RECURSOS ECONÓMICOS (PRESUPUESTO)

Para el diseño estético del guante inteligente, en la tabla 1 se presentan los costos que implicaron su diseño, así como de los materiales adicionales que fueron utilizados. Los recursos económicos que relativamente se emplearán en la investigación se especifican en la siguiente tabla:

Tabla 1
Recursos Económicos para el proyecto

CANTIDAD	DETALLE	V. UNITARIO	VALOR TOTAL
1	Guante	$ 24.90	$ 24.90
1	Velcro (20cm)	$ 0.50	$ 0.50
1	Hilo Calibre 120	$ 1.00	$ 1.00
1	Rediseño y Mano de Obra	$ 38.00	$ 38.00
5	Sensores Flexibles de 5.58 cm	$ 16.00	$ 80.00
1	LilyPad Arduino	$ 20.00	$ 20.00
1	Batería LIPO de 3.7v – 680mha	$ 16.00	$ 16.00
1	Hilo Conductor (6m)	$ 6.0	$ 6.00
20	Cables conectores Macho-Macho	$ 0.10	$ 2.00
5	Resistencias 330 kΩ 1/2W	$ 0.8	$ 0.40
1	Resistencias 1kΩ 1/2W	$ 0.8	$ 0.8
1	Pulsador	$ 0.15	$ 0.15
1	Bornera de 2 Pines	$ 0.25	$ 0.25
1	Switch	$ 0.75	$ 0.75
1	Espadines tipo Hembra	$ 1.10	$ 1.10
1	Papel Termotransferible	$ 0.75	$ 0.75
2	Impresión Láser	$ 0.30	$ 0.60
1	Estaño (1m)	$ 0.35	$ 0.35
1	Cautín	$ 2.0	$ 2.0
1	Acido	$ 1.50	$ 1.50
1	Baquelita Perforada	$ 0.75	$ 0.75
1	Baquelita de cobre	$ 1.80	$ 1.80
1	Cable Termocontraíble (1m)	$ 0.60	$ 0.60
1	Brujita	$ 0.50	$ 0.50
1	Silicona	$ 1.50	$ 1.50
2	Diodo LED	$ 0.15	$ 0.30
	Valor Total		**$201.50 USD**

1.13. CRONOGRAMA

Tabla 2
Cronograma de actividades del proyecto

	ACTIVIDAD	DURACIÓN																							
1	Revisión sistemática de literatura y revisión del estado del arte																								
2	Selección de hardware y software para el sistema del guante inteligente																								
3	Diseño electrónico y sistema de recopilación de datos																								
4	Adquisición de datos para sistemas embebidos.																								
5	Análisis de los datos de preprocesamiento y comparación de la selección de prototipos																								
6	Realizar una comparación entre los algoritmos de selección de métodos para grandes volúmenes de datos.																								
7	Determinar el algoritmo de clasificación y de reducción de dimensionalidad (RD) adecuados																								
8	Pruebas de verificación de los métodos escogidos para mejorar el rendimiento del sistema																								
9	Conclusiones y recomendaciones																								
	TIEMPO EN SEMANAS	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

CAPÍTULO II

MARCO TEÓRICO METODOLÓGICO

En este capítulo se describirá de forma detallada las características de la comunicación no verbal, lenguajes no verbales, para una comprensión inicial del tema. También se contextualizará elementos que intervienen en funcionamiento del guante inteligente (Microcontroladores, sensores, módulos de comunicación inalámbrica, elementos pasivos electrónicos, entre otros); además se hará referencia a la limpieza y balanceo de datos, pre procesamiento de datos, algoritmos de aprendizaje de máquina.

2.1. COMUNICACIÓN NO VERBAL

2.1.1. Introducción

Un millón de personas en el mundo carecen de la capacidad para oír o hablar, una persona con problemas de audición tiene la imposibilidad de recibir sonidos a través de sus oídos y que obstaculiza la capacidad de expresarse verbalmente. Sus causas pueden ser hereditarias, degenerativas, accidentales, entre otros (K. S. Abhishek, Aug 2016). Por lo tanto, esta limitación crea una barrera de entendimiento entre personas sanas y personas sordas. Es un reto para los sordos y mudos comunicarse eficazmente con la gente que encuentran en su diario vivir (Nooritawati, March 2012).

La comunicación no verbal se lleva a cabo mediante la imitación de gestos o signos sin utilizar el habla, el mensaje a querer transmitir se lo realiza mediante la expresión corporal o facial y sobre todo la señalización de los distintos objetos existentes a nuestro alrededor los cuales ayudan a descifrar el mensaje a transmitir.

Adicional algunos estudios indican que el 60% de lo que queremos transmitir lo hacemos mediante una comunicación no verbal, es decir, ayudándonos de gestos, movimientos, expresiones, miradas, señas, entre otros.

Las personas que poseen cierta capacidad especial tienen la intención de poder comunicarse con el entorno que los rodea de diferentes maneras, proporcionando información acerca de sus emociones, estados de ánimo y sobre todo de las necesidades básicas que presentan a diario; la poca estructuración y la dificultad en interpretar la información han hecho que exista una brecha de poca relación al momento de tratar con este tipo de personas que sufren una cierta capacidad especial, caso estudiado en las personas con discapacidad auditiva y de lenguaje.

2.1.2. Modos de la comunicación

"La mayor parte de los gestos y movimientos empleados y que se usan regularmente están condicionados por el entorno que nos rodea, es decir, la cultura en la cual nos hemos criado" (Flora Davis, 1993, p. 1). Esto también depende del entorno familiar ya que tiene una clara influencia en nuestro comportamiento y en nuestros modos de poder comunicarnos con nuestro cuerpo. Existen varios modos de comunicación, entre ellos se destacan las partes de nuestro cuerpo que más a menudo están en movimiento al momento de establecer una comunicación dentro del entono que nos rodea como es el olfato, la vista, el tacto, las posturas del cuerpo y sobre todo los gestos de las manos.

2.2. LENGUAJE NO VERBAL

2.2.1. Sistemas de comunicación

El componente no verbal trata de comunicar los distintos estados y actitudes

que presentan las personas al momento de querer transmitir un mensaje; mientras que el componente verbal trata de comunicar información más concreta del mensaje a transmitir. Existen dos sistemas de comunicación no verbal que pueden ser especificados para cada individuo o de forma general y estos son:

2.2.1.1. Lenguaje Corporal

También conocido como la kinesia, "es todo aquel que presentamos a diario al momento de comunicarnos con los demás como son los gestos, movimientos corporales, el tono de voz, formas de mirar, formas de vestir, y posturas del cuerpo" (Rebel y Edaf, 2001, p. 22).

2.2.1.2. Lenguaje Icónico

Colle (1998) señala que "el lenguaje icónico es conocido como un sistema de representación gráfica de la realidad a través de imágenes, en él se incluyen muchas formas de comunicación no verbal" (p. 7), como: los códigos universales (Morse, Braylle, lenguaje de los sordomudos), códigos semiuniversales (el beso, signos de luto o duelo), y códigos secretos (señales de árbitros deportivos, maestros de clase, policías de tránsito).

2.3. FUNDAMENTO DE LA COMUNICACIÓN PARA PERSONAS CON DISCAPACIDAD AUDITIVA

2.3.1. Introducción

Personas que presentan una cierta capacidad especial en el caso de los sordomudos, la comunicación con el medio que los rodea en sí, se vuelve una barrera

constante al momento de intercambiar información; hecho que se ve evidenciado aún más en el ámbito familiar, es por esto que el desarrollo de prototipos electrónicos de comunicación son alternativas que permitirán ampliar las capacidades de comunicación de las personas con discapacidad auditiva y de lenguaje, además de los métodos ya existentes como son la labio lectura y el lenguaje de signos.

2.3.2. Patología auditiva y del habla

"Podemos definir a la patología dentro de la medicina como la encargada de estudiar los trastornos anatómicos y fisiológicos de tejidos y órganos enfermos" (Hurtado, 2004, p. 2). Se llama a una persona sordomuda debido a que ha perdido la capacidad auditiva y vocal al mismo tiempo, es decir, que durante su proceso de crecimiento estas personas no han podido desarrollar la destreza de escuchar y hablar. Este tipo de enfermedad puede ser hereditario, o por consecuencia de algún traumatismo, exposición al ruido a largo tiempo o por el consumo de medicamentos que afectan al nervio auditivo.

Podemos decir que la comunicación no verbal ha permitido desarrollar habilidades en las personas con discapacidad auditiva y de lenguaje con la finalidad de ampliar sus capacidades de comunicación, el lenguaje signado o de señas ha permitido que muchas personas mejoren su comunicación dentro de su propio entorno; pero aun así existe una brecha importante entre las personas sordomudas con personas que no presentan este tipo de capacidad especial.

2.3.3. Lenguaje de señas

Lenguaje de Señas (LS) es un lenguaje de expresión natural de producción gestual y percepción visual que establece un canal de comunicación entre las personas sordas y las personas que tienen conocimiento de LS (M. B. H. Flores, 2014). Muchos trabajos de investigación han desarrollado sistemas de

reconocimiento de signos para diferentes idiomas, además hay algunos retos como clasificadores de aprendizaje de la máquina, la comprensión de la acción humana, el procesamiento del lenguaje natural, entre otros (G. Bernieri, 2015). Como conclusión, el reconocimiento eficiente del sistema de lenguaje de señas automático sigue siendo un problema abierto.

López, Rodríguez, Zamora, y Sosa (2008) señalan que el lenguaje de señas "es uno de los mecanismos de comunicación que utilizan las personas sordomudas para poder intercambiar información dentro del entorno que los rodea" (p. 24); esto les permitirá expresar sus sentimientos, pensamientos y emociones de acuerdo con los movimientos y flexibilidad de las señas que ellos emitan hacia los demás. Las personas que presentan esta capacidad especial deben estudiar una alternativa que les permita comunicarse de alguna u otra manera con el medio que los rodea, es por esto que nace la necesidad de poner a prueba las capacidades de cada persona para poder entender el lenguaje de señas que deben aplicar para su inclusión dentro de algún entorno ya sea social, laboral, educativo o familiar.

2.3.3.1. Ejercicios Previos al Lenguaje de Señas

Las personas con discapacidad auditiva y de lenguaje que inician la comprensión y manejo de las señas, "sienten a sus inicios cierta timidez y severidad al momento de formar las señas con sus manos" (López, Rodríguez, Zamora, y Sosa, 2008, p. 26).

Es necesario tener una buena coordinación de los movimientos con la finalidad de que el mensaje a transmitir sea claro y preciso para la persona que lo recibe, la práctica constante de estos movimientos hará que las personas con discapacidad auditiva y de lenguaje cierren esa brecha de inclusión dentro del entorno que los rodea y puedan comunicarse sin ningún impedimento dentro de su propio medio; es así que se presentan ciertos requisitos muy útiles y necesarios para el buen manejo del lenguaje de señas.

2.3.4. Actualidad del lenguaje de señas en ecuador

El lenguaje de señas hoy en la actualidad es una parte muy esencial para la comunicación de personas que presentan una cierta discapacidad auditiva, tal es el caso que para cada cuidad, país o región existen varias maneras de comunicación.

En el Gobierno de Ecuador ya se ha presentado el primer diccionario de señas dirigido especialmente para las personas con discapacidad auditiva y de lenguaje. El diccionario Oficial de Lengua de Señas fue lanzado el 15 de octubre del 2012 y elaborado conjuntamente con la Federación Nacional de Personas Sordas de Ecuador, el Ministerio de Educación y la Agencia de Cooperación de Estados Unidos.

2.3.4.1. Innovación Tecnológica

Tanto en el Ecuador como a nivel mundial se han desarrollado un sin número de prototipos electrónicos traductores de señas enfocados especialmente para las personas que presentan una cierta discapacidad auditiva y del habla, con la finalidad de solucionar problemas en beneficio del sector social.

Hoy en la actualidad el avance tecnológico especialmente en el área de la electrónica y tecnologías inalámbricas, han posibilitado un sin número de aplicaciones enfocadas a mantener la inclusión de personas con capacidades especiales dentro de la sociedad, devolviéndoles así la habilidad, autoestima y confianza para el desarrollo de sus actividades diarias.

La utilización de sensores, acelerómetros, módulos de comunicación inalámbrica, microcontroladores, elementos electrónicos pasivos, el uso de Smartphone y desarrollo de aplicaciones; han logrado facilitar el diseño y construcción de diferentes prototipos, desempeñando varias funciones y han sido presentados a nivel nacional provenientes de las diferentes universidades del

Ecuador, con el fin de romper la brecha tecnológica y realizar proyectos enfocados a resolver problemas dentro de la sociedad, caso particular en las personas que presentan una cierta capacidad especial. La aplicación del presente proyecto es la de ofrecer un sistema de aprendizaje diferente a los ya usuales existentes, es decir, presentar un lenguaje de signos o señas distinto y por medio de este permitir expresar las distintas necesidades básicas que presentan las personas con discapacidad auditiva y de lenguaje enfocados dentro del ámbito que se encuentren.

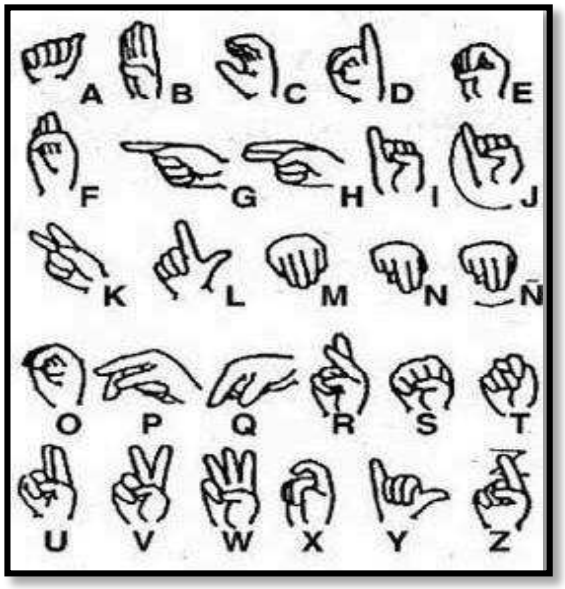


Figura 1. Alfabeto dactilológico Universal
Fuente: (Juan Oliva)

2.3.4.2. Situación Actual y Leyes de Discapacidad en Ecuador

El cambio de políticas y leyes en nuestro país últimamente ha hecho que existan avances en lo que comprende al reconocimiento de la igualdad de oportunidades e inclusión social de las personas que presentan ciertas discapacidades. Durante los últimos años se ha hecho evidente el cambio del sistema de salud, educación y empleo, en donde se garantizan los derechos de las personas con discapacidad, ofreciéndoles la debida atención, protección y cuidado de las mismas. (Orlando Caiza, 2012).

Para la Ley Orgánica de Discapacidades (2012)

Según el Art.63 se enfoca a la Accesibilidad de la Comunicación en donde el estado promocionará el uso de la lengua de señas ecuatoriana, el sistema Braille, las ayudas técnicas y tecnológicas, así como los mecanismos, medios y formatos aumentativos y alternativos de comunicación; garantizando la inclusión y participación de las personas con discapacidad en la vida en común.

Para la Ley Orgánica de Discapacidades (2012)

Según el Art.87 determina al MIES como autoridad nacional encargada de la inclusión económica y social para las personas con capacidades especiales, teniendo el objetivo de capacitar a las familias que tienen bajo su cuidado a estas personas, en el buen trato y atención que deben prestarles.

2.4. INVESTIGACIÓN DE LOS SENSORES

2.4.1. Introducción

Hoy en la actualidad los sensores se han convertido en uno de los elementos principales en el desarrollo de sistemas eléctricos y electrónicos, ya que estos son encargados de captar cualquier tipo de acción ya sea esta en magnitud física o química, es decir, el proceso o dispositivo sobre la que se ejerce el control y de la información del comportamiento del trabajo; así la señal captada es transmitida al controlador el cual la procesará para poder tomar cualquier tipo de acción específica de funcionamiento.

De esta manera, podemos encontrar a nuestro alrededor una variedad de dispositivos los cuales funcionan con sensores como pueden ser medición de distancia, velocidad, aceleración, temperatura, presión, fuerza, vibración, posicionamiento, aplicaciones en la industria como el control de máquinas, aplicaciones médicas, militares, de seguridad, además de textiles inteligentes que constan de sensores que permiten medir el pulso cardiaco, nivel de temperatura y humedad corporal; entre otros más.

2.4.2. Características de los sensores

De acuerdo con su aplicación, un sensor puede estar formado por materiales metálicos, no metálicos, orgánicos o inorgánicos, y por fluidos, gases, plasmas o semiconductores. Al usar características especiales de esos materiales, los sensores convierten la cantidad o propiedad medida en una salida analógica o digital (Letrán, 2012, p. 9); un sensor debe cumplir varios parámetros o requisitos antes de ser usado con el fin de evitar posibles fallas o errores al momento de llegar a manejarlo, los cuales son los siguientes: exactitud y precisión, rapidez de respuesta y sensibilidad, rango de funcionamiento y vida útil.

2.5. ARDUINO

2.5.1. Introducción

"Arduino es una plataforma de electrónica abierta para la creación de prototipos basada en software y hardware flexibles y fáciles de usar" (Letrán, 2012, párr. 1). En la actualidad en el mundo del desarrollo de plataformas electrónicas ha hecho que personas de toda edad se motiven a desarrollar e investigar todo tipo de proyectos o prototipos, que permitan solucionar algún problema planteado; especialmente en el campo de la "educación donde empieza a considerarse de gran

importancia dotar a las nuevas generaciones de conocimientos básicos de programación de aplicaciones o de desarrollo electrónico" (Arenas, 2014, párr. 1).

El desarrollo de proyectos interactivos ha permitido que, "Arduino pueda tomar información del entorno a través de sus pines de entrada de toda una gama de interruptores o sensores y puede afectar aquello que le rodea controlando luces, motores y otros actuadores" (Letrán, 2012, párr. 1).

2.5.2. Lenguaje de desarrollo y entorno de programación

El microcontrolador en la placa Arduino contiene un lenguaje de programación Arduino basado en Wiring, y el entorno de desarrollo Arduino basado en Processing; es decir, que los proyectos hechos con Arduino pueden ejecutarse sin necesidad de conectar a un ordenador, si bien tienen la posibilidad de hacerlo y comunicar con diferentes tipos de software.

2.5.3. Ventajas y características

La principal característica que muestra Arduino "es una plataforma electrónica de hardware libre, basada en una placa con un microcontrolador y un entorno interactivo de desarrollo, diseñada para facilitar el uso de la electrónica en el desarrollo de proyectos" (Martínez, 2014, p. 1). El entorno de desarrollo o software al ser libre se puede descargar de forma gratuita e implementa el lenguaje de programación Processing y Wiring, mientras que el hardware consiste básicamente de un microcontrolador Atmel AVR32 de las series Atmega168, Atmega328, Atmega 1280, Atmega8; con puertos de entrada y salida. Como ventajas que muestra la plataforma electrónica Arduino se puede decir que:

- Son placas electrónicas multiplataforma basados en los microcontroladores ATMEGA168, ATMEGA328 y ATMEGA1280 que "funcionan en sistemas operativos Windows, Macintosh OSX y Linux,

pero la mayoría de microcontroladores se limitan a Windows" (Arduino, 2014, párr. 6).

- Las placas electrónicas son muy utilizadas actualmente, donde los precios son relativamente accesibles para el usuario, van desde los 35 hasta los 120 dólares, dependiendo de la gama de boards arduinos que se necesiten.
- Los planos de los módulos Arduino "se basan según la licencia Creative Commons33, por lo que los diseñadores de circuitos con experiencia pueden hacer su propia versión del módulo, ampliándolo u optimizándolo" (Arduino, 2014, párr. 9). Esto quiere decir que Arduino es un hardware ampliable y de código abierto.

2.5.4. IDE de arduino

Arduino incluye su propio IDE conocido como el entorno interactivo de desarrollo basado en la aplicación escrita en Java34, la cual permite que la programación de la placa electrónica sea sencilla debido a que se basa en el código abierto de Processing y Wiring. Fue desarrollada para ser lo más amigable con el usuario.

2.5.5. Módulos arduino

Existen varios boards o placas electrónicas que presenta Arduino, entre las cuales se pueden destacar las siguientes: Arduino Uno, Arduino Mega 2560, Arduino Nano, Arduino YUN, Arduino LEONARDO, Arduino Micro, Arduino Ethernet, Arduino FIO, Arduino Robot, Arduino Explora, Arduino LilyPad. A continuación, se tiene una descripción detallada de la placa Arduino LilyPad, la cuál es la considerada para el diseño del guante electrónico.

2.5.5.1. Arduino LilyPad

El LilyPad Arduino es un "conjunto de piezas electrónicas y módulos que se

emplean para el desarrollo de piezas textiles interactivas. Sensores, altavoces o luces LED38 se cosen con hilos conductores al LilyPad Arduino y se crean prendas o accesorios dinámicos" (Sánchez, 2014, párr.1).

El microcontrolador es especial para telas y ropa inteligente. LilyPad es una tecnología electrónica textil lavable, desarrollada por Leah Buechley en cooperación con SparkFun Electronics39. "La placa electrónica está basada en el chip ATmega328V, que es una versión de bajo consumo de energía que del chip ATmega328 normalmente usado" (Arduino, 2014, párr. 1). Al ser adaptable a textiles tiene la ventaja de utilizar un hilo especial de tipo conductor, el cual ayudará a la sujeción de la placa electrónica en cualquier tipo de textil.

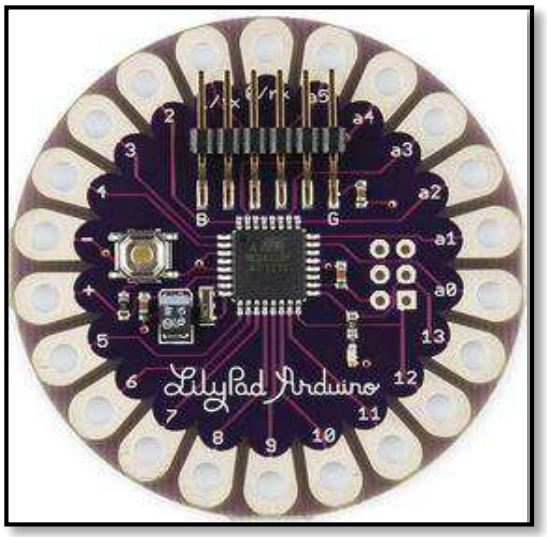


Figura 2. Placa electrónica LilyPad Arduino
Fuente: (Arduino, 2014)

2.6. TÉCNICAS DE APRENDIZAJE DE MÁQUINA

2.6.1. Introducción

El aprendizaje de máquina brinda métodos para obtener conocimiento de un conjunto de datos donde las personas no las pueden realizar por la cantidad y

complejidad de la información, con la aparición de Big Data estas técnicas se han vuelto importantes para discernir datos pocos importantes y desechar lo que no sea útil a un modelo matemático de predicción o clasificación (Berral-García, 2016).

El aprendizaje de máquina (Machine learning) tiene dos divisiones en grupos para el modelado de datos, los cuales son no supervisados y supervisados. En un aprendizaje supervisado crea modelos en base de un histórico o entrada de datos con el fin de encontrar una salida determinada, se puede dividir en pronosticar o clasificar (Brownlee, 2015).

Las técnicas no supervisadas están relacionadas a los datos que solo tienen entrada o se desconoce las variables correspondientes con el objetivo de encontrar correlaciones entre los datos que no se conocían, se pueden dividir en problemas de agrupación de datos (clústers) o la asociación entre ellos (Brownlee, 2015). Un algoritmo de machine learning debe contar con datos de entrenamiento del modelo y de validación del mismo para conocer el grado de exactitud de la clasificación o predicción en relación a los datos reales (Bishop, 2006).

Los algoritmos de aprendizaje de máquina necesitan de un pre procesamiento de datos donde se pueda preservar la información útil que represente a todas las características de un conjunto de datos de alta dimensionalidad (Bishop, 2006), es decir, de una matriz con muchas variables que no pueda el ser humano apreciar su relación (Salazar-Castro, Pena-Unigarro, & Peluffo-Órdonez, 2016) , reducir a un plano de mejor forma a un número de menos variables de su versión nominal para ser inteligible al usuario (Rosero-Montalvo, Diaz, Salazar-Castro, & Peluffo-Órdonez, 2016).

El presente trabajo solo hace relación al reconocimiento de patrones y algoritmos no supervisados de machine learning ya que no se cuentan con datos históricos para una predicción y no se conoce la forma de correlación entre la información generada por las variables a sensar desde un guante electrónico inteligente.

2.6.2. Limpieza de datos

El número de datos para el entrenamiento de un algoritmo puede tener un costo computacional alto, donde muchos datos y sus iteraciones no aportan al modelo. La selección adecuada del conjunto de entrenamiento permite aumentar la efectividad del algoritmo, reducir la carga computacional y la selección adecuada del número de datos a utilizar. En relación con la limpieza de datos existen dos tipos: eliminar el solapamiento de datos y los clasificados erróneamente (Aha, Kbiler, & Albert, 1991).

2.6.3. Reconocimiento de Patrones

En la red de sensores inalámbricos se pueden obtener diferentes variables del estado de un evento, al contar con una matriz de datos de alta dimensionalidad es presumible que exista redundancia o sea excesiva en datos, un filtrador multivariante indica que todas las variables tienen una importancia o peso equivalente que pueden representar a un conjunto de datos con menor peso. En métodos basados en modelos existen una o varias variables dependientes o independientes (Pérez López & Santín Gonzáles, 2008).

La calidad del modelo ajustado determina cuáles de ellas son las más significantes para el algoritmo. La reducción de dimensionalidad permite reducir el tamaño de la matriz de uso para modelos supervisados o no supervisados al eliminar atributos que pueden ser irrelevantes o redundantes en relación al objetivo a conseguir, permitiendo mejorar la calidad del modelo al enfocarse en correlaciones adecuadas y expresar un algoritmo con menos variables que puedan ser visualizados de mejor forma por el humano (Pérez López & Santín Gonzáles, 2008). La asociación es un concepto y la correlación es una medida de asociación (Rojas Soriano, 2006).

2.6.3.1. Algoritmos de reducción de dimensionalidad

El volumen de datos incrementa la dificultad para la detección de patrones y técnicas de machine learning, una forma de enfrentar este inconveniente es la presentación de datos en una dimensión menor que mantenga la estructura del espacio original.

Inicialmente las técnicas de reducción de dimensionalidad se basaron en métodos lineales, siendo simples y rígidos que no siempre representaban un conjunto de datos, la mayoría de variables y sus relaciones tiene un comportamiento complejo, los modelos no lineales permiten dicha detección de patrones de la naturaleza. Los métodos de reducción de dimensionalidad están orientados a la preservación de la topología de datos representado en una matriz de afinidad. (Alvarado Perez, Peluffo, & Theron, 2015).

(Alvarado Perez, Peluffo, & Theron, 2015, p. 4) Indica que:

> "Los métodos utilizados para reducción de dimensionalidad son: a *Laplacian Eigenmaps* (LE) (Belkin & Niyogi, 2003) y Locally Linear Embedding (LLE) (Roweis & Saul, 2000), los cuales son de tipo espectral. Otros métodos emergentes tienen una connotación probabilística ya que se basan en divergencias gracias a que la matriz de similitud normalizada puede ser vista como una distribución de probabilidad, entre estos métodos se pueden distinguir *Stochastic Neighbour Embedding* (SNE) (Bunte, Haase, Biehl, & Villmann, 2012), y sus variantes y mejoras, tales como t-SNE que utilizan una distribución t-student y JSE que usa la divergencia de Jensen-Shanon (Lee, Renard, Bernard, Dupont, & Verleysen, 2013). Algunos métodos utilizan criterios de conservación de la varianza y la distancia (Borg, 2005), entre estos métodos encontramos a *Principal component analysis* (PCA) y *Classical Multidimensional Scaling* (CMDS)"

2.6.4. Algoritmos no supervisados

Son conocidas por descubrir conocimiento en patrones ocultos en las bases de datos que representan información útil para una toma de decisión, los algoritmos no supervisados permiten resolver problemas de clasificación al encontrar relaciones antes no tomadas en consideración. (Moreno García & López Batista, 2004).

2.6.4.1. Clúster

Clasificar casos es una necesidad del mundo, el análisis de clúster es una colección de métodos estadísticos agrupar casos que representen características similares que puedan diferenciarse unos de otros (Bishop, 2006). Existen métodos de clúster particionales que obtienen una clasificación mediante la optimización de una función, el usuario debe fijar el número de clústeres que tendrá la partición, existen diferentes técnicas que permiten la estimación del valor de K (Número de clústeres). La partición en clúster sea la adecuada depende de la variabilidad dentro de los clústeres, mientras menor sea es mejor. Uno de los algoritmos más usados es K-medians (Sierra Araujo, 2006).

2.6.4.2. Asociación

Su funcionalidad es buscar patrones donde exista relación entre variables en grandes bases de datos con el fin de identificar reglas a base de medidas de interés y permitan tomar decisiones acertadas. Existen algoritmos como a priori, eclat, fp, entre otros que buscan parámetros de similitud de variables (Bayardo, Agrawal, & Gunopulos, 2000).

2.6.5. Validación de modelos

Estimar el porcentaje de acierto de un clasificador no indica la capacidad de predicción sobre las nuevas instancias que lleguen al modelo planteado. El método para conocer sus bondades es mediante la tasa de error que contabiliza el número de errores en relación al total de casos logrando una matriz de confusión que permite conocer los errores cometidos a lo largo de las instancias categóricas del problema (Sierra Araujo, 2006).

2.7. CLASIFICACIÓN

Los sistemas de clasificación tienen la función de usar datos estructurados en forma de instancias o ejemplos con distintos tipos de atributos como unidad de información para aprender de ellos y ser capaces de saber clasificar correctamente futuras instancias o ejemplos. En la actualidad existen sistemas de clasificación muy buenos, pero por lo general no están pensados en ser utilizados en entornos BigData. (Iñiguez Jiménez, 2016)

Ante esta creciente cantidad de dispositivos conectados a redes que generan y procesan gran cantidad de información los problemas de computación ya no se pueden resolver solo con un ordenador. En los últimos años han ido apareciendo tecnologías como la automatización de procesos, internet de las cosas, entre otras. Ahora necesitan ser resueltos por ordenadores más potentes dándose así un incremento de la popularidad del BigData. Debido a esta creciente popularidad han aparecido distintas tecnologías para resolver los problemas BigData.

2.7.1. Algoritmos de clasificación

La clasificación es uno de los temas más estudiados dentro del campo de la minería de datos y el aprendizaje automático, la razón de esto es que hay una gran

cantidad de problemas en diferentes áreas como seguridad, medicina o finanzas que necesitan clasificar muchos de los datos que manejan. El objetivo de los clasificadores es el de construir un modelo o clasificador a partir de un conjunto de ejemplos ya clasificados que permita clasificar nuevos ejemplos no vistos anteriormente.

El problema al que nos enfrentamos en este proyecto es el de un problema de clasificación supervisada. Esta clase de problemas se dividen en dos fases principales, a continuación se detallan:

1. El sistema de clasificación usará una serie de ejemplos llamados ejemplos de entrenamiento que ya estarán clasificados para aprender de ellos. Usando los datos de entrenamiento creará una serie de reglas o métodos de decisión para clasificar correctamente los ejemplos de entrenamiento.
2. Una vez creado el sistema de clasificación se pasará a clasificar ejemplos de test para ver qué tan bueno es el clasificador que se ha desarrollado.

2.7.1.1. Algoritmo KNN (k Nearest Neighbours)

Para comprobar si se han generado buenos prototipos, una de las mejores maneras es usar el algoritmo kNN. Este algoritmo de clasificación se basa directamente en las instancias para clasificar siendo idóneo para ver como mejora tanto en precisión como en rapidez.

Para saber qué tan buenos son los prototipos desarrollados se debe usar el algoritmo de clasificación kNN (k Nearest Neighbours). Dicho algoritmo se encarga de clasificar cada una de sus instancias en referencia a sus vecinos. Dicho de otro modo, en base a las instancias que tenga más cerca y por tanto las más parecidas a

esta. En el algoritmo de clasificación kNN existen 3 fases fundamentales en las que se puede decidir. (Iñiguez Jiménez, 2016)

1. El número de vecinos: Afecta directamente a la cantidad de instancias que tenemos en cuenta, si son pocos podemos vernos afectados por instancias que son ruido, si son demasiados puede que clasifiquemos mal debido a que cojamos instancias de otras clases.

2. El cálculo de la distancia: La función de distancia es la que decide de entre todas las instancias cuáles son las más cercanas. Las funciones de distancia más usadas son la distancia Euclídea y la distancia de Manhattan.

    a) Distancia de Manhattan: $d(x, y) \sum_{k=1}^{n} |X_k - Y_k|$
    b) Distancia Euclídea: $d(x, y) = \sqrt{\sum_{k=1}^{n} (X_k - Y_k)^2}$

3. Método de voto: Una vez tenemos todos los ejemplos más cercanos hay que elegir de qué clase es la instancia. Esta decisión se toma mediante distintos tipos de votaciones como votaciones simples o ponderadas.

    a) Votación simple: Seleccionar los k ejemplos con la menor distancia al ejemplo que se quiere clasificar. El ejemplo a clasificar será de la clase que más se repite de entre los k ejemplos previamente elegidos.
    b) Votación ponderada: La validez del voto viene en función de la distancia a la que se halla del ejemplo a clasificar.

    $clase = argmax_{v \in C} \sum_{y \in Ek} (\frac{1}{d(ey,ei)^2} * (v = clase(e_y)))$ EK, conjunto de los k ejemplos más cercanos a e'.

Algoritmo. kNN con votación simple

Entrada: E, conjunto de datos de entrenamiento, x, ejemplo a clasificar, C, conjunto de clases posible del ejemplo y k, número de vecinos.

Salida: $c \in C$ clase del ejemplo x

1. **Por cada** $e \in E$ **Hacer**
2. Calcular la distancia d(e,x)
3. **FIN**
4. Crear un subconjunto $N \in E$ con los k ejemplos de que más cercanos de x.
5. $clase_x = argmax_{c \in C} \sum_{n \in N} I(c = clase(n))$, donde $(\cdot)$ es una función indicador que devuelve 1 si el argumento es verdadero o 0 en otro caso.

Figura 3. Algoritmo KNN con votación simple
Fuente: (Iñiguez Jiménez, 2016)

En este proyecto se ha usado para calcular la distancia entre dos ejemplos la distancia Euclídea y el método de voto la votación simple.

2.8. EQUILIBRIO DE DATOS

En la actualidad se tiene tantos datos, que lo que preocupa es el equilibrio entre "datos de entrenamiento" y "datos para testar" y probar el modelo y su eficiencia/precisión. La optimización del rendimiento del modelo (el "Just Right" de la siguiente gráfica) ahora se puede elegir con mayor flexibilidad, dado que podemos disponer de datos para llegar a ese punto de equilibrio. (Rayón, El entrenamiento del modelo con datos y los problemas de "underfitting" y "overfitting" , 2017)

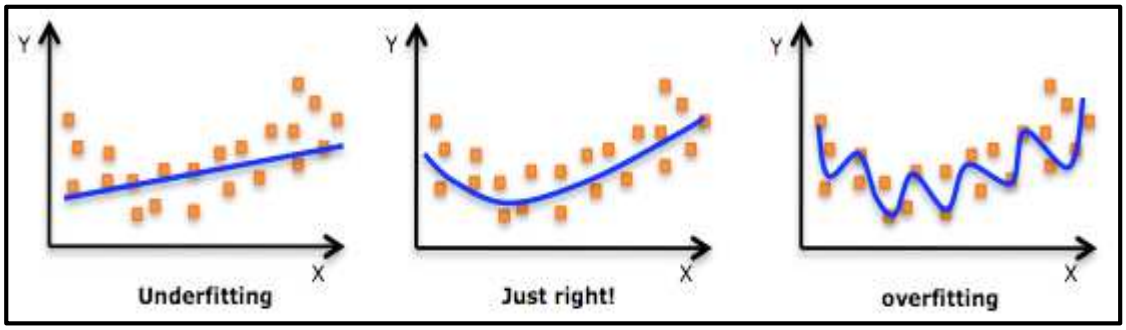


Figura 4. Equilibrio de datos con KS
Fuente: (Rayón, 2010)

2.8.1. Método Kennard-Stone

El método KENNARD STONE selecciona un subconjunto de muestras de *X* que proporcionan cobertura uniforme sobre el conjunto de datos e incluye muestras en el límite del conjunto de datos (Stone, 2006). El método comienza encontrando las dos muestras que están más alejadas usando la distancia geométrica.

Para añadir otra muestra al conjunto de selección, el algoritmo selecciona de las muestras restantes aquella que tiene la mayor distancia de separación de las muestras seleccionadas. La distancia de separación de una muestra candidata del conjunto seleccionado es la distancia desde el candidato hasta su muestra más cercana seleccionada.

Esta muestra más separada se añade entonces al conjunto de selección y se repite el proceso hasta que se ha añadido al conjunto de selección el número requerido de muestras k. En la práctica esto produce una red muy uniformemente distribuida de puntos seleccionados sobre el conjunto de datos e incluye muestras a lo largo del límite del conjunto de datos. El método funciona eficientemente porque calcula la matriz de distancias entre muestras sólo una vez.

2.8.2. Muestreo de espectros para calibración por PLS - Algoritmos de Kennard-Stone

Kennard y Stone propusieron un método secuencial que debe cubrir la región experimental de manera uniforme que es lo que se pretende al utilizar un diseño de experimentos. El procedimiento consiste en seleccionar como siguiente muestra (objeto candidato) aquel que se encuentra a mayor distancia de los objetos previamente seleccionados (objetos de calibración). La distancia utilizada normalmente es la distancia Euclidea aunque es también posible, y probablemente es mejor, utilizar la distancia de Mahalanobis. En un primer momento, se seleccionan los dos objetos que se encuentran a mayor distancia dentro del espacio experimental.

De todos los puntos candidatos, se selecciona aquel que esté más alejado de los dos primeros previamente seleccionados y se añade al conjunto de las muestras de calibración. Para ello, se determina la distancia entre cada punto candidato i_0 y cada punto i que ha sido ya seleccionado y se determina cuál es la menor distancia: $(min_i(d_{i,i_o}))$. (Restrepo, 2011). De entre estos valores se selecciona aquel para el que la distancia sea máxima $d_{seleccionado} = max_{i0}(min_{i0}(d_{i,i0}))$.

En ausencia de fuertes irregularidades en el factor espacio, el procedimiento comienza con la selección del conjunto de puntos próximos a aquellos seleccionados mediante el método D-optimal, en los límites del conjunto de datos (más el punto central).

Entonces se procede a rellenar el espacio de calibración. Kennard y Stone denominaron a su procedimiento “algoritmo de trazado uniforme”; proporciona una distribución plana de datos que, como se explicó antes, es la más adecuada para el modelo de regresión. (Restrepo, 2011).

2.9. MÉTODOS DE REDUCCIÓN DE DIMENSIONALIDAD

2.9.1. Introducción

Los MRD (métodos de reducción de dimensionalidad) son procedimientos que mapean el conjunto de datos a subespacios derivados del espacio original, de menor dimensión, en los que se encuentran en todo el conglomerado de la información, permiten una representación adecuada y significativa de estos y con un número pequeño de parámetros que logran evidenciar propiedades no observables (Lee, 2007). Los métodos de RD tienen varias ventajas, entre las cuales resaltan las siguientes: favorece la compresión, elimina la redundancia del conjunto de datos y permite mejorar procesos de visualización y clasificación de los datos a un menor uso de los recursos computacionales.

2.9.2. Clasificación de MDR

Los métodos de reducción de dimensionalidad son algoritmos que mapean el conjunto de los datos a subespacios derivados del espacio original, de menor dimensión, que permiten hacer una descripción de los datos a un menor costo. Por su importancia, son ampliamente usados en procesos asociados a aprendizaje de máquina.

En la actualidad, el creciente volumen de información generado por sistemas de información y comunicación derivados de investigaciones y procesos industriales demandan nuevas técnicas de manipulación de datos con el objetivo de extraer información no trivial que reside, de manera implícita, para facilitar la obtención de patrones y su análisis (Yeguas, 2009).

Sin embargo, evaluar esos millones de datos capturados en tiempo y espacio es altamente complejo, por lo que se busca algoritmos matemáticos que mejoren tiempos de respuesta; pero que, a su vez, la información intrínseca se pueda

recuperar (Mateos , 2016). Por esto, es imprescindible contar con métodos de reducción de dimensionalidad (MRD) eficientes que permitan simplificar la descripción del conjunto de datos y que sean capaces de abarcar grandes volúmenes de información en tiempos prudenciales.

2.9.2.1. Análisis de Componentes Principales (PCA)

El análisis de componentes principales (ACP) es una técnica lineal que se utiliza para la eliminación de la redundancia de los datos (Shlens, 2005). Es ampliamente usado, sin embargo, su mayor limitación se basa en el supuesto de linealidad.

Para Schlens (2005), ACP permite un cambio de base a una de menor dimensionalidad sobre X a través de la ecuación de transformación Y = PX, donde P es una matriz ortogonal denominada matriz de representación. El objetivo es determinar la matriz P que permita que la nube de datos pueda ser proyectada a un espacio de menor dimensión.

La estrategia es buscar P de forma que se garantice la no correlación entre vectores de Y, es decir, Cij ϵ CY, $i \neq j$ sean nulos. Si la correlación entre las distintas muestras es nula, se elimina la redundancia y el subespacio de datos puede ser descrito por P. De lo contrario, cada entrada Cij que corresponda a valores grandes que representará alta redundancia de las observaciones i y j y, por ende, habrá el ruido presente.

Según el mismo autor, el algoritmo para hallar P inicia con el centrado y estandarizado de los datos. Luego, se calcula la matriz de covarianza de X, $C_x = \frac{1}{n} XX^T$, que es simétrica y diagonalizable, y que cuantifica la covarianza entre las mediciones. Luego, se obtiene los vectores propios de Cx, que son elegidos como columnas vectores de P, ordenados de acuerdo con el valor propio y que sirven de nuevas coordenadas del sistema donde es maximizada la varianza. Se elige el número

adecuado de vectores propios que son denominados componentes principales y que describen la información del conjunto de datos de acuerdo con su coeficiente de inercia, el cual indica el porcentaje de esta, presente en cada componente principal. (Arroyo-Hernández, 2016)

2.10. PRE PROCESAMIENTO DE DATOS

2.10.1. Introducción

"La principal meta o propósito del preprocesamiento de los datos es manipular y transformar datos crudos para que el contenido de la información envuelto en el conjunto de datos pueda ser expuesto o hecho más fácilmente accesible." (Pyle, 1999).

"El Preprocesamiento de Datos" / "La Preparación de Datos" engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información (mejor porcentaje de clasificación, reglas con más completitud, entre otros) (S. Zhang, 2003)

2.10.2. Estrategias de pre procesamiento de datos

Es difícil dar una lista exacta de tareas o tópicos. Varios autores han dado a conocer las diferentes estrategias y clasificaciones. De las cuales se pueden mencionar las siguientes estrategias:

- Data collecting and integration
- Data cleaning
- Data transformation
- Data reduction (Feature Selection, Instance Selection, Discretization)

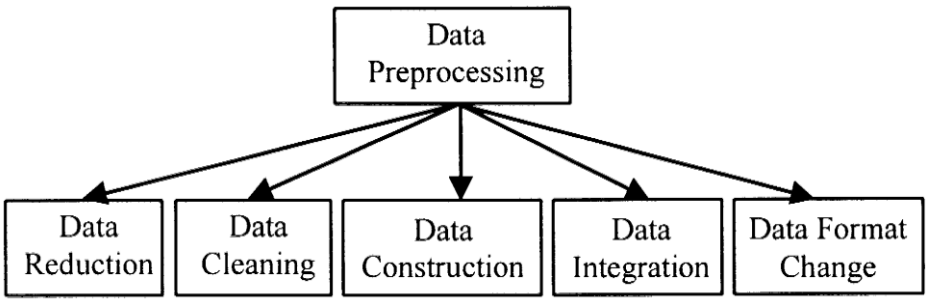


Figura 5. Estrategias de preprocesamiento de datos
Fuente: (J. Cano, 2003)

2.10.2.1. Data collecting and integration

Obtiene los datos de diferentes fuentes de información, resuelve problemas de representación y codificación, integra los datos desde diferentes tablas para crear información homogénea. (V. Detours, 2003)

2.10.2.2. Data cleaning

Las funciones que lleva a cabo Data Cleaning son: Resuelve conflictos entre datos; elimina outliers; chequea y resuelve problemas de ruido, valores perdidos. (W. Kim, 2003)

2.10.2.3. Data transformation

Los datos son transformados o consolidados de forma apropiada para la extracción de información. A continuación, se describen algunas de las formas para la transformación de datos: sumarización de datos, operaciones de agregación, entre otros. (Lin., 2002)

2.10.2.4. Data reduction

Selecciona datos relevantes para la tarea de la minería de datos/extracción de información. Existen algunas alternativas para la reducción de datos: Selección de Características, Selección de Instancias, Discretización. (H. Liu, 1998)

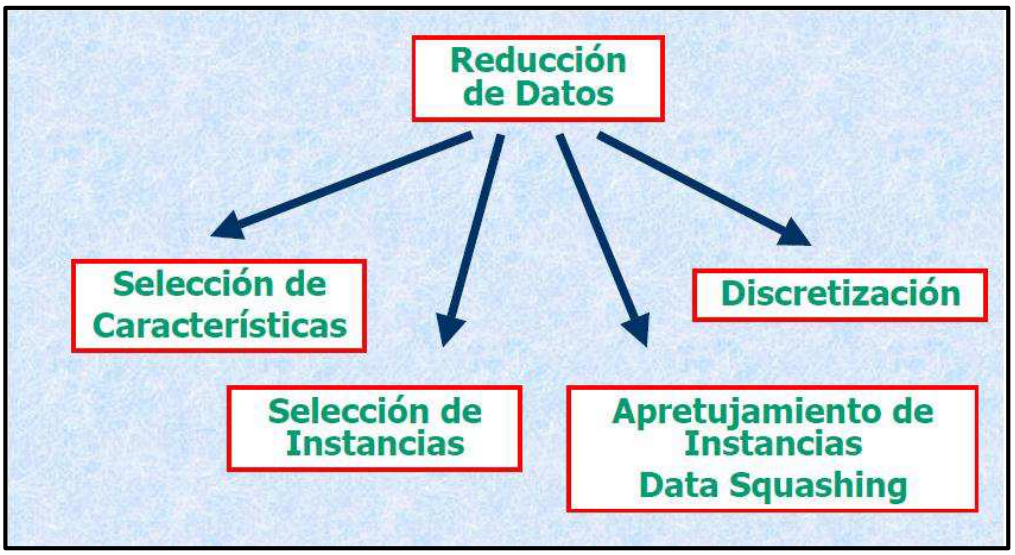


Figura 6. Estrategias de reducción de datos
Fuente: (H. Lui, 1998)

Para obtener una reducción de datos, se lo puede lograr de muchas maneras, a continuación se nombran algunas:

- ✓ Según (Motoda, 1998) seleccionando las funciones, se puede reducir el número de columnas en un conjunto de datos.
- ✓ Según (Witten, 1999) haciendo la característica discreta de valores, se puede reducir el número de posibles valores de características.
- ✓ Según (H. Brighton and C. Mellish, 2001), (Aha D. K., 2007) y (On issues of instance selection, 2002), seleccionando los casos se puede reducir el número de filas de un conjunto de datos.

La selección de instancias (IS) es una tarea centrada en la fase de preparación

de datos de KDD. Es uno de los medios más eficaces para la reducción de datos (P. Chapman, 1999). A continuación, se nombran las diferentes estrategias que se pueden seguir: muestreo, impulsando, prototipos de selección (PS), y aprendizaje activo, se va a estudiar los IS desde la perspectiva de PS.

2.11. SELECCIÓN DE PROTOTIPOS

Los métodos de selección de prototipos son una parte de los métodos de selección de instancias cuyo principal objetivo es el de seleccionar al menor número posible de instancias que permitan a un algoritmo de clasificación aprender un modelo igual o más preciso que haciendo uso del dataset original (Bush, 2001). Minimizando el tamaño del dataset se consigue disminuir la complejidad y el coste computacional de los algoritmos de minería de datos.

De una forma más acertada, la selección de prototipos se puede definir de la siguiente manera: Sea xp una instancia donde xp= (xp1,2,…,,xpc) con xp perteneciente a la clase c dado por xpc y un espacio m-dimensional donde xpi es el i-ésimo valor de la p-ésima instancia (Aha D. K., 2007). Entonces se asume que hay un conjunto de entrenamiento TR con N instancias y un conjunto de test con T instancias. Finalmente se obtiene del algoritmo de selección de prototipos un subconjunto del conjunto de entrenamiento que será usado por el algoritmo de minería de datos. Para que la selección de prototipos sea efectiva, el algoritmo de minería de datos no tiene que perder precisión respecto al conjunto de entrenamiento original.

2.12. ALGORITMOS DE SELECCIÓN Y AGREGACIÓN DE PROTOTIPOS

Los algoritmos de selección y agregación de prototipos se encargan de analizar los datos, buscar las relaciones que hay entre ellos y agruparlos según

características comunes para finalmente obtener un representante de estos al que se llama prototipo.

2.12.1. Algoritmos de selección de prototipos

La selección de prototipos se basa en seleccionar las mejores instancias para eliminar el ruido e instancias que no son necesarias. La generación de prototipos se centra en crear nuevos prototipos que representen grupos de instancias para condensar la información (Aha D. K., 2007).

Idealmente se pueden hacer algoritmos híbridos que combinen la selección de prototipos y la generación de prototipos para que el resultado sea de la mejor calidad posible. Para realizar esta labor generalmente se utilizan distintas funciones de distancia que permiten saber cuan parecidos son los datos y funciones de medias para agrupar los datos más similares y así obtener los prototipos finales.

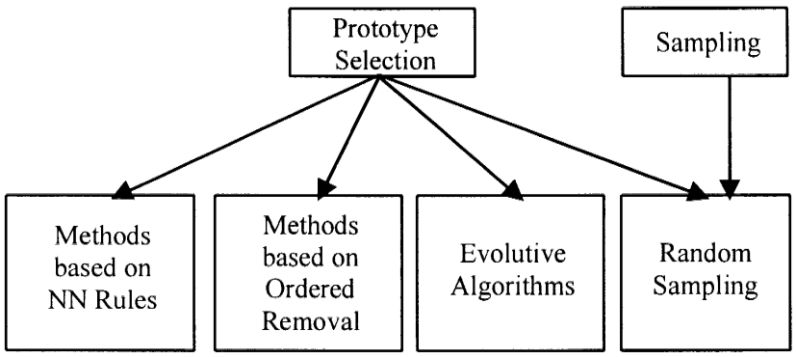


Figura 7. Algoritmos de Selección de Prototipos
Fuente: (J. Cano, 2003)

Para comprobar si se han generado buenos prototipos, una de las mejores maneras es usar el algoritmo kNN. Este algoritmo de clasificación se basa directamente en las instancias para clasificar siendo idóneo para ver como mejora tanto en precisión como en rapidez.

2.13. SELECCIÓN DE INSTANCIA

2.13.1. Selección de instancias para selección de prototipo (IS-PS).

Los clasificadores 1-NN predicen la clase de una instancia previamente no vista calculando su similitud con un conjunto de instancias almacenadas llamadas prototipos (Bush, 2001). Se ha demostrado que PS almacena un subconjunto bien seleccionado de las instancias de entrenamiento disponibles y aumenta la precisión del clasificador en muchos dominios. Al mismo tiempo, el uso de prototipos disminuye drásticamente los costos de tiempo de almacenamiento y clasificación.

Un algoritmo PS es un algoritmo IS que intenta obtener un subconjunto del conjunto de entrenamiento que permite al clasificador 1-NN alcanzar la máxima tasa de clasificación. La figura 8 muestra la forma en que actúa un algoritmo PS.

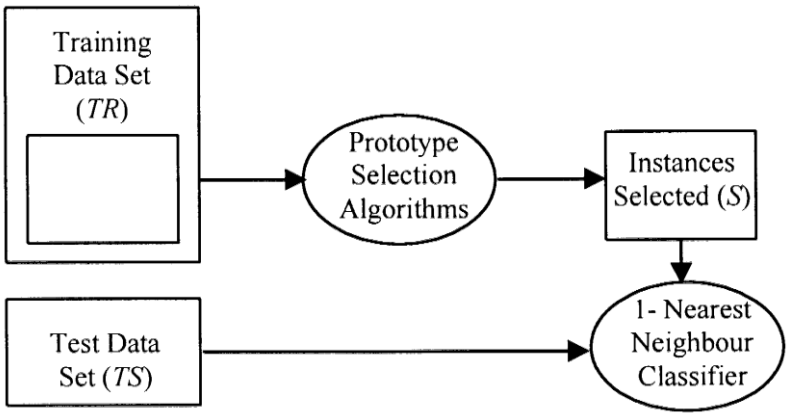


Figura 8. Estrategia IS-PS
Fuente: (J. Cano, 2003)

Cada algoritmo de PS se aplica a un conjunto de datos inicial para obtener un subconjunto de elementos de datos representativos. Evaluamos la exactitud del subconjunto seleccionado usando un clasificador 1-NN.

2.13.2. Selección de instancia para la selección del conjunto de entrenamiento (IS-TSS)

Puede haber situaciones en las que hay demasiados datos y estos datos en la mayoría de los casos no son igualmente útiles en la fase de entrenamiento de un algoritmo de aprendizaje (Bush, 2001). Mecanismos de selección de instancias han sido propuestos para elegir los puntos más adecuados en el conjunto de datos para convertirse en instancias para el conjunto de datos de entrenamiento aprendizaje.

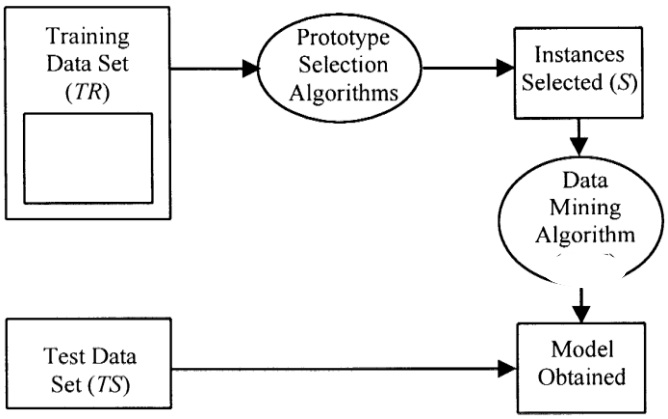


Figura 9. Estrategia IS-TSS
Fuente: (J. Cano, 2003)

La figura 9 muestra un marco general para la aplicación de un algoritmo IS para TSS. A partir del conjunto de datos TR, el algoritmo IS encuentra un conjunto adecuado, entonces un algoritmo de aprendizaje o DM es aplicado para evaluar cada subconjunto seleccionado para obtener un modelo del conjunto de datos. Este modelo se evalúa utilizando el conjunto de datos de prueba TS.

2.13.3. Descripción de los algoritmos de selección de instancia

Históricamente, IS se ha dirigido principalmente a mejorar la eficiencia del clasificador NN. El algoritmo NN es uno de los algoritmos más venerables en el aprendizaje automático. Este algoritmo calcula la distancia euclidiana (posiblemente

ponderada) entre una instancia a clasificar y cada instancia vecina de entrenamiento.

La nueva instancia por clasificar se asigna a la clase de la vecina más cercana. Más generalmente, se calcula la -NN y la nueva instancia se asigna a la clase más frecuente entre estos vecinos. El clasificador -NN también fue ampliamente utilizado y fomentado por los primeros resultados teóricos relacionados con su generalización de errores de Bayes.

Sin embargo, desde un punto de vista práctico, el algoritmo -NN no es adecuado para tratar conjuntos muy grandes de datos debido a los requisitos de almacenamiento que demanda y los costes computacionales implicados. De hecho, este enfoque requiere el almacenamiento de instancias en la memoria.

1. **Los métodos basados en reglas NN:** Cnn, Enn, Renn, Rnn, Vsm, Multiedit, Mcs, Ib2, Ib3, Icf
2. **Los métodos basados en orden o la extracción:** Drop1, Drop2, Drop3
3. **Los métodos basados en el muestreo aleatorio:** Rmhc(s), Ennrs(s)

2.14. ALGORITMOS EVOLUTIVOS

2.14.1. Introducción

Los algoritmos evolutivos son adaptables a los métodos basados en la evolución natural que pueden ser utilizados para la búsqueda y la optimización. EAs son algoritmos de búsqueda de propósito general que utilizan principios inspirados en poblaciones genéticas naturales para desarrollar soluciones a los problemas (T. Back, 2007). La idea básica es mantener una población de cromosomas, que representan soluciones plausibles al problema y evolucionan con el tiempo a través de un proceso de competencia y variación controlada.

El objetivo de este proyecto es estudiar la aplicación de un modelo representativo de EA para la reducción de datos y compararlo con algoritmos de selección de instancias. Para ello, nuestro estudio es llevado a cabo desde una doble perspectiva.

a. IS-PS: El análisis de los resultados obtenidos al seleccionar prototipos (instancias) para un algoritmo 1-NN (vecino más cercano). Este enfoque se denotará como selección de prototipo de selección de instancia (IS-PS).

b. IS-TSS: El análisis del comportamiento de EAs como selectores de instancia para la reducción de datos, al seleccionar instancias para componer el conjunto de entrenamiento que será utilizado por CHC / DROP3, un algoritmo de inducción de árbol de decisión bien conocido. En este enfoque, las instancias seleccionadas se utilizan primero para crear un árbol de decisiones y, a continuación, se utiliza el árbol para clasificar nuevos ejemplos. Este enfoque se denotará como selección de selección de instrucción de la selección de instancia (IS-TSS).

El análisis del comportamiento de EAs para la reducción de datos es, de hecho, el aspecto más importante de este estudio. Como con cualquier algoritmo, la cuestión de la escalabilidad y el efecto de aumentar el tamaño de los datos sobre el comportamiento del algoritmo están siempre presentes (T. Back, 2007). Para abordar esto, hemos llevado a cabo unas pruebas sobre IS-PS y IS-TSS con el aumento de la complejidad y el tamaño de los datos.

2.14.2. Concepto (EAs)

EAs son métodos de búsqueda estocásticos que imitan la metáfora de la evolución biológica natural. Todas las EAs se basan en el concepto de una población de individuos (representando a cada punto en el espacio de soluciones potenciales a un problema dado), que experimentan operadores probabilísticos tales como

mutación, selección y (a veces) recombinación para evolucionar hacia valores cada vez mejores de la aptitud de los individuos. La aptitud de un individuo refleja su valor de función objetivo con respecto a una función objetivo particular que se debe optimizar. El operador de mutación introduce la innovación en la población mediante la generación de variaciones de individuos y el operador de recombinación realiza típicamente un intercambio de información entre diferentes individuos de una población. El operador de selección impone una fuerza motriz al proceso de evolución prefiriendo mejores individuos para sobrevivir y reproducirse cuando se seleccionan los miembros de la siguiente generación.

2.14.3. Algoritmos de selección de instancia evolutiva

La mayor parte del éxito de EAs se debe a su capacidad de explotar la información acumulada sobre un espacio de búsqueda inicialmente desconocido. Esta es su característica clave, particularmente en espacios de búsqueda grandes, complejos y mal entendidos, donde las herramientas clásicas de búsqueda (enumerativas, heurísticas, entre otras) son inapropiadas. En tales casos, ofrecen un enfoque válido para los problemas que requieren técnicas de búsquedas eficientes y eficaces.

2.14.4. Modelos de Algoritmos Evolutivos

A continuación, describimos los cuatro modelos de EA más importantes dentro de los algoritmos evolutivos de selección de instancias. Los dos primeros son los modelos clásicos de GA; la generacional y la de estado estacionario. La tercera, la recombinación heterogénea y la mutación cataclísmica (CHC), es un modelo clásico que introduce características diferentes para obtener un equilibrio entre exploración y explotación, y el cuarto es un enfoque de EA específico, diseñado para espacios de búsqueda binaria.

2.14.4.1. Algoritmo Genético Generacional (GGA)

La idea básica en GGA es mantener una población de cromosomas, que representan soluciones plosibles al problema particular que evoluciona sobre iteraciones sucesivas (generaciones) a través de un proceso de competencia y variación controlada. Cada cromosoma en la población tiene una aptitud asociada para determinar qué cromosomas se van a utilizar para formar nuevos en el proceso de competición, esto se llama selección. Los nuevos se crean utilizando operadores genéticos como crossover y mutación.

El modelo clásico de GAs es el GGA, que consta de tres operaciones:

1. Evaluación de la aptitud individual
2. Formación de un pool genético (población intermedia) a través del mecanismo de selección
3. Recombinación a través de operadores de cruce y mutación.

El mecanismo de selección genera una nueva población $I'(t)$ con copias de los cromosomas en $I'(t-1)$. El número de copias recibidas para cada cromosoma depende de su aptitud; los cromosomas con una aptitud más alta tienen generalmente una ocasión más grande de copias que contribuyen $I'(t)$.

2.14.4.1.1. Crossover

Crossover toma dos individuos llamados padres y produce dos nuevos individuos llamados la descendencia intercambiando partes de los padres. En su forma más simple, el operador trabaja intercambiando subcadenas después de un punto de cruce seleccionado al azar. El operador de crossover no suele aplicarse a todos los pares de cromosomas en la nueva población. Se hace una elección aleatoria, donde la probabilidad de que se aplique el crossover depende de la probabilidad definida por una tasa de cruce.

2.14.4.1.2. Mutación

La mutación sirve para prevenir la pérdida prematura de la diversidad de la población mediante el muestreo aleatorio de nuevos puntos en el espacio de búsqueda. Las tasas de mutación se mantienen pequeñas, sin embargo, de lo contrario el proceso degenera en una búsqueda aleatoria. En el caso de las cadenas de bits, la mutación se aplica volteando uno o más bits aleatorios en una cadena con una probabilidad igual a la tasa de mutación. La terminación puede ser activada alcanzando un número máximo de generaciones o encontrando una solución aceptable por algún criterio.

2.14.4.2. Algoritmo Genético de Estado Estacionario (SGA)

En SGAs por lo general sólo uno o dos crías se producen en cada generación. Los padres son seleccionados para producir descendencia y luego una estrategia de reemplazo / supresión define qué miembro de la población será reemplazado por el nuevo descendiente (Whitley, 2009). Los pasos básicos del algoritmo de SGA son los siguientes:

1) Seleccione dos padres de la población I'.
2) Crear un descendiente usando crossover y mutación.
3) Evaluar la descendencia con la función de fitness.
4) Seleccionar un individuo en que puede ser reemplazado por la descendencia.
5) Decidir si este individuo será reemplazado.

En el paso 4), se puede elegir la estrategia de sustitución (por ejemplo, sustitución del peor, del más antiguo o de un individuo elegido al azar). En la etapa 5), se puede elegir la condición de reemplazo (por ejemplo, sustitución si el nuevo

individuo es mejor o reemplazo incondicional). Una combinación ampliamente utilizada es reemplazar al peor individuo sólo si el nuevo individuo es mejor.

2.14.4.3. Algoritmo de búsqueda adaptable de CHC

Durante cada generación el algoritmo CHC desarrolla los siguientes pasos.

1) Utiliza una población matriz de tamaño N para generar un intermedio población de N individuos, que son aleatoriamente emparejado y utilizado para generar N descendencia potencial.
2) Luego, se celebra una competición de supervivencia donde los mejores N cromosomas de las poblaciones de progenitores y descendientes se seleccionan para formar la siguiente generación.

CHC también implementa una forma de recombinación heterogénea utilizando HUX, un operador de recombinación especial. Intercambios HUX mitad de los bits que difieren entre los padres, donde la posición del bit a ser intercambiado es determinada al azar, CHC también emplea un método de prevención del incesto (Eshelman, 2001). Antes de aplicar HUX a dos padres, la distancia de Hamming entre ellos se mide, solamente aquellos padres que se diferencian entre sí por un cierto número de bits (umbral de apareamiento). El umbral inicial se establece en L/4, donde L es la longitud de los cromosomas. Si no hay descendientes se inserta en la nueva población, entonces el umbral se reduce por uno.

No se aplica ninguna mutación durante la fase de recombinación. En lugar, cuando la población converge o la búsqueda deja de progreso (es decir, el umbral de diferencia se ha reducido a cero y no se están generando nuevos descendientes que sean mejores que miembros de la población de padres) la población se reinicializa para introducir nueva diversidad en la búsqueda. El cromosoma que representa la mejor solución encontrada durante el curso de la búsqueda se utiliza como plantilla para volver a sembrar la población. La reasignación de la población se logra

cambiando al azar 35% de los bits en el cromosoma plantilla para formar cada uno de los otros N-1 nuevos cromosomas en la población.

2.14.4.4. Aprendizaje Incremental basado en la Población (PBIL)

PBIL es un EA específico diseñado para espacios de búsqueda binarios (Baluja, 2004). El algoritmo PBIL intenta mantener explícitamente estadísticas sobre el espacio de búsqueda para decidir dónde probar a continuación.

El objeto del algoritmo es crear un vector de probabilidad de valor real V_p, que, cuando se muestrea, revela vectores de solución de alta calidad con alta probabilidad. Por ejemplo, si una buena solución puede codificarse como una cadena de 0s y 1s alternos, una posible final V_p sería 0,01, 0,99, 0,01, 0,99, etc. Inicialmente, los valores de V_p se fijan en 0,5. El muestreo de este vector produce vectores de solución aleatorios porque la probabilidad de generar 1 o 0 es igual. A medida que progresa la búsqueda, los valores de V_p gradualmente cambian para representar vectores de alta solución de evaluación a través del siguiente proceso:

1) Una serie de vectores de solución (Nsamples) se generan sobre la base de las probabilidades especificadas en V_p.
2) V_p se empuja hacia el vector de solución generado con la evaluación más alta Sbest. $V_p[i] = V_p[i] * (1 - LR) + Sbest[i] * LR$, donde LR es la tasa de aprendizaje, que especifica cuán cerca están los pasos a la mejor solución.
3) V_p es empujado lejos de la peor evaluación, Sworse, donde Sbest y Sworse difieren. Esto se logra de la forma siguiente:

$$Si\ Sbest[i] <> Sworse[i], entonces$$

$$V_p[i] = V_p[i] * (1 - Negat_{LR}) + Sbest[i] * Negat\_LR$$

Donde Negat_LR es la tasa de aprendizaje negativo, que especifica cuán lejos están los pasos de la peor solución.

4) Después de actualizar el vector de probabilidad, el muestreo del vector de probabilidad actualizado produce un nuevo conjunto de vectores de solución y se continúa el ciclo.

Además, PBIL aplica mutaciones a V_p, con un objetivo análogo como mutación en el caso de AGs: para inhibir la convergencia prematura. Las mutaciones afectan V_p con baja probabilidad P_m en una dirección aleatoria Mut_Shif. Cabe señalar que existen problemas para establecer LR y Negat_LR en PBIL, lo que afecta a la convergencia y a la calidad de las soluciones.

2.15. IS EVOLUTIVO

EA puede aplicarse al problema de IS, ya que puede considerarse como un problema de búsqueda. La aplicación de EAs a estos dos enfoques (IS-PS, IS-TSS) se logra abordando dos cuestiones importantes: la especificación de la representación de las soluciones y la definición de la aptitud función.

1) Representación: Supongamos que un conjunto de datos denotado TR con m instancias. El espacio de búsqueda asociado con la instancia la selección de TR está constituida por todos los subconjuntos de TR. Entonces, los cromosomas deben representar subconjuntos de TR. Esto es logrado mediante el uso de una representación binaria. Un cromosoma consiste en m genes (uno para cada caso en TR) con dos posibles estados: 0 y 1. Si el gen es 1, entonces su asociado ejemplo está incluido en el subconjunto de TR representado por el cromosoma. Si es 0, esto no ocurre.
2) Función de la aptitud: Sea S un subconjunto de instancias de TR para evaluar y ser codificado por un cromosoma. Definimos una aptitud función que combina dos valores: la tasa de clasificación (clas_rat) asociado con S y el porcentaje de reducción (perc_red) de instancias de con respecto a TR.

$$Fitness(S) = a * clas_{rat} + (1 - a) * perc\_red$$

El clasificador 1-NN se utiliza para medir la tasa de clasificación, clas_rat, asociada con S. Denota el porcentaje de objetos correctamente clasificados de TR usando sólo S para encontrar el vecino más cercano. Para cada objeto y en S, el vecino más cercano se busca entre los del conjunto $S/\{y\}$.

Considerando que perc_red se define como: $perc_{red} = 100 * \frac{(|TR-|S)}{TR|}$

El objetivo de las EA es maximizar la función de aptitud definida, es decir, maximizar la tasa de clasificación y minimizar el número de casos obtenidos. En los experimentos presentados en este trabajo, hemos considerado el valor $\propto = 0.5$ en la función de aptitud, debido a que en algunas pruebas se encuentra el mejor equilibrio entre precisión y reducción con este valor.

2.16. ALGORITMOS DROP

2.16.1. Introducción

Algunos algoritmos de aprendizaje basados en la realidad se enfrentan con el problema de decidir qué instancias almacenar para su uso durante la generalización. Almacenar demasiadas instancias puede resultar en grandes requerimientos de memoria y velocidad de ejecución lenta, y puede causar una sobre-sensibilidad al ruido.

Estos algoritmos son utilizados para reducir los requisitos de almacenamiento en algoritmos de aprendizaje basado en instancias y otros algoritmos basados en ejemplares (Wilson, 1997) de los algoritmos que proporcionan una reducción sustancial del almacenamiento, los algoritmos DROP tienen la mayor precisión de generalización media en estos experimentos, especialmente en presencia de ruido de clase uniforme. Este documento ha revisado gran parte del trabajo realizado en el ámbito de la reducción de los requisitos de almacenamiento en sistemas de

aprendizaje basados en instancias (Mesa, 2008).

Algoritmos DROP (Decremental Reduction Optimization Procedure) son métodos que basan su regla de selección en términos del concepto de socio y de asociado.

- Definición: Sea X ≠ ∅, el socio de un objeto P que pertenece al conjunto X, es aquel objeto que tiene a P como uno de sus k vecinos más cercanos.
- Definición: Aquellos ejemplos que tienen a P como uno sus k vecinos más cercanos son llamados asociados de P y se denotan mediante la expresión P. A1,…,a , donde a es el número de asociados de P.

2.16.2. Tipos de algoritmos DROP

El algoritmo DROP1 elimina un objeto P de S si sus socios en S se clasifican correctamente sin P, es decir, bajo este criterio, la ausencia de P no afecta los resultados de la clasificación.

El algoritmo DROP2 verifica el efecto que causa la eliminación del objeto en los objetos de la muestra original, es decir, DROP2 elimina al objeto P de S si los socios que P tiene en TS se clasifican correctamente sin P.

Los algoritmos DROP3 y DROP4 aplican un filtrado de ruido antes de comenzar el proceso de edición. La diferencia entre ambos es el criterio empleado en la etapa de filtrado, Drop3 utiliza un filtrado de ruido antes de ordenar las instancias, esto se hace utilizando la regla: Cualquier instancia clasificada incorrectamente por sus -NN es eliminado, en cambio, DROP4 antes de eliminar el objeto ruidoso, verifica el impacto de clasificación provocado al no considerar tal objeto para determinar si será o no eliminado.

Finalmente, el método DROP5 modifica al algoritmo DROP2 de tal manera que comienza por eliminar objetos que se encuentran cerca de los enemigos más cercanos (objetos cercanos con distinta clase). El algoritmo Drop1 puede ser formalmente descrito de la manera siguiente:

```
 Algoritmo DROP1
Entrada: X → Conjunto de entrenamiento a editar
Salida: S → Conjunto editado
Método:
 1- Sea S = X
 2- Para cada objeto P in S
    2.1- Encontrar los k+1 vecinos más cercanos de P en S
    2.2- Adicionar P a cada una de las listas de sus vecinos
    asociados
3- Para cada objeto P en S
    3.1- Sea with = número de asociados de P clasificados
    correctamente con P como un vecino
    3.2- Sea without = número de asociados de P clasificados
    correctamente sin P
    3.3- Si without ≥ with
      3.3.1- Eliminar a P de S
      3.3.2- Para cada asociado A de P
      3.3.2.1- Eliminar a P de la lista de vecinos más
    cercanos de A
      3.3.2.2- Encontrar un nuevo vecino más cercano para A
      3.3.2.3- Adicionar A a la nueva lista de vecinos
    asociados
      3.3.3- Para cada vecino W de P
      3.3.3.1- Eliminar a P de la lista de asociados de W
    3.4- Fin del ciclo
4- Retornar subconjunto S
```

Figura 10. Algoritmo KNN con votación simple
Fuente: (Iñiguez Jiménez, 2016)

Cuando un ejemplo P se elimina, todos sus asociados deben eliminar a P de su lista de vecinos más cercanos y entonces deben encontrar un nuevo vecino más

cercano tal que sigan teniendo k+1 vecinos en su lista. Cuando ellos encuentran un nuevo vecino W, ellos también se adicionan a la lista de asociados de W así que siempre, cada ejemplo tiene que actualizar su lista de vecinos y de asociados. Este algoritmo elimina ejemplos ruidosos porque un ejemplo ruidoso P usualmente tiene socios principalmente de clase diferente a la suya y tales socios deben ser, probablemente, bien clasificados sin P.

El algoritmo DROP1 también elimina ejemplos en el centro de los grupos porque no hay socios cerca de sus enemigos y, por tanto, continúan siendo bien clasificados sin P. Cerca de la frontera, la eliminación de algunos ejemplos puede causar que otros sean mal clasificados porque la mayoría de sus vecinos pueden ser enemigos. Por tanto, este algoritmo tiende a almacenar puntos borde no ruidosos. En caso límite, existe una colección de ejemplos borde tales que la mayoría de los k vecinos más cercanos de cada uno de estos ejemplos está en la clase correcta.

CAPÍTULO III

DISEÑO DEL SISTEMA

En este capítulo se pretende tener una visión más clara y acertada a cerca de las siguientes etapas: diseño electrónico y sistema de recopilación de datos, adquisición de datos para sistemas embebidos.

3.1. DISEÑO ELECTRÓNICO

El lenguaje de signos representa los números y las letras con muchos gestos de las manos, para la adquisición de datos debemos conocer los movimientos de los dedos en cada número, en la figura 11 se muestra el gesto de los números de cero a nueve.

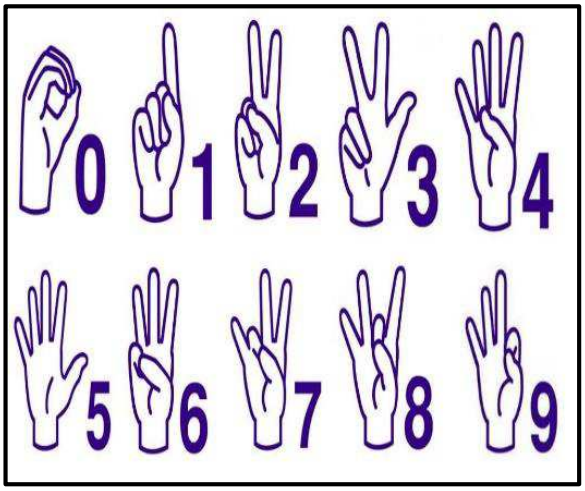


Figura 11. Gesto de mano de cada número en el lenguaje de signos
Fuente: (H. Flores, 2014)

3.1.1. Sistema del diseño electrónico

El sistema de diseño utiliza cinco sensores flexibles (uno en cada dedo), todos ellos tienen un divisor de voltaje para calibrarlos (solo se implementan con

resistencias). Estos sensores son leídos por Arduino Lilypad (usado con microcontrolador). El Arduino tiene 6 conversores analógico-digitales para convertir la resistencia del sensor en flexión para convertir a un bit. El conversor tiene una resolución de 10 bits, lo que significa que puede leer entre 0 y 1023.

Un aspecto importante es que la base de datos no necesita estar estandarizada. Finalmente, el sistema necesita una batería LiPo. La figura 12 indica el diseño del sistema electrónico con todos los componentes.

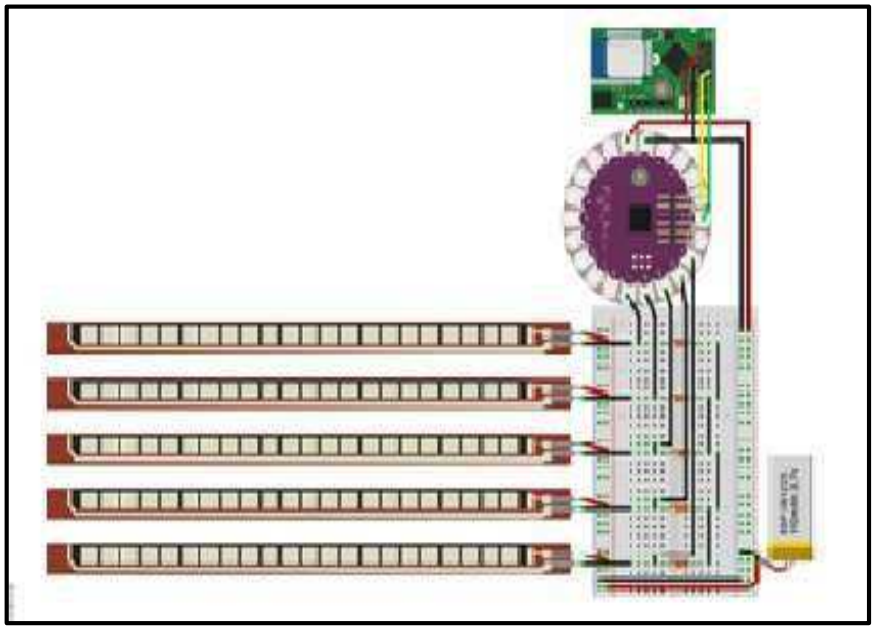


Figura 12. Sistema de diseño electrónico para el guante inteligente
Fuente: (Bernieri, 2015)

3.1.1.1. Materiales

Para la construcción del guante inteligente se han utilizado sensores, microcontroladores y otros elementos electrónicos, los cuales se detallan a continuación:

- ✓ 5 sensores flexibles,
- ✓ 1 Arduino Lylipad,

- ✓ 5 resistencias de 10 ohmios,
- ✓ 1 módulo Bluetooth hc-05
- ✓ 1 batería Li Po de 3.7 voltios a 700 miliamperios

En la figura 13 se puede observar el guante inteligente ya en funcionamiento, en el cual se ve reflejado todo el diseño, a través de este dispositivo se puede realizar la forma de los números en el lenguaje de señas y mediante el microcontrolador nos arrojará una base de datos de 5 columnas en las cuales se reflejan los movimientos de los 5 dedos (pulgar, índice, del medio, anular y meñique).

Figura 13. Visualización del guante inteligente

3.1.2. Diagrama del diseño electrónico

La fase de diseño del sistema electrónico se concluye presentando el diagrama de bloques que describe la figura 14 (S. Nunez-Godoy, 2016).

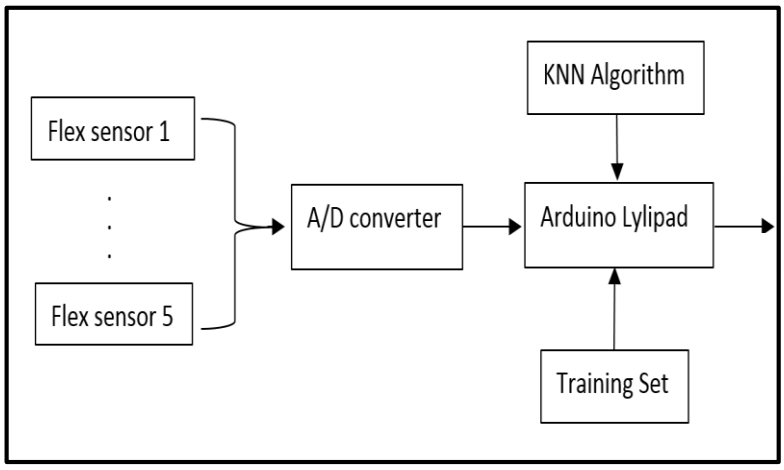


Figura 14. Diagrama de bloques para el sistema electrónico de guante inteligente

En la figura 14 se puede observar el diagrama de bloques del funcionamiento del guante inteligente. Primero se toman los datos emitidos por los sensores los cuales van a un conversor análogo/digital y después son direccionados al microcontrolador.

Cuando se obtiene el set de datos se realiza un análisis para luego poder aplicar los algoritmos de clasificación, selección de prototipos, reducción de dimensionalidad que se han estudiado durante este trabajo y se realizarán las pruebas necesarias para demostrar el mejor rendimiento y optimización de dichos algoritmos de machine learning.

3.2. ADQUISICIÓN DE DATOS

Para la adquisición de datos, veinte personas utilizaron el guante y realizaron el gesto del lenguaje de signos para cada número durante un minuto cada uno, los datos adquiridos se almacenan en una matriz T, de orden m × n, donde:

- **m:** es el número de muestras
- **n:** es el número de atributos que representan cada etiqueta de datos y una posición.

De esta manera, con la interacción de las 20 personas se logró acumular un conjunto de datos (set) de m = 5000 muestras y n = 5 atributos, como se puede observar el la figura 15.

NÚMERO	MEÑIQUE	ANULAR	MEDIO	ÍNDICE	PULGAR
1	86	100	83	30	58
1	102	135	83	31	56
1	103	135	83	33	84
1	85	137	112	30	48
1	83	125	109	30	52
1	73	126	109	27	44
1	75	114	111	29	52
1	81	112	109	29	51
1	80	115	107	27	42
1	75	113	106	30	52
1	77	116	105	28	45
1	70	106	104	30	43
1	79	109	106	30	47
1	74	113	108	27	58
1	78	104	105	30	53
1	81	113	103	30	50

Figura 15. Visualización de una parte de la base de datos obtenida del guante inteligente

3.3. DESARROLLO DE ALGORITMOS

3.3.1. Algoritmos de Clasificación

Los algoritmos de clasificación necesitan dos conjuntos de datos (entrenamiento y pruebas). El conjunto de entrenamiento es para proporcionar información al clasificador, es decir, en realidad no todos los datos son útiles.

kNN (k vecino más cercano) se considera entre los mejores algoritmos de clasificación más antiguos que utiliza la distancia más corta entre el nuevo ejemplo y todos los datos de entrenamiento para determinar el grupo correcto. El valor de k se determina normalmente mediante un conjunto de validación o mediante validación cruzada.

3.3.2. Algoritmos de Selección de prototipos.

La selección de prototipos permite elegir con diferentes criterios el mejor conjunto de datos. Según el trabajo de investigación realizado por (J. R. Cano, 2003) clasifica a los sets de datos en pequeños, medianos y grandes según el número de muestras y atributos, de acuerdo a las tablas 3 y 4.

Tabla 3
Conjuntos de datos de tamaño pequeños

	Data Set	Num. Instances	Num. Features
1	Pen-Based Recognition	297	13
2	Santimage	214	9
3	Thyroid	150	4

Fuente: (Cano, 2003)

Tabla 4
Conjuntos de datos de tamaño mediano

	Data Set	Num. Instances	Num. Features
1	Cleveland	297	13
2	Glass	214	9
3	Iris	150	4
4	LED24Digit	200	24
5	LED7Digit	500	7
6	Lymphography	148	18
7	Monk	432	6
8	Pima	768	8
9	Wine	178	13
10	Wisconsing	683	9

Fuente: (Cano, 2003)

De acuerdo con el estudio realizado en las tablas 3 y 4 nuestra base de datos se ubica en el "conjunto de datos de tamaño mediano", porque tenemos 5000 muestras y 5 atributos.

Además, estos autores realizaron una comparación de algoritmos de selección de prototipos en volúmenes de datos grandes, medianos y pequeños, donde son los mejores algoritmos para bases de datos multivariantes.

Tabla 5
Resultados promedio de IS-PS (selección de instancias para selección de prototipos) para pequeños conjuntos de datos

	Model	Execution Time (sec)	% Reduction	1-NN	
				%Ac Trn	%Ac Test
1	1-NN	0,01		89,53	71,79
2	Cnn(*)	0,01	71,98	85,45	49,10
3	Drop1(*)	0,23	86,97	87,25	53,41
4	Drop2 (*)	0,2	55,19	83,60	62,33
5	Drop 3 (*)	0,15	78,56	86,79	55,52
6	Enn	0,1	35,07	95,66	68,29
7	Ib 2 (*)	0,03	77,80	84,71	51,22
8	Icf (*)	0,29	75,81	94,59	64,02
9	Mcs	0,09	16,90	90,70	**71,82**
10	Multied	0,24	54,70	**99,83**	60,25
11	Renn	0,25	37,43	97,27	68,40
12	Rnn (*)	1,89	90,38	92,34	62,03
13	Shrink	0,08	30,41	**96,12**	**69,26**
14	Vms	0,01	74,56	90,31	63,25
15	Ib 3 (*)	0,06	65,71	79,18	65,22
16	Rmhc (*)	54,37	90,27	89,13	64,70
17	Ennsr (*)	69,95	90,27	91,31	62,04
18	GGA	70,8	87,72	90,29	62,87
19	SGA	68,6	90,50	88,94	63,36
20	CHC (*)	20,48	**96,05**	80,49	53,61
21	PBIL (*)	43,2	**93,79**	93,25	67,44

Fuente: (Cano, 2003)

Tabla 6
Resultados promedio de IS-PS (selección de instancias para selección de prototipos) para medianos conjuntos de datos

	Model	Execution Time (sec)	% Reduction	1-NN	
				%Ac Trn	%Ac Test
1	1-NN	46		94,19	94,18
2	Cnn(*)	4	97,32	79,33	80,17
3	Drop1(*)	254	96,72	78,00	76,85
4	Drop2 (*)	215	89,57	88,62	88,62
5	Drop 3 (*)	338	96,25	89,03	85,44
6	Enn	139	5,82	95,43	94,39
7	Ib 2 (*)	2	**97,78**	75,24	75,84
8	Icf (*)	386	90,04	76,77	76,68
9	Mcs	101	2,66	**95,96**	**94,38**
10	Multied	1778	13,99	92,43	92,10
11	Renn	489	6,47	95,37	94,30
12	Rnn (*)	13017	96,88	81,27	81,74
13	Shrink	206	4,89	95,19	**94,38**
14	Vms	94	1,04	94,16	94,17
15	Ib 3 (*)	42	71,67	91,49	92,61
16	Rmhc (*)	34525	90,02	91,59	91,15
17	Ennsr (*)	37802	90,02	92,79	92,75
18	GGA	66157	62,53	94,74	93,85
19	SGA	54656	62,91	95,00	93,67
20	CHC (*)	8072	**99,29**	93,31	93,53
21	PBIL (*)	32942	73,13	**96,23**	94,13

Fuente: (Cano, 2003)

3.3.3. Análisis del algoritmo evolutivo CHC

Con referencia al estudio realizado por el autor Cano, por un lado, con el criterio de reducción media el ganador es el algoritmo evolutivo CHC (Cross generational elitist selection Heterogeneous recombination Cataclysmic mutation algorithm). CHC desarrolla los siguientes pasos:

1. Utiliza una población madre de tamaño N para generar una población intermedia de N individuos, que se emparejan aleatoriamente y se usan para generar N descendencia potencial,
2. la competición de supervivencia en retenida, donde la mejor N Los cromosomas de las poblaciones de progenitores y descendientes se seleccionan de la siguiente generación.

3.3.4. Análisis del algoritmo DROP3

De los algoritmos que proporcionan una reducción sustancial del almacenamiento, los algoritmos DROP tienen la mayor precisión de generalización media en estos experimentos, especialmente en presencia de ruido de clase uniforme.

Con referencia al estudio realizado por el autor Cano, por otro lado, en el entrenamiento de precisión el mejor algoritmo es DROP3 (Decremental Reduction Optimization Procedure 3), con este método los casos ruidosos son también puntos de borde, y hacen que el orden de eliminación sea cambiado drásticamente.

Un punto ruidoso en el centro de un racimo hace que muchos puntos en ese racimo sean considerados puntos de la frontera incluso después de que el punto ruidoso sea quitado.

DROP3, por lo tanto, utiliza un filtro de ruido antes de ordenar las instancias. Esto se hace utilizando una regla similar a ENN (Cualquier instancia mal clasificada por sus k vecinos más cercanos se elimina) (S. Garcia, 2012). Esto elimina los casos ruidosos, así como los puntos fronterizos cercanos.

3.4. DESARROLLO DE LOS MODELOS

3.4.1. Criterios para la selección de algoritmos

El presente trabajo presenta un análisis de los considerados mejores algoritmos de clasificación de prototipos (DROP 3 y CHC) para determinar el adecuado para nuestro conjunto de datos, el cual se adquiere de un guante inteligente que se coloca en la mano derecha utilizado por una persona para levantar información de los números entre 0 al 9 en lenguaje de señas.

Con el entrenamiento completo, el sistema necesita un balance de datos (Kennard-Stone) para reducir el tamaño de los mismos (Martinez, 2000).

3.4.2. Reducción de dimensionaidad con PCA

Posteriormente, para verificar los resultados y poder graficarlos, se aplica una etapa de reducción de la dimensionalidad aplicando el Análisis de Componentes Principales (PCA) (P. Rosero-Montalvo, 2017), para observar agrupaciones resultantes de la selección de prototipos. Para la selección del algoritmo de selección de prototipo se consideran los siguientes criterios:

1. Instancias removidas,
2. Ejecución del tiempo del clasificador,
3. Precisión del clasificador.

Todas las pruebas se realizaron con una computadora con núcleo I7 y el microcontrolador dentro del guante inteligente.

3.4.3. Rendimiento del clasificador

Por último, para conocer el rendimiento del clasificador, se realizan dos pruebas: validación cruzada (en software) y número de éxito del número realizado por una persona al utilizar el guante desarrollado. El sistema tiene dentro de la matriz de entrenamiento y el algoritmo clasificador (hardware), como resultado, la cantidad de datos se reduce en un 98% y se logra un rendimiento de clasificación del 85% con el algoritmo evolutivo de CHC.

3.4.4. Fase de limpieza de los algoritmos

3.4.4.1. Generación de antecedentes y agrupación de prototipos

Esta fase es similar que la de los métodos descritos. La idea es volver a dividir el espacio en celdas, pero en vez de agrupar instancias se agrupan los prototipos generados. Una vez generados los antecedentes y dividido el espacio en celdas, agrupamos los prototipos. De esta manera se consigue separar los prototipos y crear conjuntos de prototipos que son muy parecidos entre ellos.

3.4.4.2. Limpieza de los prototipos por antecedente

Una vez agrupados los prototipos, cada antecedente tendrá asociados los prototipos que son más parecidos. De esta manera se pueden limpiar más fácilmente y comprobar cuáles son necesarios y cuáles no.

3.4.4.3. Limpieza mediante kNN

En esta fase se aplica la idea de kNN sobre los prototipos de cada antecedente

por separado. Una vez que se han agrupado los prototipos por celdas, se realizan operaciones individualmente por cada celda para limpiar los prototipos de ellas. Dada una celda se escoge un prototipo contenido en ella, con el prototipo escogido se calcula su distancia con el resto de ellas.

Una vez calculadas las distancias, se escoge los k prototipos con la distancia más corta al prototipo inicialmente escogido, dicho de otra manera, los k prototipos más parecidos. Si de esos k prototipos hay un número determinado que son de la misma clase que el prototipo escogido al principio, entonces se elimina este del conjunto inicial de prototipos.

Esta operación descrita se realiza con todos los prototipos de cada celda teniendo siempre en cuenta que, si se elimina un prototipo durante el proceso, este no se utiliza para realizar las comparaciones con los otros prototipos.

Algoritmo. Limpieza mediante kNN.

Entrada: E, conjunto de prototipos, k, número de vecinos a comparar $umbral$, numero límite de prototipos parecidos.
Salida: conjunto E limpiado

1. $F=E$
2. **Por cada** $e \in E$ **Hacer**
   a. $F' = F - e$
   b. $D = \emptyset$
   c. **Por cada** $f \in F'$ **hacer**
      i. $D=D+distancia\ (e, f)$
   d. **Fin Por**
   e. $D' = \min(D, F',)$
   f. **SI** $I(D', e,$ umbral$)$ **entonces** $F=F-e$
3. **Fin Por**
4. $E=F$
5. **FIN**

Figura 16. Algoritmo de limpieza mediante KNN
Fuente: (Iñiguez Jiménez, 2016)

- ✓ **min** (D, F', k)**:** Función que devuelve los k elementos de F' con las distancias asociadas D más bajas.
- ✓ $I(D', e, umbral)$**:** Función booleana que devuelve verdadero si hay un número mayor o igual que $umbral$ de elementos de D' de la misma clase que e. Falso en caso contrario.
- ✓ Tras la limpieza pasamos los prototipos que han quedado a la siguiente iteración.

3.4.4.4. Algoritmo de limpieza aleatorio

Este algoritmo sirve para hacer una limpieza breve en cada antecedente. Una vez agrupados los prototipos por celdas se realiza la misma operación por cada una de ellas. Escogemos k prototipos aleatoriamente, de los prototipos escogidos se compara sus clases con el primero de ellos, si hay un número de ellos de la misma clase que el primero mayor que cierto umbral, se elimina ese primer prototipo de la celda.

Algoritmo. Limpieza aleatoria.

Entrada: E, conjunto de prototipos asociados al mismo antecedente k, número de prototipos a comparar, $umbral$, número límite de prototipos parecidos.
Salida: conjunto E limpiado

1. F=*escogeAleatorios*(E, k)
2. e =*primero*(F)
3. $contador$=0
4. **Por cada** $f \in F$ **hacer**
   a. **SI** $eclase$=$fclase$ **entonces** contador = contador +1
5. **Fin Por**
6. **SI** $contador \geq umbral$ **entonces** eliminar el elemento e del conjunto E
7. **FIN**

Figura 17. Algoritmo de limpieza aleatoria
Fuente: (Iñiguez Jiménez, 2016)

- ✓ **La función *escoger Aleatorios* (E,):** devuelve un subconjunto de E con k elementos aleatorios.
- ✓ **La función *primero* (F):** devuelve el primer elemento del conjunto F.

3.4.4.5. Fin de la limpieza

Una vez que se han terminado todas las iteraciones, los prototipos que hayan quedado tras la fase de limpieza son los prototipos finales del algoritmo.

CAPÍTULO IV

ANÁLISIS DE LOS RESULTADOS

En este capítulo se explicarán los distintos modelos que se han creado en este proyecto junto con los detalles de implementación, parámetros usados y la idea detrás de ellos. La idea central de todas las versiones de los algoritmos es intentar mejorar la precisión y optimización.

3.5. METODOLOGÍA

En la figura 18 se indica la metodología propuesta (dividir los datos en el conjunto de entrenamiento y el conjunto de pruebas, el balance de datos, la selección de prototipos comparativos y el clasificador dentro del sistema electrónico).

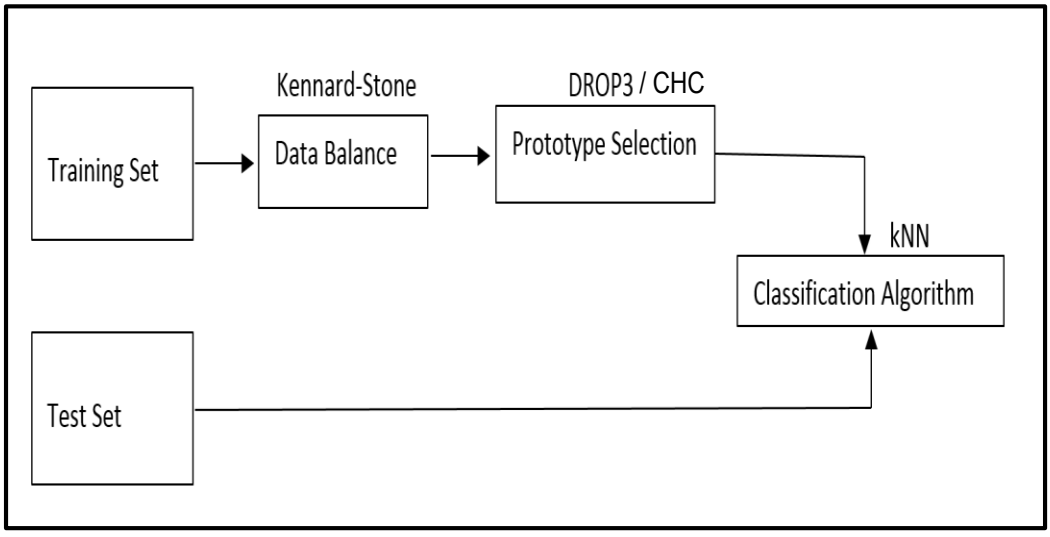


Figura 18. Diagrama de bloques para la metodología de análisis de datos para el sistema

3.6. EQUILIBRIO DE DATOS

La etapa de equilibrio de datos se realizó porque el ordenador que compiló los diferentes algoritmos de aprendizaje de máquina (específicamente con DROP3) tuvo problemas con los resultados, debido a razones de ejecución y los recursos de CPU. En esta fase se incluyen:

1. Datos de preprocesamiento
2. Selección comparativa de prototipos.

4.1. DATOS DE PREPROCESAMIENTO

4.1.1. Ruidos en el Dataset

El sistema electrónico instalado en el guante adquiere datos al flexionar cada dedo al realizar los diferentes gestos del lenguaje de señas, esto trae consigo la inconveniencia de que muchos datos sean almacenados considerados como "ruidos" por el cambio de posición de los dedos en cada gesto.

Otro problema es el tiempo que puede durar cada gesto, depende de cada usuario y de su experiencia en lenguaje de señas. La base de datos obtenida del guante ha sido llamada “T”, la cual contiene muchos de estos datos ruidosos, haciendo que la base de datos crezca en tamaño.

4.1.2. Balanceo de datos con KS

Al implementar complejos algoritmos de selección de prototipos a bases de datos de gran volumen, el tiempo de ejecución es muy alto debido a iteraciones repetidas con el conjunto de entrenamiento. Por esta razón se implementa un algoritmo de balance de datos, Kennard-Stone.

4.1.2.1. Funcionamiento de KS

El método de Kennard-Stone se encarga de seleccionar muestras de un gran volumen de datos con una distribución uniforme sobre el espacio predictor, de esta manera, la matriz que se obtiene de la base de datos llamada "T" se cambia a una matriz llamada "U" de p×n, donde p es igual a 1000.

En la figura 19 se muestra en la parte A la matriz T y en la parte B la matriz U. Además, para una mejor visualización de datos, se realizó la etapa de reducción de dimensionalidad para disminuir el número de dimensiones de cinco a dos dimensiones.

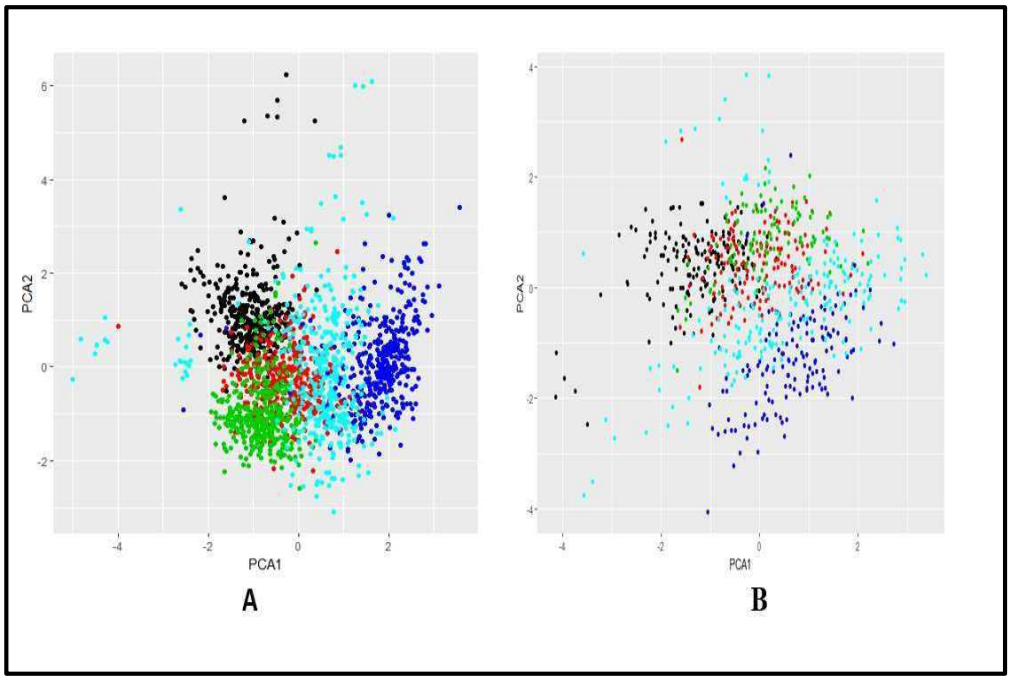


Figura 19. Base de datos de lenguaje de señas, parte A la matriz T y en la parte B la matriz U, colores: negro (número 1), rojo (número 2), verde (número 3), azul (número 4) y cian (número 5)

4.2. COMPARACIÓN DE LA SELECCIÓN DE PROTOTIPOS

Los dos algoritmos de selección de prototipos que se han estudiado en este trabajo, DROP3 y CHC, fueron compilados en R-studio con sus respectivas

bibliotecas, utilizamos la matriz U. Para realizar la función de compilar los algoritmos se utilizó una computadora con un procesador i7 y 10 gigabytes en RAM, en donde se analizaron los siguientes criterios:

1. (RI), instancias eliminadas
2. (TE), tiempo de ejecución
3. (CA), precisión del clasificador

4.2.1. División de datos (Entrenamiento y Prueba)

De acuerdo a la metodología planteada en este trabajo, se ha indicado que a la matriz U se la ha dividido en dos partes: datos de entrenamiento y datos de prueba. Para ello se ha realizado una división de forma aleatoria, quedando de la siguiente manera: entrenamiento (75%) y prueba (25%) de los datos. En la tabla 7 se muestra los resultados para para cada método, después de haber dividido los datos.

Tabla 7
Comparativa de resultados para métodos PS con el criterio: instancias removidas (RI), tiempo de ejecución (TE) y precisión del clasificador (CA)

	Algoritmo de Selección de Prototipo	RI	TE	CA
1	DROP3	64,26667%	30min	0,878
2	CHC	68,66%	2min	0,868

4.3. RESULTADOS

Los resultados de las matrices de entrenamiento se indican en las figuras 20 y 21 Para el beneficio de la visualización sólo se presentan los primeros cinco números con el proceso de reducción de dimensionalidad (DR).

4.3.1. Resultados para CHC

Por su parte, CHC descartó 515 instancias del conjunto de entrenamiento de un total de 750 puntos, es decir, se redujo un 68,8%. Además, en rendimiento del clasificador CHC obtuvo 86,8%.

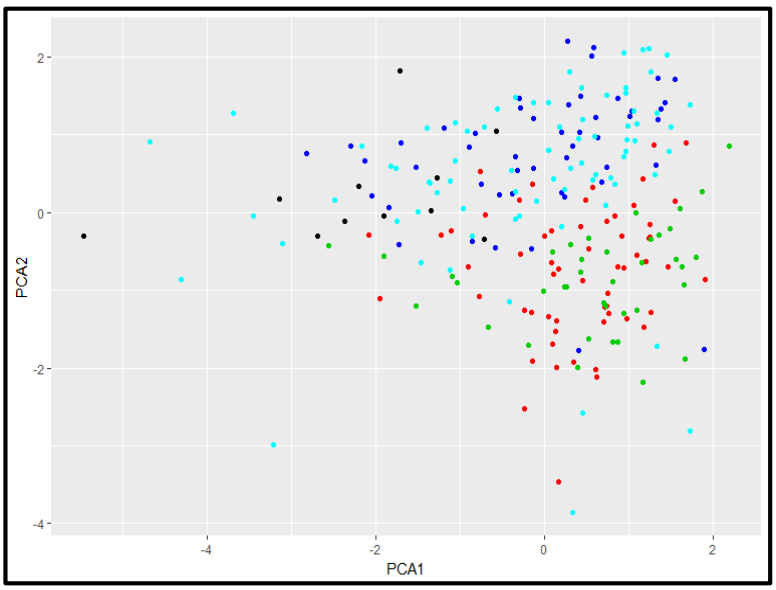


Figura 20. Conjunto de entrenamiento por algoritmo CHC, colores: negro (número 1), rojo (número 2), verde (número 3), azul (número 4) y cian (número 5)

4.3.2. Resultados para DROP3

Por su parte, DROP3 descartó 482 instancias del conjunto de entrenamiento de un total de 750 puntos, es decir, se redujo un 64,26%. Además, en rendimiento del clasificador DROP3 87,8%.

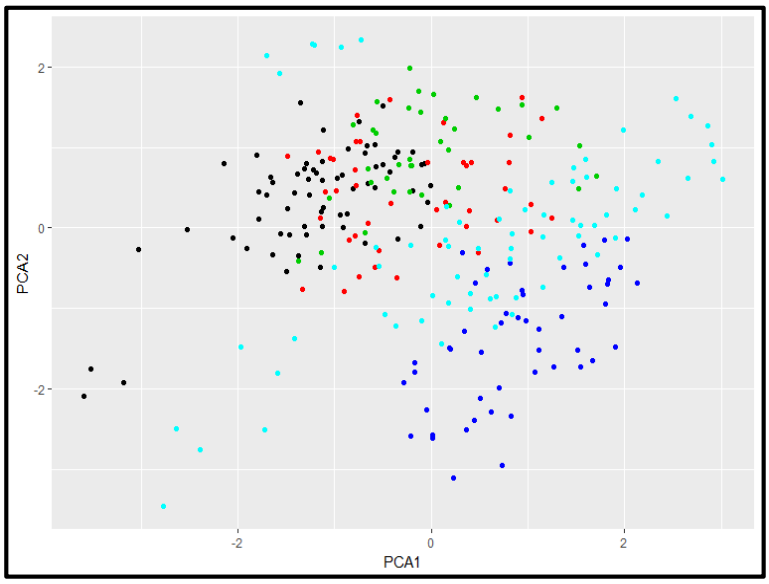


Figura 21. Conjunto de entrenamiento por algoritmo DROP3, colores: negro (número 1), rojo (número 2), verde (número 3), azul (número 4) y cian (número 5)

Para la implementación del algoritmo de clasificación, se eligió como la base de datos de entrenamiento obtenida por CHC, ya que elimina más instancias y su porcentaje de rendimiento sigue siendo alto. Además de que, al compilar el algoritmo, su tiempo de ejecución es significativamente más bajo. La matriz de entrenamiento final se llama V de qxn, donde q es igual a 268 instancias.

En referencia a la matriz de entrenamiento V (se aplica la selección de prototipos) y a la matriz U, el rendimiento del clasificador fue de 86,7% y 89,2%, respectivamente. Donde la matriz U tiene 482 casos más, es decir, 64% más de capacidad de almacenamiento.

En este caso para el tamaño de matriz U, no puede ser un almacenamiento dentro del Arduino, debido a que este microcontrolador tiene sólo 8 Kbytes de memoria.

Figura 22. Visualización del guante inteligente.

En la figura 22 se presenta el guante inteligente que va a ser usado con el conjunto de entrenamiento y el algoritmo clasificador. Como precisión para los nuevos datos entrantes el rendimiento del sistema fue 85% en cinco experimentos con cien nuevos números en lenguaje de signos.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

En este último capítulo se presentarán las conclusiones y recomendaciones de todos los puntos relevantes o de interés obtenidos del proyecto luego de haber realizado todo el proceso de desarrollo del mismo.

4.1. CONCLUSIONES

- En el trabajo se profundizó en los elementos teóricos-metodológicos de los algoritmos de machine learning para pequeños, medianos y grandes volúmenes de datos; de igual forma se realizó el análisis de los elementos electrónicos para el guante traductor, finalmente se realizó el estudio sobre pre procesamiento de datos debido a que es muy importante porque los datos reales pueden ser impuros, y, por ende, pueden conducir a la extracción de patrones/reglas poco útiles.

- Se investigó las características de cada componente electrónico que fue implementado en el guante traductor de señas básicas, donde los sensores flexibles resistivos se adaptaron de la mejor manera al mismo para el reconocimiento de las señas básicas, también se empleó la placa electrónica Lilypad Arduino y sus ventajas de uso con el hilo conductor que cumplió con los requerimientos tanto de sujeción al guante como para las respectivas conexiones; además la utilización de un módulo Bluetooth el cual permitirá establecer y mantener una comunicación inalámbrica, permitiendo al usuario tener la movilidad del guante traductor dentro de un área específica no más de 10 metros.

- El algoritmo CHC se ha convertido en una poderosa herramienta para obtener pequeños conjuntos de entrenamiento seleccionados y, por lo tanto, reducir los datos. CHC puede seleccionar las instancias más representativas independientemente de su posición en el espacio de búsqueda, satisfaciendo tanto los objetivos de alta precisión y las tasas de reducción, en cambio, la principal limitación de DROP3 es su largo tiempo de procesamiento, como se detalló en el desarrollo del trabajo, lo que dificulta la aplicación de este algoritmo a conjuntos de datos muy grandes.

- En referencia a los algoritmos de selección de prototipos CHC ha sido el ganador en el análisis con la matriz de datos propuesta por su capacidad de reducción de datos de entrenamiento y su desempeño del clasificador, su principal ventaja es la facilidad de compilación dentro del entorno de R; además, de la gran cantidad de datos adquiridos para el modelo, sólo el 2% se utiliza para entrenar el modelo con una precisión del 85%.

4.2. RECOMENDACIONES

- Este proyecto da la continuidad para el desarrollo de prototipos electrónicos que ayuden a facilitar la comunicación de las personas con discapacidad auditiva y de lenguaje dentro del entorno que los rodean; es por esto que se recomienda tener futuras investigaciones e implementaciones de prototipos electrónicos que apoyen al presente proyecto.

- Es recomendable desarrollar un prototipo que permita sensar un mayor número de señas básicas, con la finalidad de reducir el tiempo y aumentar la velocidad de conocimiento de las necesidades que presentan a diario las personas con discapacidad auditiva y de lenguaje hacia las personas que están al cuidado de las mismas que en este caso son sus familias.

- Como trabajos futuros se puede proponer realizar el proyecto para obtener una base completa de todas las letras y los números del lenguaje de signos, además extender la propuesta a otro grupo de personas con diferentes necesidades, es decir, para personas con diferentes capacidades especiales.

- A lo largo del estudio de los algoritmos se generan prototipos teniendo siempre en cuenta que no se genere ruido, sin embargo, cuando todos los prototipos generados son agrupados algunos podrían llegar a verse como ruido o innecesarios; el uso de una fase de limpieza al final del algoritmo no ha llegado a ayudar a limpiar el ruido, pero si lo hace a la hora de aumentar el porcentaje de reducción.

4.3. REFERENCIAS BIBLIOGRÁFICAS

Aha, D. K. (2007). Learning representative exemplars of concepts: An initial case of study. *4th Int. Workshop Machine Learning*, pp. 24–30.

Aha, D., Kbiler, D., & Albert, M. (1991). *Instance-Based learning algorithms.* Machine Learning.

Alvarado Perez, J., Peluffo, D., & Theron, R. (2015). Visualización y métodos kernel: Integración inteligencia natural y artificial. *SATHIRI*.

Alwakeel, S., Alhalabi, B., Aggoune, H., & Alwakeel, M. (2015). A Machine Learning Based WSN System for Autism Activity Recognition. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 771-776.

Arroyo-Hernández. (2016). Métodos de reducción de dimensionalidad: Análisis comparativo de los métodos APC, ACPP y ACPK. *UNICIENCIA*, pp. 115-122.

Baluja, S. (2004). *Population-based incremental learning*. Carnegie Mellon.

Bayardo, R., Agrawal, R., & Gunopulos, D. (2000). Reglas de Minerías de datos en grandes volúmenes de información. *Minería de datos y descubrimiento del conocimiento*, 217-240.

Berral-García, J. (2016). A Quick View on Current Techniques and Machine Learning Algorithms for Big Data Analytics. *ICTON* , 1-4.

Bishop, C. (2006). *Patttern Recognition and Machine Learning* . Cambridge: Springer.

Brownlee, J. (2015). *Machine Learning Algorithms* . Machine Learning Mastery.

Bush, C. R. (2001). Using genetic algorithms for training data selection in RBF networks. *Instance Selection and Construction for Data Mining*, pp. 339–356.

Calero, S. (2013). Nuevas tendencias mundiales en el proceso de dirección del entrenamiento deportivo. *Curso de Postgrado impartido en la Universidad de Guayaquil* (págs. 1-25). Guayaquil: Instituto de investigaciones.

Celada-Funes, E., Román, D., Asensio, M., & Beferull, B. (2014). A reliable CSMA protocol for high performance broadcast communications in a WSN. *Ad Hoc and Sensor Networking Symposium*, 473-484.

Ciudadano, E. (20 de Octubre de 2014). *Ecuador es un referente en inclusión a personas con discapacidad*. Obtenido de http://www.elciudadano.gob.ec/ecuador-es-un-referente-en-inclusion-a-personas-con-capacidades-distintas/

Eshelman, L. J. (2001). The adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination. *Foundations of Genetic Algorithms-1, G. J. E. Rawlins*, p. 265–283.

G. Bernieri, L. F. (2015). A low cost smart glove for visually impaired people mobility. *23rd Mediterranean Conference on Control and Automation (MED)*, pp. 130-135.

H. Brighton and C. Mellish. (2001). Identifying competence-critical instances for instance-based learners. *Instance Selection and Construction for Data Mining*, pp. 77–94.

H. Liu, H. M. (1998). Feature Selection for Knowledge Discovery .

Iñiguez Jiménez, L. (2016). *Generación de prototipos para clasificación en entornos Big Data.*

J. R. Cano, F. H. (2003). Using evolutionary algorithms as instance selection for data reduction in kdd: an experimental study. *IEEE Transactions on Evolutionary Computation*, pp. 561-575.

K. S. Abhishek, L. C. (Aug 2016). Glove-based hand gesture recognition sign language translator using capacitive touch sensor. *IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, pp. 334–337.

Lee, J. y. (2007). Nonlinear Dimensionality Reduction. *Springer - Science & Business*, pp. 125-32.

Lin., T. Y. (2002). Attribute Transformation for Data Mining I: Theoretical Explorations. *International Journal of Intelligent Systems*, 213-222.

M. B. H. Flores, C. M. (2014). Useroriented finger-gesture glove controller with hand movement virtualization using flex sensors and a digital accelerometer. *International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pp. 1–4.

Martinez, D. W. (2000). Reduction techniques for instance-based learning algorithms. *Machine Learning*, pp. 257–286.

Mesa, F. D. (2008). Algoritmos de Aprendizaje Continuo Mediante Selección de Prototipos para Clasificadores Basados en Distancias. En F. D. Mesa. Castellón.

Moreno García, M., & López Batista, V. (2004). Uso de Tecnicas no Supervisadas en la Construcción de modelos de Clasificación en Ingeniería del Software. *Departamente de Informática y Automática. Univerisdad de Salamanca*.

Motoda, H. L. (1998). *Feature Selection for Knowledge Discovery and.* Kluwer.

Nooritawati, M. E.-A. (March 2012). Review in sign language recognition systems. *IEEE Symposium on Computers Informatics (ISCI)*, pp. 52–57.

On issues of instance selection. (2002). *Data Mining and Knowl*, pp. 115–130.

P. Chapman, J. C. (1999). *The CRISP-DM Process Model*. Obtenido de www.crispdm.org/pub-paper.pdf

P. Rosero-Montalvo, P. D.-C.-U.-I.-P.-O. (2017). Interactive Data Visualization Using Dimensionality Reduction and Similarity-Based Representations. pp. 334–342.

Pérez López, C., & Santín Gonzáles, D. (2008). *MINERÍA DE DATOS TÉCNICAS Y HERRAMIENTAS.* Madrid, Espana: Thomson Ediciones Parainfo, S.A.

Planificación, C. N. (17 de Febrero de 2013). *Plan Nacional del buen vivir 2013-2017.* Obtenido de https://www.unicef.org/ecuador/Plan_Nacional_Buen_Vivir_2013-2017.pdf

Pyle, D. (1999). *Data Preparation for Data Mining* . Morgan Kaufmann Publishers.

Rayón, A. (31 de Agosto de 2017). *El entrenamiento del modelo con datos y los problemas de "underfitting" y "overfitting"*. Obtenido de http://i.stack.imgur.com/0NbOY.png

Rayón, A. (31 de Agosto de 2017). *El entrenamiento del modelo con datos y los problemas de "underfitting" y "overfitting"* .

Restrepo, A. M. (8 de Agosto de 2011). *Muestreo de espectros para calibración por*

PLS - Algoritmos de Kennard-Stone. Obtenido de http://quimio-metria.blogspot.com/2011/08/muestreo-de-espectros-para-calibracion.html

Rojas Soriano, R. (2006). *Guía para realizar investigaciones sociales.* San Rafael, México: Plaza y valdes.

Rosero-Montalvo, P., Diaz, P., Salazar-Castro, J., & Peluffo-Órdonez, D. (2016). Interactive Data de Visualization Using Dimensionality Reduction and Similarity-Based Representations. *IberoAmerican Congress on Pattern Recognition*, 100-108.

S. Garcia, J. D. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 417–435.

S. Nunez-Godoy, V. A.-P.-G.-C.-E.-I.-C.-M. (2016). Human-sitting-pose detection using data classification and dimensionality reduction. *IEEE Ecuador Technical Chapters Meeting (ETCM)*, pp. 1–5.

S. Zhang, C. Z. (2003). Data preparation for data mining Applied Artificial Intelligence. *Special Issue Data Cleaning and Preprocessing*, 17:5-6, 375-381.

Salazar-Castro, J., Pena-Unigarro, D., & Peluffo-Órdonez, R. (2016). Dimsensionality Reduction for Interactive Data Visualization via a Geo-Desic Approach. *Latin American Conference on Computational Intelligence*.

Shlens, J. (2005). *A Tutorial on Principal Component Analysis.* Obtenido de http://arxiv.org/pdf/1404.1100v1.pdf

Sierra Araujo, B. (2006). *Aprendizaje Automático: conceptos básicos y avanzados.* Espana: Pearson .

Stone, R. W. (2006). Computer Aided Design of Experiments, Technometrics.

T. Back, D. F. (2007). *Handbook of Evolutionary Computation.* London: Oxford Univ. Press.

V. Detours, J. E. (2003). Integration and cross-validation of high-throughout gene expression data: comparing heterogeneous data sets. *FEBS Letters*, 98-102.

Vivir, S. d. (27 de 10 de 2016). *Buen Vivir.* Obtenido de Buen Vivir: http://www.secretariabuenvivir.gob.ec/cambio-cultural-y-el-buen-vivir/

W. Kim, B. C.-K.-K. (2003). Data Mining and Knowledge Discovery 7. *A Taxonomy of Dirty Data*, 81-99.

Whitley, D. (2009). The GENITOR algorithm and selective preasure:Why rank based allocation of reproductive trials is best. *3rd Int. Conf. GAs, Schaffer*, pp. 116–121.

Wilson, D. R. (1997). *Machine learning: Proceedings of the Fourteeth International Conference.* San Francisco: Morgan Koufmann Publishers.

Witten, E. F. (1999). Making better use of global discretization. *16th Int. Conf. Machine Learning*, pp. 115–123.

Zafer, S., Turgay Tugay, B., & Cho, J. (2010). Fall Detection by Using K-Nearset Neighbor Algorithm on WSN data. *IEEE Globecom 2010 Workshop on Advances in Comunications adn Networks*, 2054-2058.