UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

ANÁLISIS DE FIABILIDAD DEL ÁRBOL DE DECISIÓN J48 EN LA CORRECTA CLASIFICACIÓN DE DOCUMENTOS EN LA UNIDAD DE GESTIÓN EDUCATIVA LOCAL EL COLLAO-ILAVE

TESIS

PRESENTADA POR:

Bach. JEANMARCO QUENTA BANEGAS

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PUNO – PERÚ

2021

DEDICATORIA

A mis padres, por haberme brindado su apoyo constante todos estos años, por sus consejos, paciencia y ayudarme a alcanzar uno de mis anhelos más deseados.

AGRADECIMIENTOS

A mi familia, por estar presente, acompañarme y brindarme su apoyo moral a lo largo de mi carrera universitaria. Agradezco también a aquellas personas que me apoyaron compartiendo su conocimiento durante este proceso.

Agradezco a la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Altiplano de Puno y a sus docentes por haber compartido sus conocimientos y enriquecer los míos durante mi formación académica.

Gracias a todas las personas que ayudaron directa e indirectamente en la realización de esta investigación.

Jeanmarco Quenta Banegas

ÍNDICE GENERAL

DEDICATORIA

AGRADECIMIENTOS

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ACRÓNIMOS

RESUMEN 10

ABSTRACT 11

CAPÍTULO I

INTRODUCCIÓN

1.1. FORMULACIÓN DEL PROBLEMA 13

1.1.1. Problema general 13

1.1.2. Problemas específicos 13

1.2. JUSTIFICACIÓN 13

1.3. PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN 14

1.3.1. Descripción del problema 14

1.4. LIMITACIONES 15

1.5. OBJETIVOS 16

1.5.1. Objetivo General 16

1.5.2. Objetivos específicos 16

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES DE INVESTIGACIÓN 17

2.2. SUSTENTO TEÓRICO 18

2.2.1. Inteligencia Artificial 18

2.2.2. Árboles de decisión 22
2.2.3. Algoritmo C4.5 24
2.2.4. Árbol de decisión J48 27
2.2.5. Bases de datos 27
2.2.6. Normalización de datos 29
2.2.7. Normas ISO 32
2.2.8. ISO 25000 32
2.2.9. Algoritmo 32
2.2.10. Pseudocódigo 34
2.2.11. Data set 35
2.2.12. Métodos de análisis de datos 37
2.3. GLOSARIO DE TÉRMINOS BÁSICOS 39
2.4. HIPÓTESIS DE LA INVESTIGACIÓN 42
2.4.1. Hipótesis General 42

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. MÉTODOS 43
3.1.1. Tipo de investigación 43
3.1.2. Diseño de investigación 43
3.2. POBLACIÓN Y MUESTRA 44
3.2.1. Población 44
3.2.2. Muestra 44
3.3. UBICACIÓN DE LA POBLACIÓN 45
3.4. INSTRUMENTOS DE RECOLECCIÓN DE DATOS 45
3.4.1. Instrumentos 45
3.4.2. Validación y confiabilidad del instrumento 46
3.5. PROCEDIMIENTO DEL EXPERIMENTO 46
3.6. PLAN DE PROCESAMIENTO Y ANÁLISIS DE DATOS 46

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. DATA SET RESULTANTE DEL TRATAMIENTO DE DOCUMENTOS 47

4.1.1. Datos obtenidos 47

4.1.2. Normalización de los datos obtenidos 49

4.1.3. Data set no analizada 63

4.1.4. Prueba Chi-cuadrada 63

4.1.5. Análisis de componentes principales categórico (CATPCA) 65

4.1.6. Data set Final 66

4.1.7. Ocurrencias encontradas 67

4.2. CONFIGURACIÓN DEL ÁRBOL DE DECISIÓN J48 68

4.2.1. Ejemplo de funcionamiento. 70

4.2.2. Ocurrencias encontradas 72

4.3. RESULTADOS DE FIABILIDAD DEL ÁRBOL DE DECISIÓN J48 73

4.3.1. Interpretación de la matriz de confusión 74

4.4. DISCUSIÓN 77

V. CONCLUSIONES 79

VI. RECOMENDACIONES 80

VII. REFERENCIAS 81

ANEXOS 84

Área : Inteligencia Artificial y Sistemas Bio-inspirados

Tema : Árboles de decisión

FECHA DE SUSTENTACIÓN: 6 DE AGOSTO DEL 2021

ÍNDICE DE FIGURAS

Figura 1: Estructura de un árbol de decisión 23

Figura 2: Formato General Pseudocódigo 35

Figura 3: K-fold validación cruzada 37

Figura 4: Diagrama simplificado de funcionamiento del árbol de decisión J48 50

Figura 5: Diagrama proceso de eliminación/reemplazo de información irrelevante 54

Figura 6: Diagrama simplificado, proceso de identificación de palabras clave 56

Figura 7: Esquema de datos de entrada para la data set 60

Figura 8: Proceso de creación de los datos de entrada de la data set 61

Figura 9: Configuración de la data set no analizada..... 63

Figura 10: Nivel de significancia, Características vs Resultados..... 64

Figura 11: Coeficientes de Correlación de las Características 65

Figura 12: Configuración de la data set final 67

Figura 13: Configuración árbol de decisión J48 obtenido..... 69

Figura 14: Representación del paso 1. CASO 1 71

Figura 15: Representación del paso 2, primera iteración. CASO 1..... 71

Figura 16: Representación del paso 2, segunda iteración. CASO 1 72

Figura 17: Representación del paso 3. CASO 1 72

Figura 18: Matriz de confusión obtenida..... 73

ÍNDICE DE TABLAS

Tabla 1: Formato de resumen de datos obtenidos de los documentos 48

Tabla 2: Reemplazo y/o eliminación de información irrelevante de la data set 52

Tabla 3: Proceso de identificación de palabras clave de la data set 55

Tabla 4: Característica Tipo Monetaria 57

Tabla 5: Característica Relación de afinidad con la persona 58

Tabla 6: Característica Tratamiento del documento 58

Tabla 7: Característica Tipo Entidad Involucrada 59

Tabla 8: Característica Tipo Proceso 59

Tabla 9: Característica Estado 60

Tabla 10: Resultados de obtención de los datos de entrada de la data set 62

Tabla 11: Resultados datos de salida de la data set 62

Tabla 12: Explicación de figuras presentes en el árbol de decisión J48 70

Tabla 13: Resumen de exactitud árbol J48 por Oficina 76

ÍNDICE DE ACRÓNIMOS

UGEL: Unidad de Gestión Educativa Local.

G.A.: Área de Gestión Administrativa de la UGEL.

I.A.: Inteligencia Artificial (Artificial Intelligence)

CETPRO: Centros de Educación Técnico-Productiva

IDE: Integrated Development Environmen (Entorno de desarrollo integrado).

RESUMEN

En la actualidad la necesidad de automatizar los procesos se han vuelto una prioridad para mejorar el rendimiento y eficiencia de una persona u organización, este proceso puede ser abordado desde un punto de vista de hardware o software, por esta razón la presente investigación tuvo como objetivo analizar la fiabilidad del uso del árbol de decisión J48 como una herramienta fiable de clasificación de documentos en la Unidad de Gestión Educativa Local El Collao-Ilave, entidad dedicada al manejo del aspecto educativo de la provincia El Collao, la principal característica del J48 es la gran adaptabilidad a los datos y la gran precisión que tiene dentro de los demás árboles de decisión en problemas de clasificación, resolviendo así necesidades específicas de la institución y mejorando el flujo de información y/o documentos dentro de la misma. El presente trabajo de investigación implicó primeramente la obtención de los datos de entrada a partir de los documentos que ingresaron a la UGEL siendo necesaria una normalización o procesamiento de los mismos para la adecuación al formato de datos de entrada del árbol J48, después de ello se da paso a la programación del árbol de decisión J48 siendo la base del trabajo, además una evaluación de la calidad del programa que será dado por la norma ISO 25000. Para este trabajo de investigación fue necesario apoyarse en una metodología cuantitativa. Adicionalmente se evaluó el grado de precisión que tuvo en la clasificación de documentos contenidos en la data set creada, el cual fue desarrollado a partir de los documentos que ingresaron a la UGEL El Collao-Ilave.

Palabras Clave: Árbol de decisión, J48, Clasificación, Data set, Stemming.

ABSTRACT

Currently the need to automate processes have become a priority to improve the performance and efficiency of a person or organization, this process can be approached from a hardware or software point of view, for this reason the present research aimed to analyze the reliability of the use of the J48 decision tree as a reliable tool for document classification in the Local Educational Management Unit El Collao-Ilave, The main characteristic of the J48 is the great adaptability to data and the great precision that it has within the other decision trees in classification problems, thus solving specific needs of the institution and improving the flow of information and/or documents within it. The present research work involved first obtaining the input data from the documents that entered the UGEL being necessary a normalization or processing of the same for the adequacy to the input data format of the J48 tree, after that it gives way to the programming of the J48 decision tree being the basis of the work, in addition to an evaluation of the quality of the program that will be given by the ISO 25000 standard. For this research work it was necessary to rely on a quantitative methodology. Additionally, the degree of accuracy in the classification of documents contained in the created data set was evaluated, which was developed from the documents that entered the UGEL El Collao-Ilave.

Key words: Decision tree, J48, Classification, Data Set, Stemming.

CAPÍTULO I

INTRODUCCIÓN

Actualmente los programas de seguimiento y/o administración de documentos usados en la UGEL El Collao-Ilave permiten identificar, ordenar y monitorear el estado de un documento presentado en mesa de partes de la UGEL. Cada vez son más los documentos que ingresan a la UGEL para ser atendidos, pero muchas veces dichos documentos son enviados a la oficina equivocada por el personal encargado de recepcionarlos y clasificarlos. Una solución al problema es la automatización de dicho proceso haciendo uso de una herramienta que clasifique de manera correcta los documentos.

Por tal efecto la necesidad de automatizar ciertos procesos en pos de mejorar la calidad de servicio que se puede ofrecer se incrementa y por consiguiente se buscan soluciones que son mayormente herramientas informáticas o algoritmos. Ya que los algoritmos son imparciales y una vez implementados cumplen sus funciones según lo establecido por el programador pueden realizar una tarea concreta de una manera más eficiente.

Es por ello que surgió la idea de evaluar la fiabilidad del árbol de decisión J48 en la correcta clasificación de los documentos del área de Gestión Administrativa de la UGEL El Collao-Ilave, ello con el objetivo de evaluar una posible implementación del modelo en la clasificación de los documentos que ingresen a la UGEL.

La presente investigación consta de: Capítulo I, se hace referencia al planteamiento del problema de investigación, en base a ello se formula el problema, la justificación de la investigación, las limitaciones del estudio y los objetivos; en el

Capítulo II, está referido a los antecedentes que tienen relación con el trabajo de investigación, el sustento teórico y glosario de términos. El Capítulo III se menciona la metodología de la investigación, la población, los métodos, técnicas de medición de datos. En el Capítulo IV se muestran los resultados y discusión de la misma, a los cuales se llegó durante la investigación, así como la presentación de conclusiones y recomendaciones.

Finalmente, se hace mención de las referencias bibliografía y anexos.

1.1. FORMULACIÓN DEL PROBLEMA

1.1.1. Problema general

¿El árbol de decisión J48 será una solución fiable para la correcta clasificación de los documentos de la UGEL El Collao-Ilave?

1.1.2. Problemas específicos

a) ¿Cuál será el proceso para convertir los documentos en una data set usable por el árbol de decisión J48?
b) ¿Qué configuración tendrá el árbol de decisión J48 para desarrollar el proceso de clasificación?
c) ¿En qué medida es fiable usar el árbol de decisión J48 en la correcta clasificación de documentos en la UGEL El Collao-Ilave?

1.2. JUSTIFICACIÓN

Los arboles de decisión, una herramienta creciente de la Inteligencia Artificial, son usados actualmente para solucionar problemas de clasificación, y en el ámbito de la UGEL El Collao-Ilave gran cantidad de documentos usualmente son extraviados o enviados a una oficina equivocada lo cual genera desorden dentro de la institución, por

consiguiente, un árbol de decisión nos podría ayudar en la solución de los problemas de correcta clasificación de documentos.

Así pues, antes de usar un árbol de decisión como una solución al problema, debemos comprobar que dicho árbol de decisión produzca resultados fiables por tanto se estableció como ámbito de estudio el área de Gestión Administrativa de la UGEL El Collao-Ilave y de ahí generalizar para toda la UGEL. El árbol de decisión que se evaluará será el J48, se escoge este tipo de árbol de decisión ya que dicho árbol de decisión fue desarrollado por Ross Quinlan, expresamente para resolver problemas de clasificación estadística.

Hoy en día es necesario evaluar las posibles soluciones que se implementaran para resolver un problema específico, en especial cuando lo que se quiere mejorar es la calidad del servicio a las personas a través de pequeños detalles, tales como el correcto envío de sus documentos a las oficinas que darán respuesta al documento.

1.3. PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.3.1. Descripción del problema

Muchas veces un documento que ingresa al UGEL El Collao-Ilave es enviado a una oficina equivocada, pues el porcentaje de exactitud o precisión en la clasificación documentos en el área de Gestión Administrativa de la UGEL El Collao-Ilave es del 80.24% y el porcentaje de error en la clasificación de los mismos es del 19.76% (véase anexo 6), generando así malestar en la población usuaria de los servicios de la UGEL.

En tal sentido, la necesidad de implementar un modelo de clasificación de documentos se incrementa, pero antes de ello, es necesario evaluar si dicho modelo cumple con la tarea de clasificar los documentos de manera correcta. Para este trabajo de investigación, el modelo evaluado fue el árbol de decisión J48; el cual fue configurado

teniendo en cuenta el esquema de funcionamiento u organigrama de la UGEL El Collao-Ilave y del área de Gestión Administrativa (véase anexo 1 y 2).

Una de las ventajas más útiles que proporciona el análisis de un modelo antes de su implementación final radica en las diferentes características que se le pueden añadir durante el proceso de desarrollo de dicho modelo. Además de brindar una visión generalizada del problema y sus posibles soluciones.

Por lo antes mencionado surge la necesidad de analizar una posible solución antes de ser implementada, dicho de otra forma, es necesario evaluar si el árbol de decisión J48 garantiza una correcta clasificación de los documentos dentro del área de Gestión Administrativa y establecer una correlación con los documentos que ingresan a la UGEL El Collao-Ilave.

1.4. LIMITACIONES

Algunas de las limitaciones encontradas fueron escasez bibliográfica sobre estudios relacionados al tema específico, además de no existir ninguna data set disponible para una evaluación rápida de la fiabilidad del árbol de decisión J48 por lo cual se tuvo que elaborar uno propio con el cual se trabajó, también al momento de realizar la verificación de los resultados obtenidos en la clasificación se tuvo que ir oficina por oficina para verificar si el resultado era correcto, pues de dicha evaluación se establece si el árbol J48 cumple con lo requerido. El tiempo para la conformación de la data set es relativamente corto para aseverar que la data generada es altamente representativa. Pues el tiempo para generar una data así sería superior a 1 año. Finalmente, la falta de tiempo por parte de las personas a cargo de las oficinas para realizar un análisis más profundo de los documentos que su oficina atiende y mejorar así el resultado obtenido durante el desarrollo del trabajo de investigación.

1.5. OBJETIVOS

1.5.1. Objetivo General

Analizar la fiabilidad del uso del árbol de decisión J48 en la clasificación de documentos de la UGEL El Collao-Ilave.

1.5.2. Objetivos específicos

- Generar una data set a partir de los documentos de la UGEL El Collao-Ilave, el cual será usado por el árbol de decisión J48.
- Configurar el árbol de decisión J48 a partir de la data set generada a partir de los documentos de la UGEL El Collao-Ilave.
- Determinar la fiabilidad del árbol de decisión j48 en la correcta clasificación de documentos.

CAPÍTULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES DE INVESTIGACIÓN

Según Karabadji et al. (2017), en el artículo científico "An Evolutionary Scheme for Decision Tree Construction" los investigadores presentan un esquema de cómo construir un árbol de decisión mejorando los ajustes que puede sufrir este durante su creación. Dicho artículo contribuyo en la creación del árbol de decisión J48 usado en el trabajo de investigación, pues brindó una visión de modificaciones posibles a la configuración básica que presenta un árbol de decisión. Además, dicha investigación mostro un incremento en el accuracy de clasificación del árbol resultante, dichos resultados coinciden la hipótesis planteada en el presente trabajo, el cual busca establecer al árbol de decisión J48 como una herramienta fiable en la clasificación de los documentos de la UGEL El Collao-Ilave por lo cual se toma en cuenta las recomendaciones presentadas en dicho artículo científico.

(Karabadji, Seridi, Bousetouane, Dhifli, 2016), en el artículo científico "The value of decisión tree analysis in planning anaesthetic care in obstetrics" los investigadores estudian los beneficios y las limitaciones del uso de análisis del árbol de decisión se revisan y se discute su aplicación en la anestesia obstétrica. El objetivo de su investigación era estudiar la viabilidad de usar los arboles de decisión en su uso para la aplicación de anestesia obstétrica, y sus resultados brindan recomendaciones a considerar para el uso de variables de entrada para un árbol de decisión.

(Kastrati et al., 2019), en el artículo científico "The impact of Deep learning on document classification using semantically rich representations" los investigadores

exponen como el enriquecimiento de la semántica ayuda mejorar la clasificación usando aprendizaje profundo. El artículo brindó un paisaje más esclarecedor en lo que respecta al tratamiento de las oraciones o frases que componían los datos de entrada para el árbol de decisión J48 y como el tratamiento o normalización de la entrada ayudaba a generar mayor accuracy lo cual coincide con la razón de analizar la fiabilidad del árbol de decisión J48 en la clasificación de documentos antes de su implementación

(Zhao et al., 2016), en el artículo científico "Uncertain XML documents classification using Extreme Learning Machine" los investigadores exponen como usaron el Machine Learning para la clasificación de documentos XML inciertos. En esta investigación abordan el uso de Inteligencia Artificial para generar una herramienta de clasificación de documentos, los resultados de esa investigación sirvieron de precedente para realizar modificaciones a los datos de entrada que usamos en el presente trabajo de investigación.

2.2. SUSTENTO TEÓRICO

2.2.1. Inteligencia Artificial

"La IA es la rama de la ciencia que se encarga del estudio de la inteligencia en elementos artificiales y, desde el punto de vista de la ingeniería, propone la creación de elementos que posean un comportamiento inteligente. Dicho de otra forma, la IA pretende construir sistemas y máquinas que presenten un comportamiento que si fuera llevado a cabo por una persona, se diría que es inteligente."(Romero et al., 2007)

En el actual contexto nos encontramos inmersos en una sociedad que se orienta, cada vez más, hacia el proceso de la tecnificación masiva. Cada cierto tiempo, y con enormes avances, todos los sectores que la estructuran están, en cierta medida sometiéndose en algunos caos o adecuándose en otros a los avances de la tecnología y,

de acuerdo a su nivel de desarrollo alcanzado, adaptándose frente a tan inevitable tendencia. El área de la educación (que es sensible a los cambios en la sociedad ya que avanza a la par de la misma) también se encuentra atravesando dicha tendencia ineluctable de adaptación a las novedosas comunidades de interacción tecnológica; proceso que está orientado a nuevas tendencias y perfiles en relación a las nuevas propuestas en el sector. Pero, cabe la interrogante crucial ¿Hasta qué nivel la tecnología es capaz de revolucionar el universo de la educación?

El asumir de forma estructural un parámetro tan novedoso y a su vez vertiginoso, requiere del desarrollo y aplicaciones cada vez más impactantes, tanto así como las discrepancias y temores que se suscitan en relación a la aplicación de la inteligencia artificial (IA), debe ser punto clave en las discusiones de trascendencia en relación a la novedosas propuestas en educación superior y asumir al mismo tiempo los parámetros que permitan una mejor administración de este importante mecanismo, así como la aplicabilidad de políticas efectivas, cada vez más adecuadas que vitalicen de forma equilibrada las posibilidades de la IA, en función de las necesidades de las instituciones más representativas de la sociedad (tal como es el caso de las universidades) y por ende, sean los ciudadanos los beneficiarios de estas mediadas acertadas.(Ocaña-Fernández et al., 2019)

Comportamiento humano: el enfoque de la Prueba de Turing

La Prueba de Turing, propuesta por Alan Turing (1950), se diseñó para proporcionar una definición operacional y satisfactoria de inteligencia. En vez de proporcionar una lista larga y quizá controvertida de cualidades necesarias para obtener inteligencia artificialmente, él sugirió una prueba basada en la incapacidad de diferenciar entre entidades inteligentes indiscutibles y seres humanos. El computador supera la

prueba si un evaluador humano no es capaz de distinguir si las respuestas, a una serie de preguntas planteadas, son de una persona o no. Hoy por hoy, podemos decir que programar un computador para que supere la prueba requiere un trabajo considerable. El computador debería poseer las siguientes capacidades:

- **Procesamiento de lenguaje natural** que le permita comunicarse satisfactoriamente.
- **Representación del conocimiento** para almacenar lo que se conoce o siente.
- **Razonamiento automático** para utilizar la información almacenada para responder a preguntas y extraer nuevas conclusiones.
- **Aprendizaje automático** para adaptarse a nuevas circunstancias y para detectar y extrapolar patrones.

La Prueba de Turing evitó deliberadamente la interacción física directa entre el evaluador y el computador, dado que para medir la inteligencia es innecesario simular física- mente a una persona. Para superar la Prueba Global de Turing el computador debe estar dotado de:

- **Visión computacional** para percibir objetos.
- **Robótica para manipular** y mover objetos.

Estas seis disciplinas abarcan la mayor parte de la IA, y Turing merece ser reconocido por diseñar una prueba que se conserva vigente después de 50 años. Los investigadores del campo de la IA han dedicado poco esfuerzo a la evaluación de sus sistemas con la Prueba de Turing, por creer que es más importante el estudio de los principios en los que se basa la inteligencia que duplicar un ejemplar. La búsqueda de un ingenio que «volara artificialmente» tuvo éxito cuando los hermanos Wright, entre otros, dejaron de imitar a los pájaros y comprendieron los principios de la aerodinámica. Los

textos de ingeniería aerodinámica no definen el objetivo de su campo como la construcción de «máquinas que vuelen como palomas de forma que puedan incluso confundir a otras palomas».(Russell, 2004)

Los **campos de aplicación** de la inteligencia artificial son muchos, y algunos están orientados a satisfacer necesidades muy distintas:

a. **Machine learning o aprendizaje automático**

   El Machine Learning es la rama de la ciencia que busca el desarrollo de técnicas que permitan a los ordenadores aprender por sí mismos. Para ello se crean programas que pueden **generalizar ciertas respuestas** a partir de información sin estructurar, que se suministra como ejemplos. Con ello, se induce al conocimiento por parte del ordenador.

b. **Sistemas expertos**

   Hace referencia a un sistema de información que se basa en el conocimiento de un área de aplicación de gran **complejidad** y muy específica. Sirve como asistente consultor y experto para los usuarios de su interfaz.

   Son entornos que proporcionan respuestas sobre problemáticas muy específicas, pudiendo realizar inferencias muy parecidas a las de un ser humano acerca de los conocimientos concretos consultados.

c. **Redes neuronales artificiales**

   Estas redes son un paradigma del aprendizaje y los procesamientos automáticos, inspirado todo ello en el modo en que funciona el sistema nervioso de los animales. Consiste en un sistema de interconexión de neuronas en una red que colaboran entre ellas para crear una respuesta de salida.

d. Procesamiento del lenguaje natural

Es una disciplina de la rama de la ingeniería para la lingüística computacional. Se utiliza para la **formulación e investigación** de mecanismos de eficacia informática para servicios de comunicación entre las personas o entre ellas y las máquinas usando lenguajes naturales.

Los campos de desarrollo e investigación de la inteligencia artificial sirven para el desarrollo de nuevos mecanismos y aplicaciones que permitan diseñar nuevos métodos de trabajar y comunicar con las máquinas y los entornos informáticos.

2.2.2. Árboles de decisión

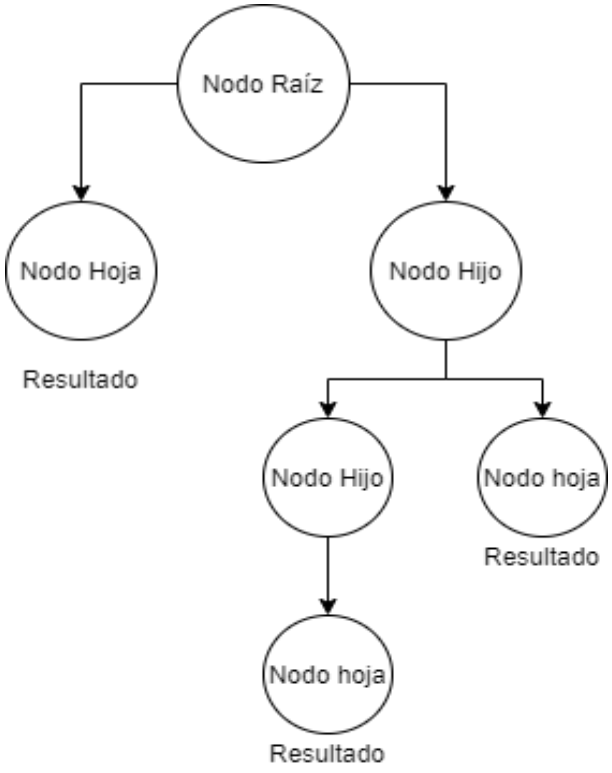
Un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema. Constituyen probablemente el modelo de clasificación más utilizado y popular.

El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol. Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema.

Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver (Ver Figura 1).

Este modelo se construye a partir de la descripción narrativa de un problema, ya que provee una visión gráfica de la toma de decisión, especificando las variables que son evaluadas, las acciones que deben ser tomadas y el orden en el que la toma de decisión será efectuada. Cada vez que se ejecuta este tipo de modelo, sólo un camino será seguido dependiendo del valor actual de la variable evaluada. Los valores que pueden tomar las variables para este tipo de modelos pueden ser discretos o continuos.(Barrientos et al., 2009)

Figura 1: Estructura de un árbol de decisión



Elaboración propia

Un algoritmo de generación de árboles de decisión consta de 2 etapas: la primera corresponde a la inducción del árbol y la segunda a la clasificación.

a. En la primera etapa se construye el árbol de decisión a partir del conjunto de entrenamiento; comúnmente cada nodo interno del árbol se compone de un

atributo de prueba y la porción del conjunto de entrenamiento presente en el nodo es dividida de acuerdo con los valores que pueda tomar ese atributo. La construcción del árbol inicia generando su nodo raíz, eligiendo un atributo de prueba y dividiendo el conjunto de entrenamiento en dos o más subconjuntos; para cada partición se genera un nuevo nodo y así sucesivamente. Cuando en un nodo se tienen objetos de más de una clase se genera un nodo interno; cuando contiene objetos de una clase solamente, se forma una hoja a la que se le asigna la etiqueta de la clase.

b. En la segunda etapa del algoritmo cada objeto nuevo es clasificado por el árbol construido; después se recorre el árbol desde el nodo raíz hasta una hoja, a partir de la que se determina la membresía del objeto a alguna clase. El camino a seguir en el árbol lo determinan las decisiones tomadas en cada nodo interno, de acuerdo con el atributo de prueba presente en él.

2.2.3. Algoritmo C4.5

El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero (depth-first). C4.5 construye árboles de decisión desde un grupo de datos de entrenamiento de la misma forma en que lo hace ID3, usando el concepto de entropía de información.

El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. En cada

nodo, el sistema debe decidir cuál prueba escoge para dividir los datos.(López Takeyas, 2005)

Algoritmo que aprende a partir de la diferencia que existe entre los datos para analizar, esto es, un procedimiento de divide y vencerás, que maximiza la información obtenida, la cual se utiliza como una métrica para seleccionar el mejor atributo que divida los datos en clases homogéneas.

Pseudocódigo

- Comprobar los casos base.
- Para cada atributo "**a**".
- Encontrar la ganancia de información normalizada de la división de a.
- Dejar que **a_best** sea el atributo con la ganancia de información normalizada más alta.
- Crear un nodo de decisión que divida **a_best.**
- Repetir en las sublistas obtenidas por división de **a_best**, y agregar estos nodos como hijos de nodo.

Desarrollo del pseudocódigo

Para un mayor entendimiento del pseudocódigo es mejor un ejemplo gráfico, desarrollando cada paso del pseudocódigo y explicándolo (véase anexo 3).

2.2.3.1. Ganancia de información

La ganancia de información es una propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con su clasificación objetivo.(CIAT, 1999)

Ecuación para el cálculo de ganancia de información

$$Ganancia(S,A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{S_v}{S} Entropia(Sv) \quad (1)$$

2.2.3.2.Entropía

Shannon inventó El concepto de entropía, que mide la impureza del conjunto de entrada. En física y matemáticas, la entropía se conoce como aleatoriedad o impureza en el sistema. En teoría de la información, se refiere a la impureza en un grupo de ejemplos. La ganancia de información es una disminución de la entropía.(Varela Arregocés Edwin Campbells Sánchez & Simón Bolívar Barranquilla -Atlántico, 2011)

2.2.3.3. Overfitting

Según Ying (2019), el sobreajuste es un error de modelado en las estadísticas que se produce cuando una función está demasiado alineada con un conjunto limitado de puntos de datos. Como resultado, el modelo es útil en referencia solo a su conjunto de datos inicial y no a ningún otro conjunto de datos.

El sobreajuste del modelo generalmente toma la forma de hacer un modelo demasiado complejo para explicar las idiosincrasias en los datos en estudio. En realidad, los datos que se estudian a menudo contienen algún grado de error o ruido aleatorio.

2.2.3.4. Poda

La poda es una técnica de compresión de datos en algoritmos de búsqueda y aprendizaje automático que reduce el tamaño de los árboles de decisión al eliminar secciones del árbol que no son críticas y son redundantes para clasificar instancias. La poda reduce la complejidad del clasificador final y, por lo tanto, mejora la precisión predictiva al reducir el sobreajuste. (Frank, 2000)

Técnicas

- **Pre-poda**

  Consiste en detener la expansión de un nodo en un momento dado de la construcción del árbol.

- **Post-poda**

  Se genera el árbol completo y luego se buscan sub-ramas a podar con algún criterio.

2.2.4. Árbol de decisión J48

Este algoritmo construye un árbol a partir de datos. Se construye iterativamente al ir agregando nodos o ramas que minimicen la diferencia entre los datos. Este algoritmo es un descendiente del ID3 y se extiende en el sentido de su capacidad de utilizar atributos numéricos y vacíos para generar reglas del árbol. Con el propósito de clasificación de una nueva instancia, J48 prueba cada uno de los valores del atributo de acuerdo con su estructura hasta que encuentra una hoja, la cual contiene los valores de la clase para cada instancia.

Adicionalmente el árbol de decisión J48 es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta WEKA de minería de datos. J48 sigue cada paso del algoritmo C4.5 para la generación de un árbol de decisión.

2.2.5. Bases de datos

En la actualidad, las bases de datos se usan tan ampliamente que se pueden encontrar en organizaciones de todos los tamaños, desde grandes corporaciones y agencias gubernamentales, hasta pequeños negocios e incluso en hogares. Las actividades

diarias con frecuencia lo ponen en contacto con las bases de datos, ya sea directa o indirectamente.

El término de bases de datos fue escuchado por primera vez en 1963, en un simposio celebrado en California, USA. Una base de datos se puede definir como un conjunto de información relacionada que se encuentra agrupada o estructurada.(M. Ricardo, 2009)

Desde el punto de vista informático, la base de datos es un sistema formado por un conjunto de datos almacenados en discos que permiten el acceso directo a ellos y un conjunto de programas que manipulen ese conjunto de datos.

¿Por qué surgieron las bases de datos?

Antes de las bases de datos se utilizaban los archivos para guardar la información, sin embargo, estos presentaban varios problemas (Silberschatz et al., 2002):

- **Redundancia e inconsistencia de los datos**. - Redundancia significa tener el mismo dato guardado varias veces. Inconsistencia significa que hay contradicción en el contenido de un mismo dato, es decir, que un mismo dato tiene un valor en una parte de la memoria, mientras que en otra parte contiene otro valor diferente.
- **Dificultad en el acceso a los datos**. - Era difícil que el usuario encontrara rápidamente un dato en especial.
- **No existía el aislamiento de los datos**. - Debido a que los datos estaban dispersos en varios archivos y podían estar en diferentes formatos, era difícil escribir programas nuevos de aplicación para recuperar los datos apropiados.
- **Problemas de integridad.** - Era complicado asegurarse que los valores almacenados satisficieran ciertos tipos de restricciones, por ejemplo, que tuvieran un valor mínimo y/o un valor máximo.

- **Problemas de atomicidad**. - Era muy difícil asegurar que una vez que haya ocurrido alguna falla en el sistema y se ha detectado, los datos se restauraran al estado de consistencia que existía entes de la falla.
- **Anomalías en el acceso concurrente**. - La cuestión de asegurar la consistencia de los datos se complica todavía más cuando se trata de sistemas en los que hay varios usuarios accediendo a un mismo archivo desde diferentes computadoras.
- **Problemas de seguridad**. - No todos los usuarios de un sistema de información deberían poder acceder a todos los datos. En un sistema de archivos es muy difícil garantizar las restricciones de seguridad.

Lenguaje de bases de datos

Un sistema de bases de datos proporciona un lenguaje de definición de datos para especificar el esquema de la base de datos y un lenguaje de manipulación de datos para expresar las consultas a la base de datos y las modificaciones.

Un esquema de base de datos se especifica mediante un conjunto de definiciones expresadas mediante un lenguaje especial llamado lenguaje de definición de datos (LDD).(Silberschatz et al., 2002)

2.2.6. Normalización de datos

Según (M. Ricardo, 2009). La normalización de datos es el modelo bajo el cual se minimiza una cantidad extensa de información adaptándola a las necesidades requeridas respecto a la información tratada.

El objetivo básico del modelado lógico es desarrollar una “buena” descripción de los datos, sus relaciones y sus restricciones. Para el modelo relacional, esto significa que debe identificar un conjunto adecuado de relaciones. Sin embargo, la tarea de elegir las relaciones es difícil, porque existen muchas opciones para que el diseñador las considere.

El propósito de la normalización es producir un conjunto estable de relaciones que sea un modelo fiel de las operaciones de la empresa. Al seguir los principios de la normalización, se logra un diseño que es muy flexible, lo que permite al modelo extenderse cuando necesite representar nuevos atributos, conjuntos de entidades y relaciones.

La base de datos se diseña en tal forma que se pueden fortalecer con facilidad ciertos tipos de restricciones de integridad. También se puede reducir la redundancia en la base de datos, tanto para ahorrar espacio como para evitar inconsistencias en los datos.

También asegura que el diseño esté libre de ciertas anomalías de actualización, inserciones y borrado. Una anomalía es un estado inconsistente, incompleto o contradictorio de la base de datos. Si estas anomalías estuvieran presentes sería incapaz de representar cierta información, podría perder información cuando ciertas actualizaciones se realicen y correría el riesgo de que los datos se vuelvan inconsistentes con el tiempo.

Reglas de inferencia: axiomas de Armstrong.

Para comenzar con el abordaje más formal de la normalización es necesario un conjunto de axiomas que proporcionen reglas para trabajar con las dependencias funcionales. Las reglas de inferencia para dependencias funcionales, llamados axiomas de inferencia o axiomas de Armstrong, en honor a su desarrollador, se pueden usar para encontrar todas las DF lógicamente implicadas por un conjunto de DF.

Estas reglas son sonoras, lo que significa que son una consecuencia inmediata de la definición de dependencia funcional y que cualquier dependencia funcional que se pueda derivar a partir de un conjunto dado de DF, usándolas, es verdadera. También son completas, lo que significa que se pueden usar para derivar toda inferencia válida acerca

de las dependencias, de modo que, si una DF particular no se puede derivar a partir de un conjunto dado de DF usando estas reglas, entonces el conjunto dado de DF no implica dicha DF particular.(M. Ricardo, 2009)

Sean A, B, C y D subconjuntos de atributos de una relación R. Los siguientes axiomas se sostienen (note que aquí AC significa la unión del conjunto A y el conjunto C):

- **Reflexividad**. Si B es un subconjunto de A, entonces A → B. Esto también implica que A → A siempre se sostiene. Las dependencias funcionales de este tipo se llaman **dependencias funcionales triviales**.
- **Aumento**. Si A → B, entonces AC → BC.
- **Transitividad**. Si A → B y B → C, entonces A → C.

Las siguientes reglas se pueden derivar a partir de las tres anteriores:

- **Aditividad o unión.** Si A → B y A → C, entonces A → BC.
- **Proyectividad o descomposición.** Si A → BC, entonces A → B y A → C.
- **Pseudotransitividad.** Si A → B y CB → D, entonces AC → D.

Estas reglas se pueden usar para desarrollar una teoría formal de las dependencias funcionales, pero en vez de ello el texto se concentrará en sus aplicaciones prácticas.

Proceso de normalización

Como se afirmó al comienzo del capítulo, el objetivo de la normalización es encontrar un conjunto estable de relaciones que sea un modelo fiel de la empresa. Se encontró que la normalización elimina algunos problemas de la representación de datos y resulta en un buen esquema para la base de datos.

- Análisis
- Síntesis
- Normalización desde un diagrama entidad-relación

2.2.7. Normas ISO

Las normas ISO son documentos que especifican requerimientos que pueden ser empleados en organizaciones para garantizar que los productos y/o servicios ofrecidos por dichas organizaciones cumplen con su objetivo. Hasta el momento ISO (International Organization for Standardization), ha publicado alrededor de 19.500 normas internacionales que se pueden obtener desde la página oficial de ISO (http://www.iso.org/).

2.2.8. ISO 25000

ISO/IEC 25000, conocida como SQuaRE (System and Software Quality Requirements and Evaluation), es una familia de normas que tiene por objetivo la creación de un marco de trabajo común para evaluar la calidad del producto software.

La familia ISO/IEC 25000 es el resultado de la evolución de otras normas anteriores, especialmente de las normas ISO/IEC 9126, que describe las particularidades de un modelo de calidad del producto software, e ISO/IEC 14598, que abordaba el proceso de evaluación de productos software. Esta familia de normas ISO/IEC 25000 se encuentra compuesta por cinco divisiones.

2.2.9. Algoritmo

Un Algoritmo es una serie ordenada de instrucciones, pasos o procesos que llevan a la solución de un determinado problema. Los algoritmos pueden ser tan simples como sumar dos números o tan complejos como graficar figuras en tercera dimensión.

Los Algoritmos permiten describir claramente una serie de instrucciones que debe realizar el computador para lograr un resultado previsible. Vale la pena recordar que un procedimiento de computador consiste de una serie de instrucciones muy precisas y escritas en un lenguaje de programación que el computador entienda. (Carlos et al., 2009)

Luego de analizar detalladamente el problema hasta entenderlo completamente, se procede a diseñar un algoritmo (trazar un plan) que lo resuelva por medio de pasos sucesivos y organizados en secuencia lógica. Procesos, rutinas o biorritmos naturales como la gestación, las estaciones, la circulación sanguínea, los ciclos cósmicos, etc. Son algoritmos naturales que generalmente pasan desapercibidos.(Carlos et al., 2009)

Ejemplo

Consideremos el algoritmo de Euclides para hallar el Máximo Común Divisor (MCD) de dos números enteros positivos dados. Obsérvese que no se especifica cuáles son los dos números, pero si se establece claramente una restricción: deben ser enteros y positivos.

ALGORITMO EN SEUDOCÓDIGO

- Paso 1: Inicio.
- Paso 2: Leer los dos números (“a” y “b”). Avanzar al paso 3.
- Paso 3: Comparar “a” y “b” para determinar cuál es mayor. Avanzar al paso 4.
- Paso 4: Si “a” y “b” son iguales, entonces ambos son el resultado esperado y termina el algoritmo. En caso contrario, avanzar al paso 5.
- Paso 5: Si “a” es menor que “b”, se deben intercambiar sus valores. Avanzar al paso 6; si “a” no es menor que “b”, avanzar al paso 6.
- Paso 6: realizar la operación “a” menos “b”, asignar el valor de “b” a “a” y asignar el valor de la resta a “b”. Ir al paso 3.

2.2.10. Pseudocódigo

El pseudocódigo es una forma de expresar los distintos pasos que va a realizar un programa, de la forma más parecida a un lenguaje de programación. Su principal función es la de representar por pasos la solución a un problema o algoritmo, de la forma más detallada posible, utilizando un lenguaje cercano al de programación. El pseudocódigo no puede ejecutarse en un ordenador ya que entonces dejaría de ser pseudocódigo, como su propio nombre indica, se trata de un código falso (pseudo = falso), es un código escrito para que lo entienda el ser humano y no la máquina.(Duque et al., 2017)

FORMATO GENERAL DEL PSEUDOCÓDIGO

Todo programa escrito en pseudocódigo está conformado por los siguientes tres bloques:

- La Cabecera donde se coloca el nombre del programa escrito en pseudocódigo:

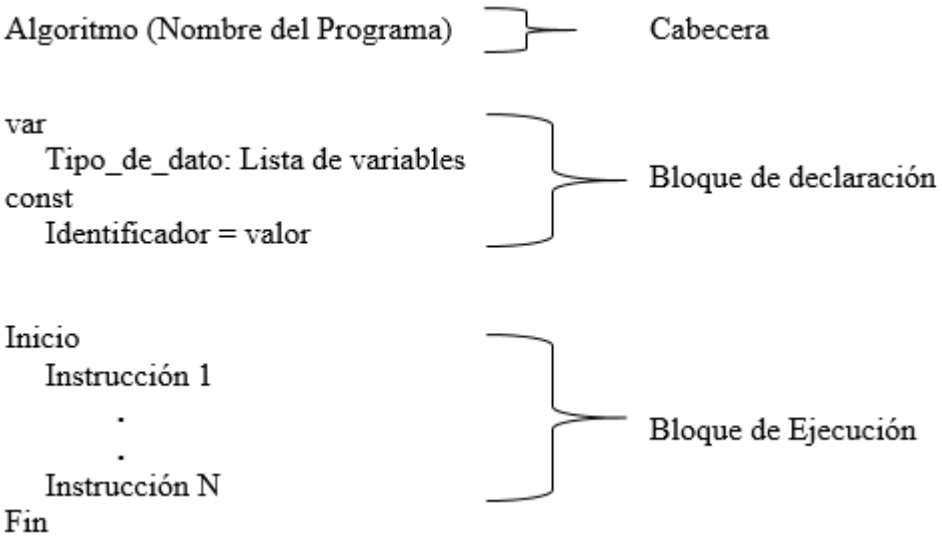
  Algoritmo_nombre_del_programa

- El Bloque de Declaración es donde se informa al computador cuántas celdas y de qué tipo se van a necesitar, aquí se deben declarar todos los identificadores que se utilizarán en el transcurso de la ejecución del programa.

  El Bloque de Declaración puede contener dos secciones: Declaración de Variables y Declaración de Constantes.

- El Bloque de Ejecución es el cuerpo de instrucciones que debe ejecutar el computador. Comienza y termina con las palabras reservadas Inicio y Fin, y en su interior podrá contener la repetición de las otras instrucciones las veces que sea necesario.

Figura 2: Formato General Pseudocódigo



Elaboración propia

2.2.11. Data set

Según Akujuobi & Zhang (2017) una Data Set es un conjunto de datos (o conjunto de datos) es una colección de datos. En el caso de los datos tabulares, un conjunto de datos corresponde a una o más tablas de la base de datos, donde cada columna de una tabla representa una variable particular y cada fila corresponde a un registro dado del conjunto de datos en cuestión.

El conjunto de datos enumera los valores de cada una de las variables, como la altura y el peso de un objeto, para cada miembro del conjunto de datos. Cada valor se conoce como dato. Los conjuntos de datos también pueden consistir en una colección de documentos o archivos.

La inteligencia artificial se entrena utilizando datos preparados que hemos recopilado. Este conjunto de datos debe adaptarse a cada IA concreta modelo. Por lo tanto, debemos preparar ese conjunto de datos para adaptarlo. Además, este conjunto suele debe dividirse en tres subconjuntos: conjunto de datos de entrenamiento, conjunto de datos de

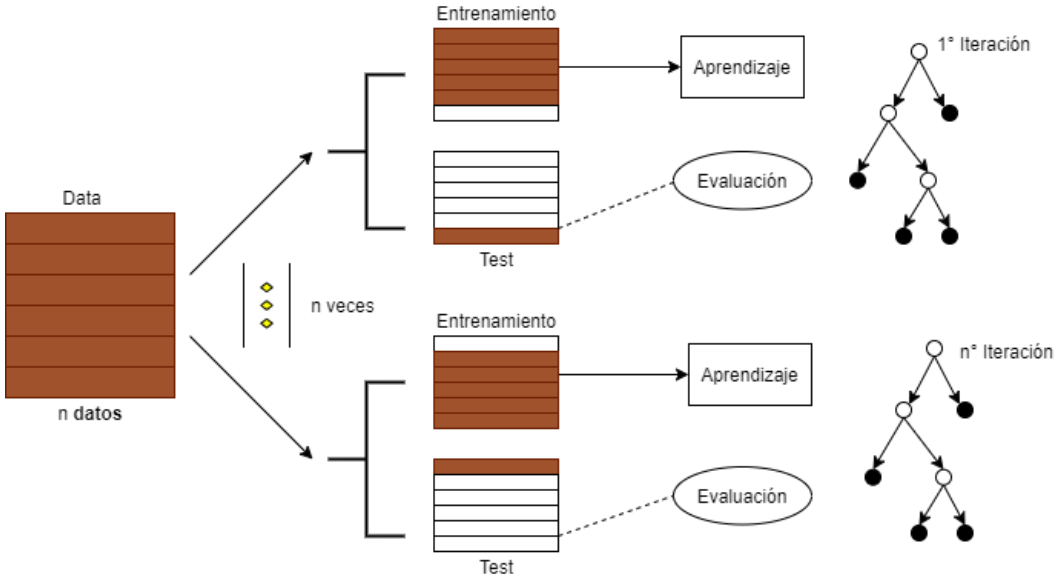
validación y conjunto de datos de prueba con los siguientes porcentajes del conjunto total 60%, 20%, 20%.

- **Datos de entrenamiento**: Los datos de entrenamiento son los datos que pertenecen a un conjunto de entrenamiento que se usa para aprender y se ajusta a los parámetros del clasificador.
- **Datos de validación**: Los datos de validación son los datos que pertenecen al conjunto de datos de validación que se utiliza para ajustar los parámetros de un clasificador. Estos parámetros se denominan híper-parámetros y se ajustan durante las sucesivas repeticiones de entrenamiento.
- **Datos test:** Los datos de prueba son los datos que pertenecen al conjunto de prueba. Es el último subconjunto utilizado en el proceso y su principal objetivo es evaluar el desempeño de un modelo entrenado.(Savira & Suharsono, 2013)

K-Fold cross validation

Cuando la cantidad de datos es relativamente pequeña, surge un dilema, que es cómo debemos dividir el conjunto de datos para poder entrenar el modelo sin usar los mismos datos en el entrenamiento y en validación ya que esto sería un error. Por otro lado, si los dividimos en el camino propuesto anteriormente, habrá una cantidad muy pequeña de datos en uno de los conjuntos, que puede hacer que el modelo no esté lo suficientemente entrenado o que la validación no sea confiable.

Figura 3: K-fold validación cruzada



Elaboración propia

2.2.12. Métodos de análisis de datos

2.2.12.1. Chi-cuadrada

Es una prueba estadística para evaluar hipótesis acerca de la relación entre dos variables categóricas.

- Se simboliza: x^2.
- Hipótesis por probar: correlacionales.
- Variables involucradas: dos. La prueba Chi cuadrada no considera relaciones causales.
- Nivel de medición de las variables: nominal u ordinal (o intervalos o razón reducidos a ordinales).

Procedimiento: se calcula por medio de una tabla de contingencia o tabulación cruzada, que es un cuadro de dos dimensiones y cada dimensión contiene una variable.(Carlos Fernández Collado, 2014)

2.2.12.2. Métodos de correlación

Los métodos de correlación de Pearson y Spearman son técnicas bivariadas que se emplean en situaciones donde el investigador quiere observar representaciones de la información, que permitan establecer similaridades o disimilaridades entre las variables e individuos, para hacer evidente la variabilidad conjunta y por tanto tipificar lo que sucede con los datos. Ejemplos clásicos de correlación podrían ser la relación entre peso y talla, la relación entre horas dedicadas al deporte y percepción de calidad de vida, la relación entre la cantidad suministrada de un fármaco y su correlación con los valores de signos vitales, entre otras. (Mondragon, 2014)

Tanto el coeficiente de correlación de Pearson como el de Spearman siguen las mismas normas de interpretación

- Solamente toma en cuenta valores entre 1 y -1.
- El 0 indica que no existe correlación.
- El valor numérico indica la magnitud de la correlación.
- El coeficiente de correlación cuantifica la correlación entre dos variables, cuando está realmente existe.
- El hecho de que exista correlación entre las variables no implica que exista causalidad o dependencia entre ellas.
- El signo indica la dirección de la correlación.
- Los valores cercanos a 1 nos indican una correlación muy buena y los cercanos a cero una correlación mínima o nula.

2.3. GLOSARIO DE TÉRMINOS BÁSICOS

Accuracy

El accuracy(**Exactitud**) se refiere a que tan cerca está el resultado de una medición del valor verdadero. En términos simples, el accuracy es el porcentaje de acierto con el clasifica los datos de entrada una herramienta de I.A. En forma práctica la Exactitud es la cantidad de predicciones positivas que fueron correctas.

Validación Cruzada

La validación cruzada o cross-validation es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones.(Pérez-Planells et al., 2015)

Matriz de Confusión

En el campo de la inteligencia artificial y en especial en el problema de la clasificación estadística, una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.(Ariza-lópez et al., 2018)

Anaconda

Anaconda Individual Edition es un administrador de paquetes, administrador de entorno y distribución de Python gratuito y fácil de instalar con una colección de más de 1,500 paquetes de código abierto con soporte gratuito de la comunidad. Anaconda es independiente de la plataforma, por lo que puede usarla ya sea que esté en Windows, macOS o Linux.

PYTHON

Python es la navaja suiza de los programadores. Se trata de un veterano lenguaje de programación presente en multitud de aplicaciones y sistemas operativos. Podemos encontrarlo corriendo en servidores, en aplicaciones iOS, Android, Linux, Windows o Mac. Esto es debido a que cuenta con una curva de aprendizaje moderada ya que su filosofía hace hincapié en ofrecer una sintaxis de código legible.

IDE

Un entorno de desarrollo integrado, también conocido por sus siglas IDE, puede considerarse como un entorno digital utilizado para desarrollar software, juegos o cualquier cosa relacionada con la codificación. Un IDE ofrece integración desde los pasos más básicos del desarrollo de software, como escribir su código, depurar o incluso compilar sus aplicaciones en un lenguaje que las computadoras puedan entender.

Tasa de Error

La tasa de error se refiere a que tan lejos está el resultado de una medición del valor verdadero. En términos simples, la tasa de error es el porcentaje de error con el que clasifica los datos de entrada una herramienta de I.A.

WEKA

Es un software de código abierto proporciona herramientas para el preprocesamiento de data, la implementación de varios algoritmos de aprendizaje automático y herramientas de visualización para que las desarrolle técnicas de aprendizaje automático y aplicarlas a problemas de minería de data del mundo real.

JAVA

Java es un lenguaje de programación y una plataforma informática comercializada por primera vez en 1995 por Sun Microsystems. Hay muchas aplicaciones y sitios web que no funcionarán a menos que tenga Java instalado y cada día se crean más. Java es rápido, seguro y fiable. Desde portátiles hasta centros de datos, desde consolas para juegos hasta súper computadoras, desde teléfonos móviles hasta Internet, Java está en todas partes.

Desviación Estándar

La desviación estándar es la medida de dispersión más común, que indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos. El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido. La desviación estándar se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso.

SPSS

SPSS es un formato que ofrece IBM para un análisis completo. Es el acrónimo de Producto de Estadística y Solución de Servicio. Existen otros productos diferentes en la suite, cada uno de ellos ofrecen sus propias características únicas.

SPSS es un software popular entre los usuarios de Windows, es utilizado para realizar la captura y análisis de datos para crear tablas y gráficas con data compleja. El SPSS es conocido por su capacidad de gestionar grandes volúmenes de datos y es capaz de llevar a cabo análisis de texto entre otros formatos más.

CATPCA

Este procedimiento cuantifica simultáneamente las variables categóricas a la vez que reduce la dimensionalidad de los datos. El análisis de componentes principales categórico se conoce también por el acrónimo CATPCA, del inglés CATegorical Principal Components Analysis.

2.4. HIPÓTESIS DE LA INVESTIGACIÓN

2.4.1. Hipótesis General

El árbol de decisión J48 es una herramienta fiable en la clasificación de documentos de la UGEL El Collao-Ilave.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. MÉTODOS

3.1.1. Tipo de investigación

El presente trabajo se caracteriza por ser una investigación del tipo cuantitativa. "La investigación cuantitativa nos ofrece la posibilidad de generalizar los resultados más ampliamente, nos otorga control sobre los fenómenos, así como un punto de vista de conteo y las magnitudes de éstos. Asimismo, nos brinda una gran posibilidad de réplica y un enfoque sobre puntos específicos de tales fenómenos, además de que facilita la comparación entre estudios similares.". (Hernandez, Roberto; Fernández, Collado;Baptista, 2010)

3.1.2. Diseño de investigación

Según Hernandez, Roberto; Fernández, Collado; Baptista, (2010) una investigación cuantitativa nos ofrece la posibilidad de generalizar los resultados más ampliamente.

Diseño con pre test y post test para poder realizar las comparaciones entre los valores obtenidos y medir el efecto que provoca, el esquema es el siguiente:

$$O1-X-O2$$

Donde:

$O1$= Medición antes (Pre test).

X = Implementación del árbol de decisión J48.

$O2$= Medición después (Post test).

Este diseño permite la comparación de resultados de clasificación de documentos de la UGEL El Collao-Ilave; sin el uso del árbol J48 y usando el árbol J48. Esta comparativa permite establecer el nivel de fiabilidad de la clasificación de documentos usando el árbol J48 y de esa forma confirmar o desmentir la hipótesis inicial que establece que el árbol de decisión J48 sería una herramienta fiable para la clasificación de los documentos de la UGEL El Collao-Ilave.

3.2. POBLACIÓN Y MUESTRA

3.2.1. Población

La población es definida por Hernández (2014) como un conjunto de todos los casos que concuerda con una serie de especificaciones. De la cual, definida la unidad de análisis, se procede a delimitar la población de estudio sobre la cual se pretende generalizar los resultados y que cumpla con los criterios de selección.

La población está constituida por los documentos que ingresaron al área de Gestión Administrativa de la UGEL El Collao-Ilave para su clasificación durante el periodo de duración designado para la recolección de datos (2 meses). Los cuales conforman la cantidad de 1013 documentos.

3.2.2. Muestra

Según Hernández (2014), la muestra es un subgrupo de elementos que pertenecen a ese grupo definido en sus características. La muestra fue conformada mediante la técnica de muestreo no probabilístico intencional: el cual permite seleccionar casos característicos de una población limitando la muestra sólo a estos, seleccionando aquellos que más convengan al equipo investigador para conducir la investigación.(Otzen & Manterola, 2017)

La muestra está constituida por los documentos que ingresaron al área de Gestión Administrativa de la UGEL El Collao-Ilave y fueron clasificados para oficinas dentro del área de Gestión Administrativa durante el periodo de duración designado para la recolección de datos (2 meses). Los cuales conforman la cantidad de 334 documentos.

3.3. UBICACIÓN DE LA POBLACIÓN

El presente trabajo de investigación se realizó en el área de Gestión Administrativa el cual pertenece a la UGEL El Collao-Ilave, la cual está ubicada en la ciudad de Ilave capital de la provincia de El Collao la cual pertenece al departamento de Puno.

3.4. INSTRUMENTOS DE RECOLECCIÓN DE DATOS

3.4.1. Instrumentos

Data Set

Para conseguir un conjunto de datos con los cuales trabajar se elaboró una Data Set propia a partir de los documentos que ingresaron al área de Gestión Administrativa de la UGEL El Collao-Ilave durante el periodo de 2 meses. Dicho Data Set era un registro que contenía las palabras que se iban a usar para la clasificación con el árbol de decisión J48. Dicho data set adicionalmente fue etiquetado con los resultados esperados para cada documento y de esa manera realizar el proceso tanto de clasificación como verificación de fiabilidad del árbol.

Listas

Se utilizó como instrumento de medición de resultados un conjunto de listas, los cuales constan de una lista de documentos y las oficinas a las cuales fueron clasificadas usando el árbol de decisión J48.

Las listas no son arreglos (arrays), aunque ambos representan secuencias de elementos de un tipo, los arreglos tienen longitud fija; las listas, no; es decir, las listas son flexibles y permiten cambio de implementación.

3.4.2. Validación y confiabilidad del instrumento

Para la validación del instrumento se hizo uso de métodos estadísticos, tales como Chi-Cuadrado, CATPCA y Coeficientes de correlación los cuales se obtuvieron a través del software SPSS para analizar y depurar los datos. Todo ello para lograr una buena confiabilidad para una evaluación de los datos obtenidos (Data set obtenida).

3.5. PROCEDIMIENTO DEL EXPERIMENTO

Para el procedimiento del experimento se tomó en cuenta los siguientes pasos:

- Definir los instrumentos para la obtención de datos.
- Definir qué parte del documento es relevante para usarlo en el árbol de decisión.
- Realizar el tratamiento de los datos de entrada.
- Configurar el árbol de decisión J48.
- Realizar la clasificación usando el árbol de decisión J48.
- Análisis de los resultados obtenidos.

3.6. PLAN DE PROCESAMIENTO Y ANÁLISIS DE DATOS

Para obtener los datos requeridos para el estudio, se procedió a examinar los documentos y extraer la información. El procedimiento se realizó mediante el uso de una tabla de información, algunos de los resultados están presentados en tablas a fin de realizar su análisis e interpretación. Para el tratamiento estadístico se utilizó el método descriptivo, expresada en porcentajes. Adicionalmente, se hizo uso de técnicas propias de los árboles de decisión para la interpretación de los datos.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. DATA SET RESULTANTE DEL TRATAMIENTO DE DOCUMENTOS

Según Akujuobi & Zhang (2017), una Data Set es un conjunto de datos, llamado también una colección de datos. Y en el caso de la investigación que se llevó a cabo fue desarrollado y conformado a partir del conjunto de documentos que conformo la muestra a evaluar para esta investigación. Para la elaboración de la data set se siguieron una serie de pasos y/o procesos con los cuales se convirtió la información de los documentos en elementos de la data set. Antes de la conformación de la data set final fue necesario aplicar métodos estadísticos de análisis, para validar la data set obtenida. A continuación, se explica detalladamente el proceso de obtención de la data set.

4.1.1. Datos obtenidos

Después de inspeccionar los documentos que llegaban al área de gestión administrativa durante el periodo de 2 meses, se procedió a desarrollar un pequeño data set que simplificaba los datos obtenidos, cabe destacar que el armado de la data set fue hecho de manera continua, ya que los documentos presentados son confidenciales.

Total, documentos inspeccionados: 1013 documentos. Lo cual conformo la población de estudio, y tras aplicar muestreo bajo la forma probabilístico aleatorio simple por conveniencia se conformó la muestra de 334 documentos.

Los datos de salida o posibles respuestas son las oficinas a las cuales pueden ser enviados los documentos, dichas salidas se obtienen de evaluar el organigrama del área de Gestión Administrativa de la UGEL El Collao-Ilave (véase anexo 2).

Los datos obtenidos de los documentos fueron los siguientes:

- Código de documento
- Asunto del documento
- Información suplementaria
- Nombre del solicitante
- Fecha de presentación

Esta información está presente en todos los documentos, y son fáciles de identificar. (véase anexo 7). Una vez obtenidos los datos relevantes, se procedió a armar una data set provisional usando una tabla de información. Se obtuvo la oficina receptora haciendo un seguimiento de los documentos, en la siguiente tabla se presenta un ejemplo y la razón por la cual fueron elegidos los datos.

Tabla 1: Formato de resumen de datos obtenidos de los documentos

	Código de documento	Asunto del documento	Información suplementaria	Nombre del solicitante	Fecha de presentación	Oficina receptora
Descripción de los datos obtenidos	Cuando ingresa un documento a la UGEL, este se registra con un código para su seguimiento.	Es la razón del documento presentado.	Son ideas o razones complementarias que explican el asunto del documento.	Es el identificador por el cual se reconoce quien presentó el documento.	Es el identificador por el cual se reconoce en qué fecha se presentó el documento.	Es la oficina que atendió o dio respuesta al documento presentado.

Elaboración propia

4.1.2. Normalización de los datos obtenidos

Los datos obtenidos de los documentos son irregulares y no contienen ningún patrón o secuencia por la cual puedan ser identificados o usados por un árbol de decisión, es por ello que se necesita normalizarlos, para realizar el proceso de normalización o tratamiento de datos se hizo uso de dos técnicas.

a. **Reducción de datos:** Eliminar información irrelevante.
b. **Separación de datos:** Separar la información original en Datos de Entrada y Datos de salida. Cada una de las partes tiene una forma diferente de normalización para su futura integración en la data set final.

Reducción de datos

Algunos de los datos obtenidos de los documentos tales como Nombre del solicitante, fecha de presentación no formaron parte de la data set final. Ya que dichos datos no aportan información sobre la intención o asunto del documento.

Datos eliminados de la data set.

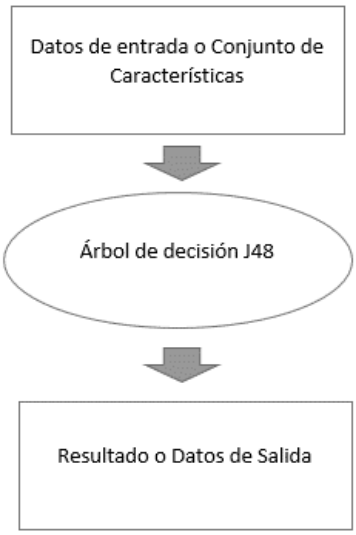
- Nombre del solicitante. La persona que presento el documento no era información relevante a tomar en cuenta para la conformación de la data set final ya que dicha información no puede ser convertida en un dato de entrada para el árbol de decisión J48. Ya que existen cientos de personas que presentan documentos y muchas veces esas personas presentan documentos sin usar su cargo público.
- Fecha de presentación del documento. La fecha de ingreso de un documento al área de Gestión Administrativa de la UGEL El Collao-Ilave no fue tomada en cuenta para la conformación de la data set final, pues la fecha solo simboliza una prioridad en la atención más no una entrada que pueda ayudar al árbol J48 a decidir su correcta clasificación.

Separación de datos

El árbol de decisión J48 recibe como parámetros datos de entrada y datos de salida. Los cuales son extraídos una data set. Es por ello que se procedió a dividir los datos en entrada y salida.

- Los datos de entrada de un árbol de decisión son un conjunto de características ya definidas los cuales el árbol evaluará y decidirá como clasificarlos.
- Los datos de salida son los resultados del proceso de clasificación usando el árbol de decisión J48.

Figura 4: Diagrama simplificado de funcionamiento del árbol de decisión J48



Elaboración propia

Datos de Entrada

Los datos de entrada para un árbol de decisión son el conjunto de características o atributos de un documento que harán posible su evaluación y posterior clasificación. Dicho de otra forma, son las materias primas que usará o en el caso de un árbol de decisión evaluará para generar un resultado. Para la conformación de los datos de entrada se tomó en cuenta que dichos datos deben ser útiles para el proceso de clasificación.

Datos del documento usados para conformar los datos de entrada:

- Asunto del documento e Información Suplementaria: Estos datos obtenidos anteriormente de los documentos son los que pueden ser tratados como datos de entrada ya que en base a estos son clasificados por las secretarias y el personal de mesa de partes de la UGEL El Collao-Ilave.

Procesamiento de los datos obtenidos para convertirlos en datos de entrada para el árbol de decisión J48.

Para procesar los datos de entrada se llevó a cabo un proceso de normalización que consistió en:

i. Reemplazo/Eliminación de información irrelevante
ii. Identificación de palabras clave
iii. Reducción de palabras clave a características.

Para ejemplificar el proceso por el cual se normalizo los datos obtenidos de los documentos para convertirlos en datos de entrada para el árbol de decisión J48, se procedió a elaborar un ejemplo simple basado en 1 documento que forma parte de la data set obtenida. Como anteriormente ya se eliminó nombre del solicitante y fecha de presentación del documento, para este ejemplo solo se extrajo el Asunto del Documento y la Información Suplementaria.

Caso 1:

- Asunto del documento: Solicito materiales faltantes para el inicio de clases de la I.E.P. 70315.
- Información Suplementaria: El personal a cargo no hizo el informe de conformidad cuando se recibieron los materiales.

Los datos que se muestran en forma de oraciones corresponde a la información relevante obtenida del documento, los cuales eran “Asunto del documento” e “Información Suplementaria”. Dicha información fue tratada de manera manual mientras se procesaba la información de cada documento que ingresó al área de Gestión Administrativa de la UGEL El Collao-Ilave.

PROCESOS

Reemplazo/Eliminación de información irrelevante

Tabla 2: Reemplazo y/o eliminación de información irrelevante de la data set

Identificador	Oración Original	Oración tratada
Caso 1	Solicito materiales faltantes para el inicio de clases de la I.E.P. 70315. El personal a cargo no hizo el informe de conformidad cuando se recibieron los materiales.	Solicito materiales faltantes para el inicio de clases de la **escuela**. El personal a cargo no hizo el informe de conformidad cuando se recibieron los materiales.

Elaboración propia

Explicación breve del cuadro, en la primera columna tenemos al identificador del documento, para nuestro ejemplo es **CASO 1**, la segunda columna contiene el Asunto del documento y la información suplementaria, es decir las entradas de este proceso, finalmente la tercera columna es la salida del proceso de reemplazo/eliminación de información irrelevante.

Detalles del proceso

Este proceso de eliminación y/o reemplazo se realiza con el fin de simplificar los datos de entrada y formar una data set que pueda ser leído por el árbol de decisión.

Para desarrollar este proceso de manera automática se hizo un pequeño algoritmo en Python usando el IDE Anaconda, el cual se centra en hacer una identificación de

conjuntos de palabras usadas de manera repetida y las cuales pueden ser reemplazadas por una sola palabra o ser eliminadas sin afectar el proceso de normalización de los datos.

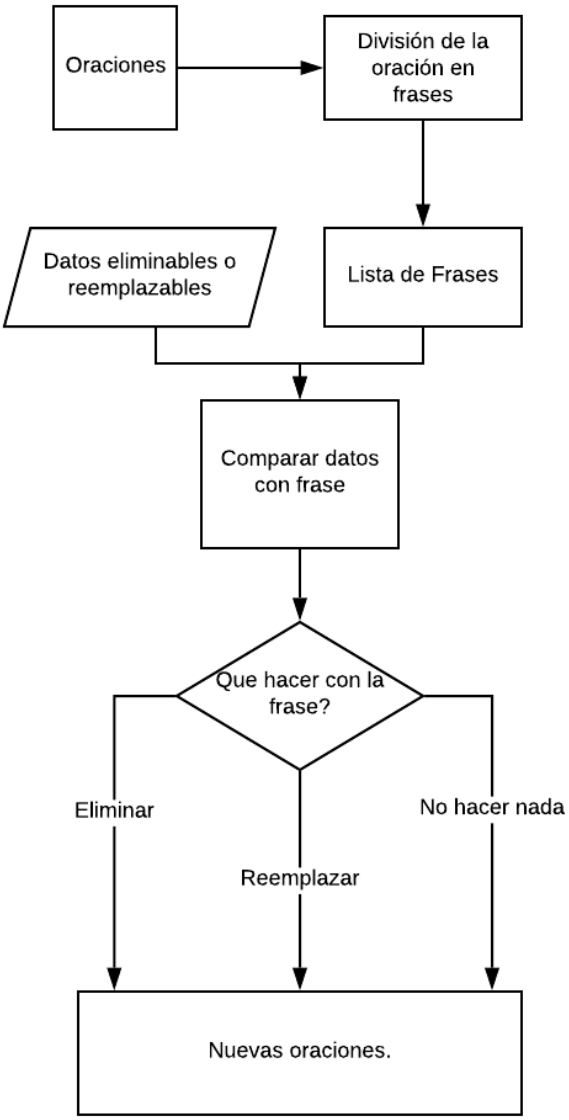
Los pasos que realizó el algoritmo para realizar el proceso fueron los siguientes:

i. División de la oración en frases: La entrada para este paso fueron el Asunto del Documento y la Información Suplementaria, de los cuales se obtuvo una lista de frases u oraciones acortadas los cuales se usaron para el siguiente paso del algoritmo.

ii. Comparar datos con frase: Este proceso se dio con el fin de evaluar si existían dentro de las oraciones datos que podrían ser eliminados y/o reemplazados de manera segura.

iii. Una vez comparados los datos con la frase se obtenía una respuesta del algoritmo, las tres posibles respuestas eran: Eliminar la frase, Reemplazar la frase o no hacer nada con la frase.

Figura 5: Diagrama proceso de eliminación/reemplazo de información irrelevante



Elaboración propia

Se muestra en la figura un diagrama que muestra el ciclo de vida del algoritmo para Eliminar/Reemplazar información innecesaria de la entrada de este proceso y generar una salida normalizada.

Identificación de palabras clave

Aquí se establece la forma bajo la cual se convirtió la entrada tratada anteriormente a palabras clave que serán analizadas más adelante.

Tabla 3: Proceso de identificación de palabras clave de la data set

Identificador	Oración Original	Lista palabras clave
Caso 1	Solicito materiales faltantes para el inicio de clases de la **escuela**. El personal a cargo no hizo el informe de conformidad cuando se recibieron los materiales.	a. Solicito b. Materiales c. Faltantes d. Escuela e. Personal f. Conformidad g. Informe h. Recepción

Elaboración propia

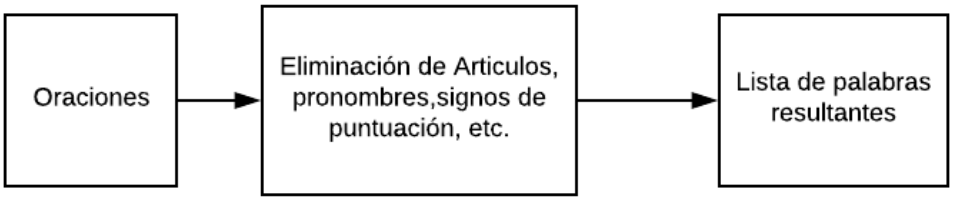
Explicación breve del cuadro, en la primera columna tenemos al identificador del documento, para nuestro ejemplo es **CASO 1**, la segunda columna contiene el resultado del proceso de Reemplazo/Eliminación de información irrelevante, finalmente la tercera columna es la salida del proceso de identificación de las palabras clave.

Detalles del proceso

La identificación de las palabras clave son el resultado de un proceso de eliminación de información innecesaria de la oración tales como artículos, pronombres, conectores y entre otros. Los cuales no son información relevante para el conjunto de datos de entrada del árbol de decisión J48. Este proceso se realizó con un pequeño algoritmo hecho en Python usando el IDE Anaconda.

El trabajo o funcionamiento del algoritmo para este proceso fue simple. Eliminar toda palabra o signo que fuera innecesaria. Para de esa forma obtener solo aquellas palabras clave que se necesitaban para formar los datos de entrada.

Figura 6: Diagrama simplificado, proceso de identificación de palabras clave



Elaboración propia

Se muestra en la figura un diagrama que muestra el ciclo de vida del algoritmo de Identificación de las palabras clave, dichas palabras clave se conformarían en una lista la cual representa la salida de este proceso.

Reducción de palabras clave a características.

Este proceso se da con el fin de evaluar cada palabra clave e inferir si va a continuar como un dato de entrada o ser eliminado. Para ello cada palabra clave va a ser comparada con un conjunto de características que forman los datos de entrada del árbol de decisión J48. Primero debemos conocer que características o atributos conformaran los datos de entrada. Es por ello que se elaboró un listado de la información más relevante y que puede ser usado por el árbol de decisión J48 para generar una clasificación.

Lista de Características

- Tipo Monetario
- Relación Afinidad con la persona
- Tratamiento del documento
- Tipo de entidad involucrada
- Tipo proceso
- Estado

Las características elegidas representan los atributos en los cuales puede ser divididos la información de un documento, para llegar a esa conclusión se hizo el análisis de más de 300 documentos y formando así un listado de que atributos o características que son relevantes para el proceso de clasificación usando el árbol de decisión. Los valores que puede tomar cada característica fueron establecidos en base a la agrupación de los conceptos relacionados a cada característica y la ayuda de expertos de la UGEL.

Descripciones de las características establecidas

Tipo Monetaria

Esta característica como su nombre lo indica, explora el aspecto económico ligado al documento.

Tabla 4: Característica Tipo Monetaria

Posibles Valores	Descripción
Saldo	Cuando la razón del documento involucra el desenvolvimiento de dinero por parte de la UGEL
Materiales	Cuando la razón del documento se relaciona con materiales de la UGEL o administrados por la misma
No monetario	Cuando no forma parte de ninguno de los dos posibles valores anteriores.

Elaboración propia

Relación de afinidad con la persona

Esta característica establece qué relación tiene la persona que presenta el documento con la persona que es involucrada en el documento.

Ejemplificando lo anterior, si un tutor o apoderado presenta un documento de Visación este afecta al estudiante al cual pertenece dicho documento (Certificado de Estudios).

Tabla 5: Característica Relación de afinidad con la persona

Posibles Valores	Descripción
Titular	La persona a la que afecta o involucra el documento es la misma que la presentó.
Familiar	Muchas veces los familiares, ya sean los padres o hermanos, presentan un documento en nombre de su familiar para realizar algún trámite.
No especificado	Cuando no forma parte de ninguno de los dos posibles valores anteriores.

Elaboración propia

Tratamiento del documento

Esta característica vislumbra cual será el trato que recibirá el documento una vez sea enviado a la oficina correcta.

Tabla 6: Característica Tratamiento del documento

Posibles Valores	Descripción
Contra-respuesta	Cuando el documento presentado requiere una contestación en forma de otro documento oficial.
Respuesta	Cuando el documento presentado no requiere una contestación escrita o en forma de documento.
Lectura	Cuando la función del documento es únicamente informar o poner en conocimiento algo.
No especifica	Cuando el documento no forma parte de ninguna de las tres posibles respuestas.

Elaboración propia

Tipo Entidad Involucrada

Esta característica sirve para establecer, que entidad está involucrada con el documento ya que dentro de la UGEL existen áreas designadas para atender a cada ente que presente algún documento, desde personas hasta instituciones.

Tabla 7: Característica Tipo Entidad Involucrada

Posibles Valores	Descripción
Escuela	Cuando la institución educativa involucrada con el documento es una Escuela.
Colegio	Cuando la institución educativa involucrada con el documento es un Colegio.
CETPRO	Cuando la institución educativa involucrada con el documento es un CETPRO.
No especifica	Cuando el documento no forma parte de ninguna de las tres posibles respuestas.

Elaboración propia

Tipo Proceso

Con esta característica se establece que tipo de proceso espera realizar la persona que presenta el documento, para esta característica en especial se necesita entender que algunos procesos surgen o siempre existen o existirán.

Tabla 8: Característica Tipo Proceso

Posibles Valores	Descripción
Contratación	Proceso por el cual las instituciones educativas pertenecientes a la jurisdicción de la UGEL contratan docentes. Dicho proceso se da por medio de concurso.
Adjudicación	Proceso por el cual el docente ganador de un concurso público toma posesión de la plaza ganada.
Visación	Proceso por el cual un certificado o documento es validado por la UGEL El Collao-Ilave.
No especifica	Cuando el documento no forma parte de ninguna de las tres posibles respuestas.

Elaboración propia

Estado

Con esta característica se establece el tiempo del evento (pasado, presente y futuro) al cual el documento hace referencia, pues muchos documentos hacen referencia a un evento pasado, otros a un evento futuro. Esto resulta importante, pues muchas de las

oficinas dentro de la UGEL toman control de algunos eventos durante su realización y una vez este finaliza otra oficina toma el control del mismo evento.

Tabla 9: Característica Estado

Posibles Valores	Descripción
Normal	Cuando el documento no está sujeto a ninguna convocatoria o proceso.
Próxima	Cuando el documento menciona un proceso que aún no ha comenzado.
En Curso	Cuando el documento menciona un proceso que se está realizando en la actualidad.
Finalizado	Cuando el documento menciona un proceso que ha finalizado.

Elaboración propia

Detalles del proceso

La conformación de los datos de entrada son el resultado de un proceso de comparación y adición, para la realización de tal fin es necesario primero contar con la lista de las características o atributos y sus respectivos posibles valores, una vez obtenido ello se procede a usar otro algoritmo programado en Python usando el IDE Anaconda para convertir las palabras clave a características. Para definir que entraba en los datos de entrada, se eligió el valor de característica que mejor se ajustaba teniendo dentro de esos posibles valores el valor "No específica".

Figura 7: Esquema de datos de entrada para la data set

Valor Tipo Monetario	Valor Relación Afinidad con la persona	Valor Tratamiento del documento	Valor Tipo Entidad Involucrada	Valor Tipo Proceso	Valor Estado

Elaboración propia

Los pasos que realizo el algoritmo para este proceso fueron los siguientes:

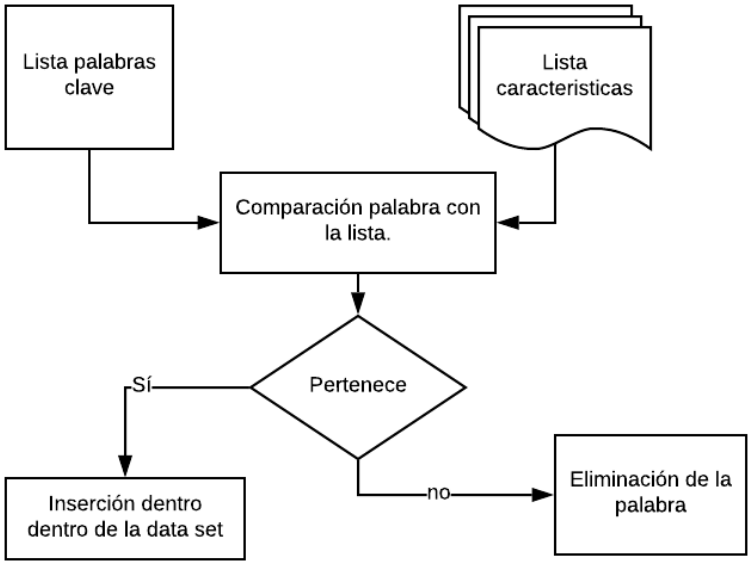
I. Comparación de cada palabra clave con cada característica. Ya que las palabras clave no siempre van a ser iguales a las características o sus posibles valores, fue

necesario el uso de una función de sinónimos dentro del proceso de comparación, así como también el uso de un identificador de contexto el cual fue una función creada para este trabajo de investigación.

II. Si la palabra clave luego de ser comparada con todas las características no tomaba ningún valor, esta era eliminada o no tomada en cuenta para su inserción dentro de la data set como un atributo de un dato de entrada.

III. Si la palabra clave luego de ser comparada con las características era afín a alguna de ellas, se la colocaba en la insertaba dentro de la data set como atributo de un dato de entrada.

Figura 8: Proceso de creación de los datos de entrada de la data set



Elaboración propia

En la anterior figura se muestra el diagrama del funcionamiento del proceso de obtención de los datos de entrada y al mismo tiempo su inclusión dentro de la data set a partir de las palabras clave obtenidas en el proceso de **Identificación de palabras clave.**

Tabla 10: Resultados de obtención de los datos de entrada de la data set

Identificador	Lista palabras clave	Datos de Entrada resultante	
		Característica	Valor
Caso 1	a. Solicito	Tipo Monetario	Materiales
	b. Materiales	Relación Afinidad con la persona	Titular
	c. Faltantes	Tratamiento del documento	Respuesta
	d. Escuela	Tipo de entidad involucrada	Escuela
	e. Personal	Tipo proceso	No especifica
	f. Conformidad	Estado	Finalizado
	g. Informe		
	h. Recepción		

Elaboración propia

Explicación breve del cuadro, en la primera columna tenemos al identificador del documento, para nuestro ejemplo es **CASO 1**, la segunda columna contiene el resultado del proceso de **Identificación de palabras clave**, finalmente la tercera columna es la salida del proceso de **Reducción de palabras clave a características**.

Datos de salida.

La data set inicialmente obtenida de los documentos nos sirve para crear los datos de salida que formaran parte de la data set final. Esto se debe a que muestran la correcta clasificación de los documentos, por tanto, pueden ser usados como parte del entrenamiento del árbol de decisión J48.

Tabla 11: Resultados datos de salida de la data set

Identificador del documento	Oficina receptora (Resultado Clasificación)
Caso 1	Almacén

Elaboración propia

4.1.3. Data set no analizada

Una vez conocidos los datos de entrada y salida, así como el identificador del documento para una posible evaluación individual, se procede a realizar el armado de la data set no analizada, es decir una data set formada a partir de la información de los documentos. La data set fue elaborada de la siguiente forma.

Figura 9: Configuración de la data set no analizada

Identificador	Características						Resultado
	Tipo Monetario	Relación afinidad con la persona	Tratamiento del documento	Tipo de entidad involucrada	Tipo proceso	Estado	
Caso 1	Materiales	Titular	Respuesta	Escuela	No especifica	Finalizado	Almacén

Elaboración propia

Después de haber realizado todo el tratamiento a la información de los documentos se obtuvo una data set con 334 elementos.

4.1.4. Prueba Chi-cuadrada

La prueba Chi-cuadrada nos permite, determinar la asociación o independencia de dos variables cualitativas. En nuestro caso, tenemos dentro de nuestra data set, dos grupos de variables cualitativas o categóricas: Características y Resultados. A raíz de que el proceso a realizar será clasificar o encontrar los resultados en base a las características es necesario analizar si existe una relación entre cada una de las características y el resultado que se obtiene de analizarlas o usarlas en el proceso de clasificación.

Como tenemos 6 características, evaluaremos cada una con el Resultado.

Figura 10: Nivel de significancia, Características vs Resultados

Variable 1	Variable 2	Nivel de significancia
Tipo Monetario	Resultados	0
Relación Afinidad con la Persona	Resultados	0.738
Tratamiento del Documento	Resultados	0
Tipo Entidad Involucrada	Resultados	0
Tipo Proceso	Resultados	0
Estado	Resultados	0

Elaboración propia

Por lo general, un nivel de significancia (denotado como α o alfa) de 0.05 funciona adecuadamente. Un nivel de significancia de 0.05 indica un riesgo de 5% de concluir que existe una asociación entre las variables cuando no hay una asociación real.

Según nuestros cálculos, el p-valor o nivel de significancia entre Relación Afinidad con la Persona y Resultados es 0.738. Puesto que el valor p es mayor que α, se acepta la hipótesis nula. Concluyendo que las variables no están asociadas. Es decir, son independientes: Para esta investigación nosotros calculamos el resultado en base a las características y si el resultado es independiente de la característica, ello nos indica que esta característica no es útil para calcular el resultado.

Según nuestros cálculos, el p-valor o nivel de significancia entre todas las características a excepción de Relación Afinidad con la Persona y Resultados es 0. Puesto que el valor p es menor que α, se rechaza la hipótesis nula y concluye que las variables están asociadas.

Es decir, son dependientes: En nuestro análisis la variable Resultados es dependiente de todas las variables a excepción de Relación Afinidad con la Persona.

Tras el análisis realizado, se concluye que se debe eliminar o quitar de la data set la variable de Relación Afinidad con la Persona, pues no está relacionada con el Resultado y su conservación en la data set podría entorpecer el proceso de clasificación.

4.1.5. Análisis de componentes principales categórico (CATPCA)

Antes de reducir la dimensionalidad de los datos es necesario saber si dicha reducción es viable, pues si se reduce la dimensionalidad sin analizar la correlación entre las variables se podría perder información. Por ello se hizo el análisis de correlación de las Características.

Figura 11: Coeficientes de Correlación de las Características

Variable 1	Variable 2	Coeficiente de Correlación
Tipo Monetario	Tratamiento del Documento	-0.297
Tipo Monetario	Tipo Entidad Involucrada	0.433
Tipo Monetario	Tipo Proceso	-0.035
Tipo Monetario	Estado	-0.027
Tratamiento del Documento	Tipo Entidad Involucrada	-0.095
Tratamiento del Documento	Tipo Proceso	0.114
Tratamiento del Documento	Estado	0.017
Tipo Entidad Involucrada	Tipo Proceso	0.035
Tipo Entidad Involucrada	Estado	-0.051
Tipo Proceso	Estado	-0.073

Elaboración propia

Una vez obtenidos los coeficientes de correlación se procedió a interpretarlas.

- Un valor menor que 0 indica que existe una correlación negativa, es decir, que las dos variables están asociadas en sentido inverso. Cuánto más se acerca a -1, mayor es la fuerza de esa relación invertida (cuando el valor en una sea muy alto, el valor

en la otra será muy bajo). Cuando es exactamente -1, eso significa que tienen una correlación negativa perfecta.

- Un valor mayor que 0 indica que existe una correlación positiva. En este caso las variables estarían asociadas en sentido directo. Cuanto más cerca de +1, más alta es su asociación. Un valor exacto de +1 indicaría una relación lineal positiva perfecta.
- Finalmente, una correlación de 0, o próxima a 0, indica que no hay relación lineal entre las dos variables.

El coeficiente de correlación más cercano a -1 o 1 es 0.433 (Coeficiente de correlación de las variables Tipo Monetario y Tipo Entidad Involucrada). La correlación entre estas 2 variables es regular, por tanto, una reducción de dimensiones asociando estas 2 variables no es viable.

Luego del análisis de correlación de las variables se llegó a la conclusión que no es posible reducir dimensiones de la data set asociando variables.

4.1.6. Data set Final

Luego de realizar el análisis de Chi-cuadrada y el CATPCA, se obtuvo la configuración final de la data set, la cual será usada por el árbol de decisión J48. La data set final consta de 334 elementos.

Figura 12: Configuración de la data set final

Identificador	Características					Resultado
	Tipo Monetario	Tratamiento del documento	Tipo de entidad involucrada	Tipo proceso	Estado	
Caso 1	Materiales	Respuesta	Escuela	No especifica	Finalizado	Almacén

Elaboración propia

La primera columna presenta el identificador del documento, para proteger la confidencialidad de los documentos se les asigno un identificador ajeno al que usa la UGEL para darles seguimiento.

De la segunda hasta la séptima columna representan los datos de entrada que recibirá el árbol de decisión J48 para realizar el proceso de clasificación.

Para finalizar, en la última columna se encuentran los datos de salida que recibirá el árbol de decisión J48 para realizar el proceso de entrenamiento y posterior verificación de clasificación. Esta data set puede ser leído perfectamente por el árbol desarrollado, ya que fue creado y personalizado expresamente para esta investigación. (véase anexo 4)

4.1.7. Ocurrencias encontradas

- El proceso de recopilación de la información de los documentos es muy tedioso para ser realizado de forma manual, en esta ocasión fue especialmente difícil ya que los documentos no entraban en el área de Gestión Administrativa a una hora especifica.
- El proceso de obtención de datos de los documentos se puede agilizar haciendo uso de otras herramientas de I.A., tales como reconocedores de texto y/o voz para reducir el tiempo necesario para realizar esta tarea.

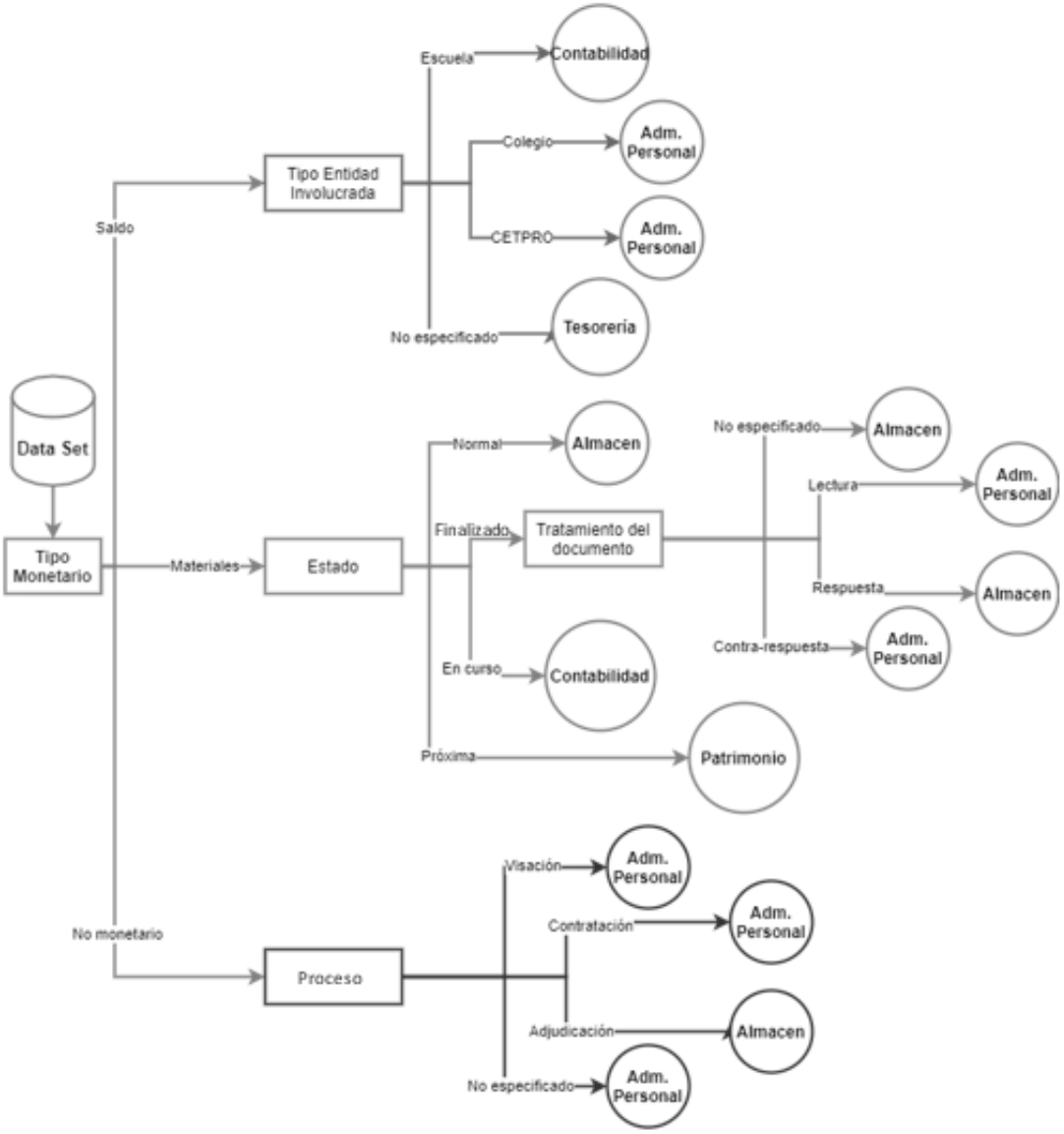
- Antes de crear la data set final y las características que estarían presentes en los datos de entrada, se debe hacer un proceso de análisis de los datos obtenidos de los documentos. Dicho análisis debe ser realizado con la ayuda de expertos de la UGEL a fin de crear variables o características representativas de los documentos.

4.2. CONFIGURACIÓN DEL ÁRBOL DE DECISIÓN J48

Por definición un árbol de decisión se genera a partir de un conjunto de datos previamente normalizados. Para esta investigación se hace uso de la herramienta WEKA el cual tiene incorporado el algoritmo del árbol de decisión J48, y la data set generada, lográndose encontrar una configuración óptima para el análisis de la fiabilidad del árbol J48 en la correcta clasificación de documentos. A continuación, se explica detalladamente la configuración del árbol de decisión J48, así como un ejemplo del funcionamiento del árbol generado.

Usando el programa WEKA y la data set se obtuvo el siguiente árbol de decisión J48. (véase anexo 5)

Figura 13: Configuración árbol de decisión J48 obtenido

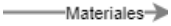


Elaboración propia

Algunas consideraciones tomadas en cuenta para la obtención del árbol J48 fue establecer el número de objetos mínimos que contendría cada hoja del árbol, pues si bien se podía mejorar ligeramente el resultado generando hojas con 1 solo documento clasificado el árbol resultante de ese proceso no sería representativo para solucionar el problema inicial ya que dicho árbol hubiera sido creado con Overfitting y por definición un árbol con Overfitting no sirve para predecir nuevos datos ajenos al data set original.

Explicación de figuras

Tabla 12: Explicación de figuras presentes en el árbol de decisión J48

Figura	Descripción
	Es la data set ya normalizada que entrara en el árbol de decisión.
	Característica por la cual será clasificada en ese nodo.
	Valor de la característica por la cual fue dividida o clasificada en ese nivel.
	Resultado de la clasificación u oficina a la cual sería enviada un documento de acuerdo a la clasificación del árbol de decisión J48.

Elaboración propia

4.2.1. Ejemplo de funcionamiento.

Caso 1

Recepción de los datos de entrada. En este caso son los valores de las características.

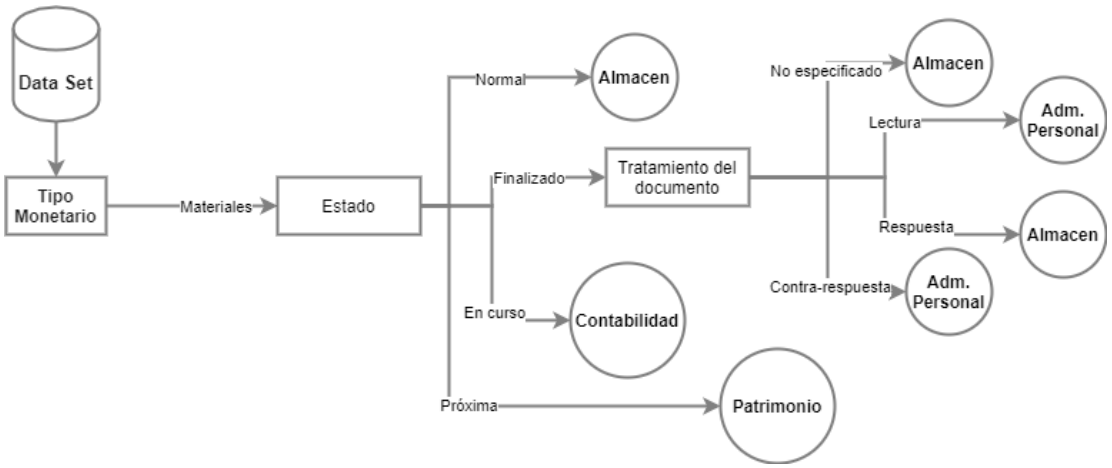
- Materiales
- Titular
- Respuesta
- Escuela
- No especifica
- Finalizado

<u>Paso 1: Nodo de entrada, clasificación.</u>

En cada nodo se da el proceso de clasificación en base al valor de la característica eligiendo así una rama u hoja del árbol de decisión J48. Para este paso la característica del nodo es: Tipo monetario y el valor para esa característica de nuestra entrada es

"Materiales". Dicho esto, se eliminan las otras ramas salientes de Tipo Monetario a excepción de "Materiales".

Figura 14: Representación del paso 1. CASO 1



Elaboración propia

Paso 2: Repetición del paso anterior, hasta llegar al resultado.

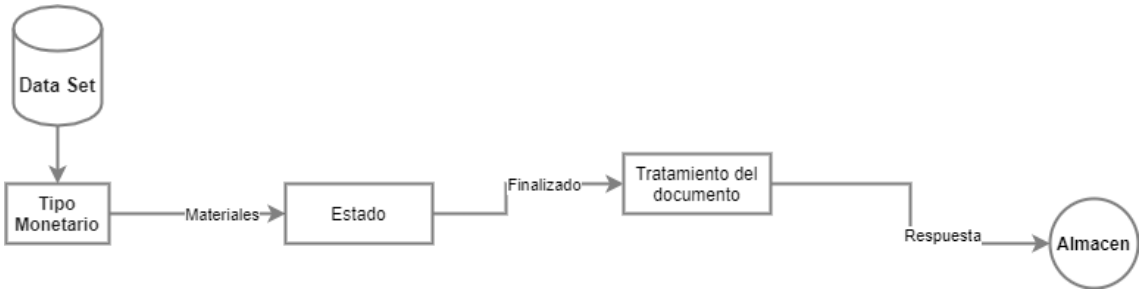
Ahora la característica a evaluar es Estado y su valor en nuestra entrada es "Finalizado". Por tanto, se eliminan las otras ramas salientes de Estado a excepción de "Finalizado".

Figura 15: Representación del paso 2, primera iteración. CASO 1

Elaboración propia

Ahora la característica a evaluar es Tratamiento del documento y su valor en nuestra entrada es "Respuesta". Por lo tanto, se eliminan las otras ramas salientes de la característica Tratamiento del documento a excepción de "Respuesta".

Figura 16: Representación del paso 2, segunda iteración. CASO 1



Elaboración propia

Paso 3: Interpretación del resultado.

Cuando ya no hay más características que evaluar el árbol nos muestra la respuesta para la clasificación del documento. En este caso la respuesta es “Almacen”.

Figura 17: Representación del paso 3. CASO 1

Elaboración propia

4.2.2. Ocurrencias encontradas

Por la forma en que fue configurado el árbol J48 algunas características no son usadas en el proceso de generación del árbol. Lo cual indica que se puede mejorar la data set mediante un mejor análisis de los datos de entrada mejorando así el árbol J48 final.

Características no usadas:

- Relación Afinidad con la persona

4.3. RESULTADOS DE FIABILIDAD DEL ÁRBOL DE DECISIÓN J48

Para determinar la fiabilidad en la correcta clasificación de los documentos del árbol J48, el proceso se limitó a interpretar la matriz de confusión resultante y el uso de las fórmulas de cálculo de exactitud. Adicionalmente, por el hecho de usar la validación cruzada para el proceso de generación del árbol, la fiabilidad o precisión obtenida se conserva incluso para nuevos datos o entradas.

Matriz de confusión resultante

La matriz de confusión se obtuvo del uso de la herramienta WEKA y su método de clasificación Árbol de decisión J48 usando la data set generada para esta investigación.

Figura 18: Matriz de confusión obtenida

		Valores predichos				
		Almacen	Adm. Personal	Tesorería	Patrimonio	Contabilidad
Valores Reales	Almacen	60	5	2	0	3
	Adm. Personal	9	86	1	0	5
	Tesorería	3	5	81	2	2
	Patrimonio	0	4	1	21	2
	Contabilidad	5	6	0	3	28

Elaboración propia

Esta figura presenta los datos obtenidos a través del proceso de clasificación de documentos usando el árbol de decisión J48 (Matriz de confusión). Los datos obtenidos son los siguientes:

- 60 documentos fueron clasificados para Almacen de forma correcta.
- 86 documentos fueron clasificados para Adm. Personal de forma correcta.

- 81 documentos fueron clasificados para Tesorería de forma correcta.
- 21 documentos fueron clasificados para Patrimonio de forma correcta
- 28 documentos fueron clasificados para Contabilidad de forma correcta.

Adicionalmente, el total de documentos enviados originalmente a cada oficina se obtuvo de la **Figura A. *16*** del anexo 6.

4.3.1. Interpretación de la matriz de confusión

Interpretación General

Para interpretar resultados obtenidos respecto a la clasificación de documentos usando el árbol de decisión J48, primero se debe calcular la exactitud y tasa de error en la clasificación. Para ello, se hace uso de las siguientes ecuaciones.

Ecuaciones

Ecuación: Cálculo de exactitud o fiabilidad en la clasificación.

$$Exactitud = \frac{Documentos\ Clasificados\ de\ manera\ correcta}{Documentos\ Clasificados\ Incorrectamente + Documentos\ Clasificados\ de\ manera\ correcta} \quad (2)$$

Ecuación: Cálculo de tasa de error en la clasificación.

$$Tasa\ de\ error = \frac{Documentos\ Clasificados\ Incorrectamente}{Documentos\ Clasificados\ Incorrectamente + Documentos\ Clasificados\ de\ manera\ correcta} \quad (3)$$

Ecuación: Total de documentos clasificados de manera correcta.

$$Documentos\ Clasificados\ de\ manera\ correcta = \sum Documentos\ Clasificados\ de\ manera\ correcta\ por\ oficina \quad (4)$$

Ecuación: Total documentos clasificados de manera incorrecta.

$$Documentos\ Clasificados\ Incorrectamente = \sum Documentos\ clasificados\ de\ manera\ erronea\ por\ oficina \quad (5)$$

Cálculo de exactitud

Para el cálculo de la exactitud o fiabilidad del árbol de decisión en la correcta clasificación de documentos, primero hay que calcular el total de documentos clasificados correcta e incorrectamente haciendo uso de las ecuaciones **(** 4) y **(** 5).

$$Documentos\ Clasificados\ de\ manera\ correcta = 60 + 86 + 81 + 21 + 28$$

$$Documentos\ Clasificados\ de\ manera\ correcta = 276$$

$$Documentos\ Clasificados\ Incorrectamente = 10 + 15 + 12 + 7 + 14$$

$$Documentos\ Clasificados\ Incorrectamente = 58$$

Una vez calculados el total de documentos clasificados correcta e incorrectamente, procedemos a calcular la exactitud y tasa de error en la clasificación de documentos haciendo uso de las ecuaciones **(** 2) y **(** 3).

$$Exactitud = \left(\frac{276}{276 + 58}\right) * 100$$

$$Exactitud = 82.63\ \%$$

$$Tasa\ de\ error = \left(\frac{58}{276 + 58}\right) * 100$$

$$Tasa\ de\ error = 17.37\%$$

- ✓ Fiabilidad o Exactitud de la clasificación de documentos usando el árbol de decisión J48 = 82.63%

- ✓ Tasa de error en la clasificación de documentos usando el árbol de decisión J48 = 17.37%

Los resultados muestran que a partir de diversas características empleadas tales como: Tipo Monetario, Tipo Proceso, Estado, Tratamiento del documento y Tipo Entidad Involucrada el porcentaje de aciertos del árbol de decisión J48 es superior al 80%.

Interpretación especifica

Para interpretar los resultados obtenidos respecto a la clasificación de documentos usando el árbol de decisión J48 de forma específica (por oficina), se hizo uso de la ecuación **(** 6) y los datos obtenidos en la matriz de confusión.

$$Exactitud\ por\ oficina = \left(\frac{Total\ documentos\ clasificados\ a\ la\ oficina}{Total\ documentos\ dirigidos\ a\ la\ oficina}\right) * 100 \qquad (6)$$

Ejemplo

$$Exactitud\ (Almacen) = (60/70)$$

$$Exactitud\ (Almacen) = 85.71$$

Tabla 13: Resumen de exactitud árbol J48 por Oficina

Oficina	Total de documentos	Clasificados Correctamente	Exactitud por Oficina
Almacen	70	60	85.71
Adm. Personal	101	86	85.15
Tesorería	93	81	87.1
Patrimonio	28	21	75
Contabilidad	42	28	66.7

Elaboración propia

- Los documentos dirigidos a las oficinas de “Almacen”, “Adm. Personal” y “Tesorería” fueron clasificados con una exactitud mayor al 85% lo cual indica que el árbol J48 realiza un buen trabajo clasificando los documentos dirigidos a estas oficinas.
- Los documentos dirigidos a las oficinas de “Patrimonio” y “Contabilidad” fueron clasificados con una exactitud menor al 80% lo cual indica que el árbol J48 realiza un trabajo regular clasificando los documentos dirigidos a estas oficinas.
- Por tanto, el análisis especifico nos muestra que hay menor fiabilidad o exactitud en los documentos que tienen poca representación en la data set obtenida.

4.4. DISCUSIÓN

En la presente investigación, se muestra de manera clara y detallada el proceso por el cual se realizó el análisis de fiabilidad del árbol de decisión J48 en la correcta clasificación de los documentos del Área de Gestión Administrativa de la UGEL El Collao-Ilave.

Haciendo uso de las herramientas estadísticas, se ha medido la fiabilidad del árbol de decisión J48, los resultados indican que el árbol de decisión J48 puede ser una herramienta fiable para la correcta clasificación de documentos de la UGEL El Collao-Ilave pues el porcentaje de exactitud de su clasificación es de 82.63% y obteniendo una Tasa de error del 17.37%.

Haciendo uso de las herramientas estadísticas se ha medido la variabilidad en el porcentaje de clasificación correcta de los documentos del área de Gestión Administrativa con destino oficinas dentro del misma área, dando como resultado una precisión o fiabilidad en el pre test del 80.24% y una precisión o fiabilidad en el post test del 82.63%, evidenciando de esta forma una mejora en la correcta clasificación de los documentos.

Adicionalmente haciendo una inferencia aún más detallada se infiere que las oficinas que tienen un mayor error en su clasificación son aquellas las cuales no tienen muchos documentos en la data set, mientras que las oficinas que tienen menor error en su clasificación son aquellas que tienen muchos elementos en la data set.

Por lo anterior, a partir de los resultados encontrados, en consecuencia, se confirma la hipótesis general que establece que el árbol J48 puede ser una herramienta fiable en la correcta clasificación de documentos de la UGEL El Collao-Ilave.

V. CONCLUSIONES

PRIMERO: La fiabilidad o exactitud obtenida del uso del árbol de decisión J48 en la correcta clasificación de documentos del área de Gestión Administrativa de la UGEL El Collao-Ilave fue del 82.63%, mientras que el mismo proceso realizado sin el árbol fue del 80.24%. Concluyendo así, que el árbol de decisión J48 mejoró ligeramente la fiabilidad de la clasificación de documentos. Y su uso extendido en toda la UGEL sería viable, pues la exactitud o fiabilidad del árbol en la correcta clasificación de los documentos puede ser mejorado con el incremento en el tamaño de la data set y aplicando mejores métodos para la conversión de documentos a datos de entrada.

SEGUNDO: La generación de la data set necesaria para el árbol J48 evaluado en este trabajo de investigación fue el proceso más largo y complicado, siendo necesarios múltiples técnicas de normalización de datos y ayuda por parte de las personas a cargo de clasificar los documentos de la UGEL sin el uso de ninguna herramienta. La data set generada tuvo que ser adecuada para su uso por parte del árbol J48.

TERCERO: Para la correcta configuración del árbol de decisión J48 fue necesario una data set normalizada y conocimiento previo del proceso de configuración de un árbol de decisión en la herramienta WEKA. Adicionalmente, fue necesario el uso de la validación cruzada durante el proceso de entrenamiento para generar un árbol de decisión J48 que no tenga problemas de Overfitting.

CUARTO: Determinar la fiabilidad o exactitud del árbol de decisión J48 en la correcta clasificación de documentos fue simple, pues el proceso se limitó a interpretar la matriz de confusión resultante de la clasificación de documentos usando el árbol de decisión J48, obteniendo así una fiabilidad del 82.63%.

VI. RECOMENDACIONES

PRIMERO: Antes de implementar una herramienta de I.A. como solución a un problema, se recomienda analizarla y evaluarla. Y de ser ineficaz en su trabajo, optar por usar una herramienta de I.A. distinta o usar simultáneamente varias.

SEGUNDO: La elaboración de una propia data set permite personalizar una herramienta de I.A. para de esa forma obtener el mejor rendimiento que este pueda ofrecer. La elaboración de una data set propia a partir de documentos requiere un trabajo extenso. Por lo tanto, se recomienda el uso de Reconocimiento de texto o Reconocimiento de voz para optimizar el tiempo en el cual se obtienen los datos de los documentos.

TERCERO: Al tratarse de documentos pertenecientes a una UGEL, debe considerarse que por temporadas o trimestres existen procesos que se dan exclusivamente en ese periodo de tiempo. Por ejemplo: navidad, año nuevo, fiestas patrias, viernes santo, entre otros. Es por ello que para mejorar la representatividad de la data set, se recomienda que está debe ser conformada tras un análisis de los documentos por un periodo no menor a 1 año.

CUARTO: Se recomienda la planificación y diseño de un método eficiente para la obtención de los datos, ya que para validar o establecer un análisis de una herramienta de I.A. se debe contar con una cantidad considerable de datos.

VII. REFERENCIAS

Akujuobi, U., & Zhang, X. (2017). Delve: a dataset-driven scholarly search and analysis system. *ACM SIGKDD Explorations Newsletter*, *19*(2), 36–46.

Ariza-lópez, F. J., Rodríguez-avi, J., Las, C., España, J., Estadística, D. De, & Universidad, O. (2018). *CONTROL ESTRICTO DE MATRICES DE CONFUSIÓN POR MEDIO DE DISTRIBUCIONES MULTINOMIALES Dep . de Ingeniería Cartográfica , Geodésica y Fotogrametría . Universidad de Jaén . CM (i , j) = [# items of class (j) of the RDS classified as class (i) of the CDS*. 215–227.

Barrientos, R., Cruz, N., Acosta, H., Rabatte, I., Gogeascoechea, M., Pavón, P., & Blázquez, S. (2009). Árboles De Decisión Como Herramienta En El Diagnóstico Médico. *Articulo Original*, 20–24. https://www.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf

Carlos Fernández Collado, P. B. L. (2014). *Metodología de la Investigación* (6th ed.).

Carlos, J., García, L., & Gabriel, F. (2009). *ALGORITMOS Y PROGRAMACIÓN*. 1–96.

CIAT, C. I. D. A. T. (1999). Guia Para El Uso De “Árboles De Decision.” *Ministerio de Agricultura y Desarrollo Rural*, 1–46.

Collao-Ilave, U. El. (2016). *Manual de Organización y Funciones de la Unidad de Gestión Educativa Local El Collao 2016*. https://www.peru.gob.pe/docs/PLANES/16384/PLAN_16384_2016_MOF_2016_UGEL_EL_COLLAO.PDF

Duque, F., Saint-Priest Velásquez, Y., Segovia, P., & Loaiza, D. F. (2017). *Algoritmos Y Programación En Pseudocódigo*. www.editorialtecnologica.tec.ac.cr

Frank, E. (2000). Pruning decision trees and lists. *Science*, *300*(January), 204. http://www.cs.waikato.ac.nz/~ml/publications/2000/thesis.final.pdf

Karabadji, Seridi, Bousetouane, Dhifli, & A. (2016). The value of decision tree analysis in planning anaesthetic care in obstetrics. *International Journal of Obstetric Anesthesia*. https://doi.org/https://doi.org/10.1016/j.ijoa.2016.02.007

Karabadji, N. E. I., Seridi, H., Bousetouane, F., Dhifli, W., & Aridhi, S. (2017). An evolutionary scheme for decision tree construction. *Knowledge-Based Systems*, *119*, 166–177. https://doi.org/10.1016/j.knosys.2016.12.011

Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *Information Processing and Management*, *56*(5), 1618–1632. https://doi.org/10.1016/j.ipm.2019.05.003

López Takeyas, B. (2005). *Inteligencia Artificial: Algoritmo C4.5*. 1–15.

M. Ricardo, C. (2009). *Bases de Datos*.

Mondragon, M. (2014). Uso de la correlación de Spearman en un estudio de intervención en fisioterapia. *Movimiento Científico*, *8*(1), 98–104. https://dialnet.unirioja.es/servlet/articulo?codigo=5156978

Ocaña-Fernández, Y., Valenzuela-Fernández, L. A., & Garro-Aburto, L. L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones*, *7*(2), 536–552. https://doi.org/10.20511/pyr2019.v7n2.274

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio. *International Journal of Morphology*, *35*(1), 227–232. https://doi.org/10.4067/S0717-95022017000100037

Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J. P., & Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de Teledeteccion*, *2015*(44), 55–65. https://doi.org/10.4995/raet.2015.4153

Romero, J. J., Dafonte, C., Gòmez, Á., & Penousal, F. J. (2007). Inteligencia Artificial Y Computación Avanzada. In *Inteligencia Artificial …*. http://fmachado.dei.uc.pt/wp-content/papercite-data/pdf/ms07.pdf#page=9

Russell, S. (2004). *Inteligencia Artificial Un Enfoque Moderno* (2° Edición). https://www.iberdrola.com/te-interesa/tecnologia/que-es-inteligencia-artificial

Savira, F., & Suharsono, Y. (2013). Pattern Recognition and Neural Networks. *Journal of Chemical Information and Modeling*, *01*(01), 1689–1699.

Silberschatz, A. (Bell L., Korth, H. F. (Bell L., & Sudarshan, S. (Instituto Indio de Tecnología, B. (2002). Fundamentos de bases de datos. In *Victoria*.

Varela Arregocés Edwin Campbells Sánchez, E., & Simón Bolívar Barranquilla - Atlántico, U. (2011). Redes Neuronales Artificiales: Una Revisión del Estado del Arte, Aplicaciones Y Tendencias Futuras Artificial Neural Networks: A Brief Review. *Investigación y Desarrollo En TIC*, *2*, 18–27. http://publicaciones.unisimonbolivar.edu.co/rdigital/inovacioning/index.php/identic/article/viewFile/21/29

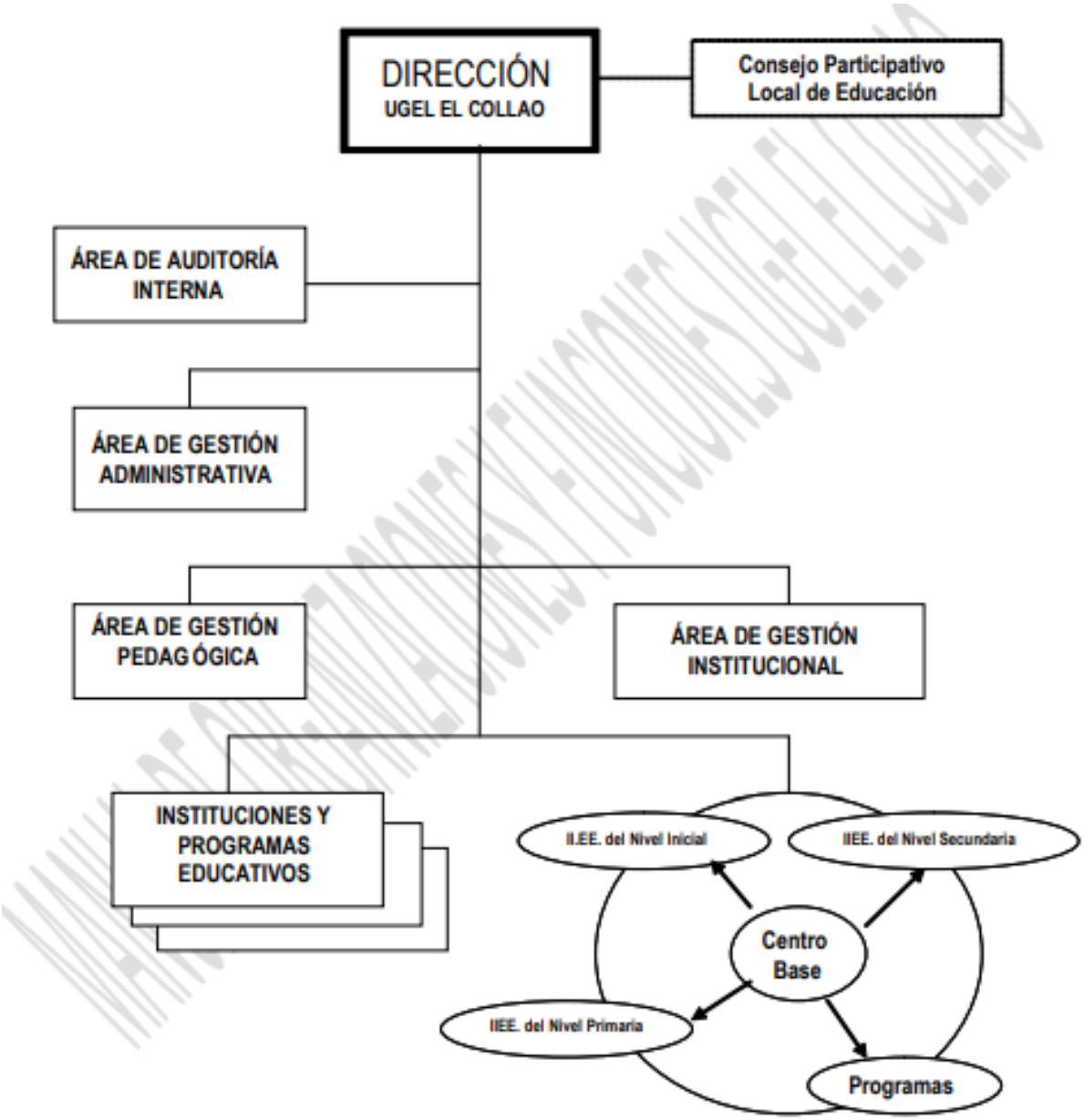
Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2). https://doi.org/10.1088/1742-6596/1168/2/022022

Zhao, X., Bi, X., Wang, G., Zhang, Z., & Yang, H. (2016). Uncertain XML documents classification using Extreme Learning Machine. *Neurocomputing*, *174*, 375–382. https://doi.org/10.1016/j.neucom.2015.02.095

ANEXOS

Anexo 1. Estructura de la Unidad de Gestión Educativa Local El Collao

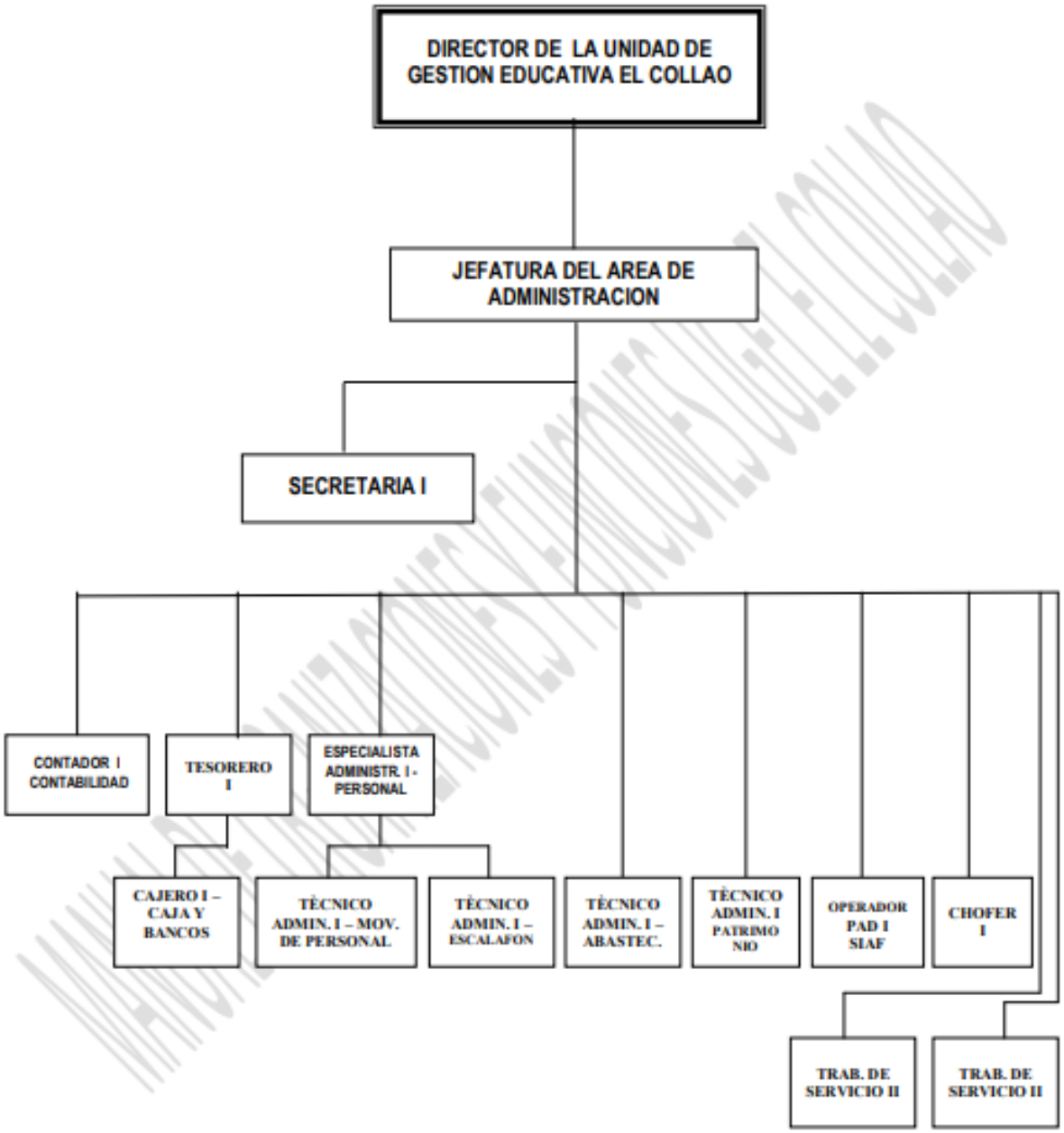
Figura A. 1: Estructura de la Unidad de Gestión Educativa Local El Collao



Fuente: (Collao-Ilave, 2016)

Anexo 2. Estructura del área de Gestión Administrativa

Figura A. 2: Estructura del área de Gestión Administrativa



Fuente: (Collao-Ilave, 2016)

Anexo 3. Desarrollo pseudocódigo del algoritmo C4.5

Para explicar el desarrollo del proceso de creación de un árbol de decisión, haciendo uso del pseudocódigo del algoritmo C4.5 se pone como ejemplo el siguiente conjunto de datos. Y el árbol generado al final contestara la siguiente pregunta.

¿Cuál aplicación recomendamos?

- Para una mujer que trabaja.
- Para un hombre que trabaja.
- Para alguien que va al colegio

Figura A. 3: Data de entrada, caso muestra

Genero	Ocupación	Aplicación
F	Colegio	Chess
F	Trabajo	Facebook
M	Trabajo	Whatsapp
F	Trabajo	Facebook
M	Colegio	Chess
M	Colegio	Chess

Elaboración propia

Paso 1: Comprobar los casos base.

- ✓ Caso 1: Mujer, va al colegio y elige la aplicación Chess.
- ✓ Caso 2: Mujer, trabaja y elige la aplicación Facebook.
- ✓ Caso 3: Hombre, trabaja y elige la aplicación Whatsapp.
- ✓ Caso 4: Mujer, trabaja y elige la aplicación Facebook.

- ✓ Caso 5: Hombre, va al colegio y elige la aplicación Chess.
- ✓ Caso 6: Hombre, va al colegio y elige la aplicación Chess.

Paso 2: Que atributo al ser dividido maximiza la ganancia de información

Los atributos son, las características por las cuales se puede dividir los datos de entrada para convertirlos en una salida, en este caso como la salida es que aplicación se recomienda, por lo cual los atributos a evaluar son **Genero** y **Ocupación**.

Primero, se calcula la entropía general del nodo raíz.

$$Entropia\ nodo\ raíz = \left(-\frac{3}{6}\right) log_2 \left(\frac{2}{6}\right) - \left(\frac{2}{6}\right) log_2 \left(\frac{2}{6}\right) - \left(\frac{1}{6}\right) log_2 \left(\frac{1}{6}\right) = 1.46$$

Segundo, visualización de las posibles divisiones y cálculo de las entropías de estas.

Caso 1: División por Genero.

Nodo hijo Género Femenino

Figura A. 4: Nodo hijo género femenino, caso muestra

Genero	Ocupación	Aplicación
F	Colegio	Chess
F	Trabajo	Facebook
F	Trabajo	Facebook

Elaboración propia

Nodo hijo Género Masculino

Figura A. 5: Nodo hijo género masculino, caso muestra

Genero	Ocupación	Aplicación
M	Trabajo	Whatsapp 
M	Colegio	Chess 
M	Colegio	Chess 

Elaboración propia

Cálculo de entropía por cada nodo hijo generado.

$$Entropy\ Nodo\ Genero\ Femenino = 0.92$$

$$Entropy\ Nodo\ Genero\ Masculino = 0.92$$

Cálculo de ganancia de información.

$$Ganancia\ información = Entopia\ Nodo\ padre - \frac{\sum Entropia\ Nodos\ Hijos}{\#\ Nodos\ Hijos}$$

$$Ganancia\ información = 1.46 - \frac{0.92 + 0.92}{2} = 0.54$$

Caso 2: División por ocupación.

Nodo hijo Ocupación Trabajo

Figura A. 6: Nodo hijo ocupación trabajo, caso muestra

Genero	Ocupación	Aplicación
F	Trabajo	Facebook
M	Trabajo	Whatsapp
F	Trabajo	Facebook

Elaboración propia

Nodo Hijo Ocupación Colegio

Figura A. 7: Nodo hijo ocupación colegio, caso muestra

Genero	Ocupación	Aplicación
M	Colegio	Chess
M	Colegio	Chess
F	Colegio	Chess

Elaboración propia

Cálculo de entropía por cada nodo hijo generado.

$$Entropy\ Nodo\ Ocupación\ Trabajo = 0.92$$

$$Entropy\ Nodo\ Ocupación\ Colegio = 0$$

Cálculo de ganancia de información.

$$Ganancia\ información = Entopia\ Nodo\ padre - \frac{\sum Entropia\ Nodos\ Hijos}{\#\ Nodos\ Hijos}$$

$$Ganancia\ información = 1.46 - \frac{0 + 0.92}{2} = 1$$

Paso 3: Establecer cuál atributo tiene mayor ganancia de información al dividirlo.

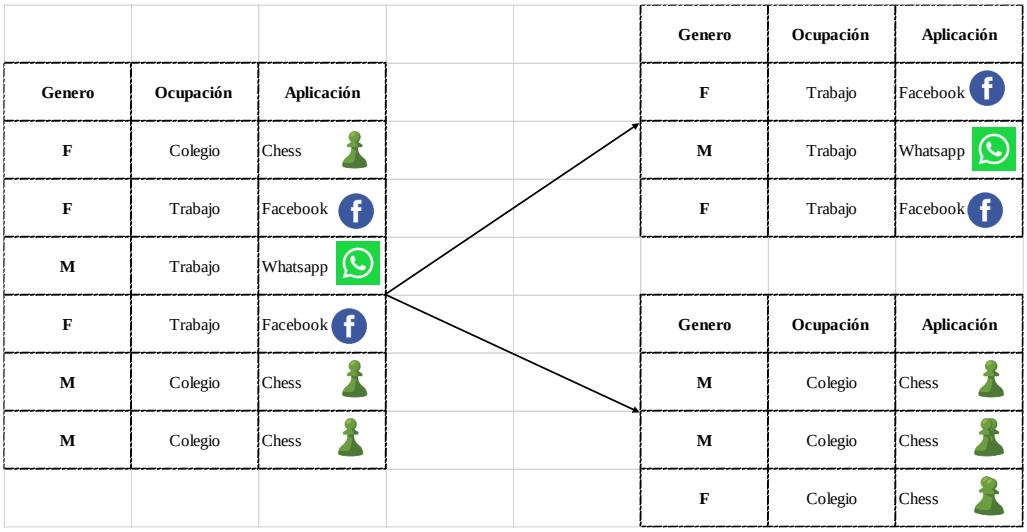
$$Ganancia\ de\ información\ dividiendo\ por\ Genero <$$

$$Ganancia\ de\ información\ dividiendo\ por\ Ocupación$$

Se establece que la mayor ganancia de información se da cuando el proceso de división en hijos se usa el atributo de Ocupación.

Paso 4: Crear la división en el nodo en base al atributo seleccionado.

Figura A. 8: Resultado obtenido, primera iteración del caso muestra



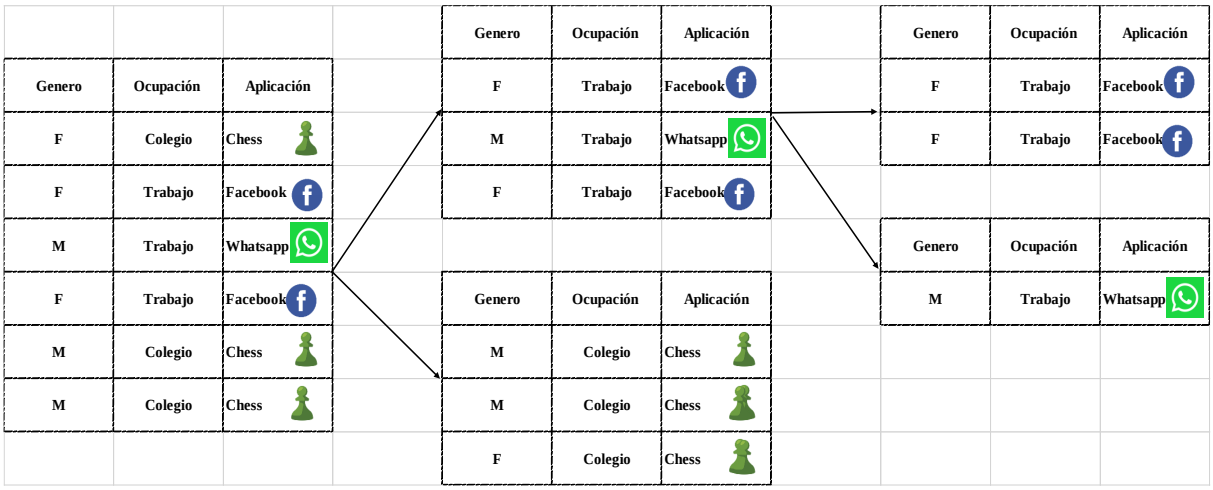
Elaboración propia

Paso 5: Repetir el procedimiento hasta que el nodo resultante tenga una entropía de "1" o no pueda dividirse más.

Dato a tomar en consideración. Solo en la primera iteración se calcula la entropía del nodo raíz, a excepción del primer nodo todos los nodos heredan la entropía que fue calculada anteriormente. Por ejemplo, en nuestro caso de estudio, el nodo hijo de ocupación trabajo tiene por entropía 0.92, mientras que el nodo hijo de ocupación colegio tiene una entropía de 1, por tanto, este nodo no necesita ser evaluado otra vez.

Este proceso de cálculo y división de los datos se realiza de forma automática implementando los algoritmos que sean necesarios. Generando así un árbol de decisión.

Figura A. 9: Resultado final, obtenido de la segunda iteración del caso muestra



Elaboración propia

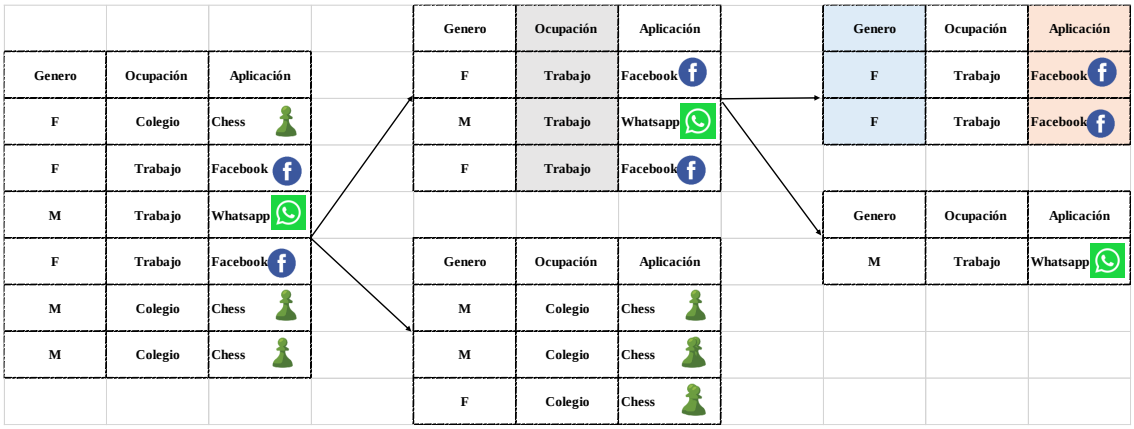
Prueba de ello es que podemos responder a la pregunta inicial, que aplicación se recomienda.

- Para una mujer que trabaja.
- Para un hombre que trabaja.
- Para alguien que va al colegio

Prueba del árbol de decisión generado:

- **Para una mujer que trabaja.**

Figura A. 10: Resultado obtenido usando el árbol para una mujer que trabaja

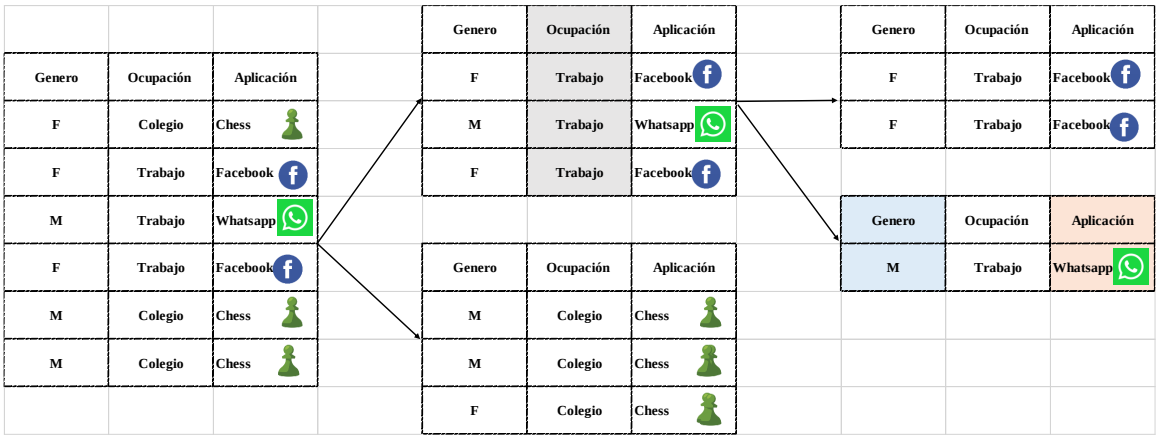


Elaboración propia

Para una mujer que trabaja, primero entra en el árbol y se va a la rama que dividió por ocupación, luego una vez allí se divide por género, siendo el caso una mujer, llegando así a un nodo final. Por tanto, la aplicación que se le recomendaría será “Facebook”.

- **Para un hombre que trabaja.**

Figura A. 11: Resultado obtenido usando el árbol para un hombre que trabaja

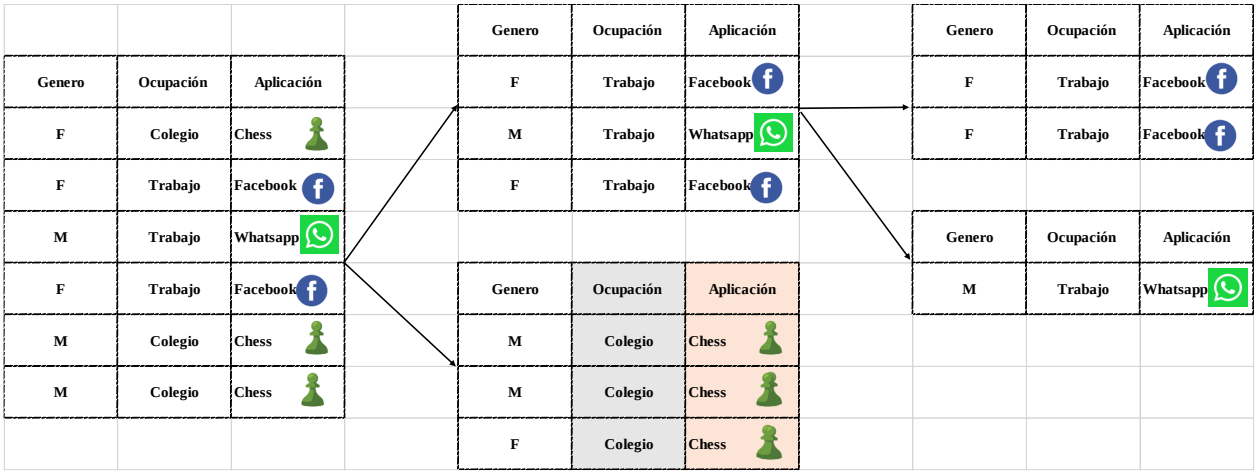


Elaboración propia

Para un hombre que trabaja, primero entra en el árbol y se va a la rama que dividió por ocupación, luego una vez allí se divide por género, siendo el caso un hombre, llegando así a un nodo final. Por tanto, la aplicación que se le recomendaría será “Whatsapp”.

- **Para alguien que va al colegio**

Figura A. 12: Resultado obtenido usando el árbol para alguien que va al colegio

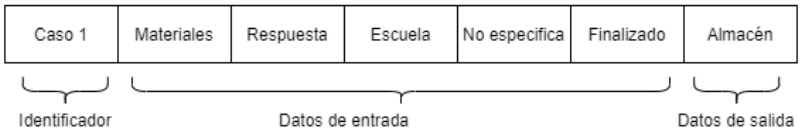


Elaboración propia

Para alguien que va al colegio, primero entra en el árbol y se va a la rama que dividió por ocupación, llegando así a una hoja o nodo final. Por tanto, la aplicación que se le recomendaría será “Chess”.

Anexo 4: Descripción detallada data set.

Figura A. 13: Descripción data set



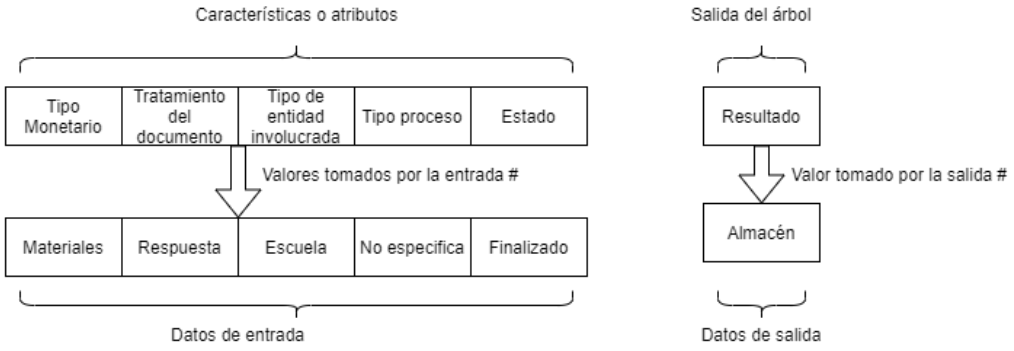
Elaboración propia

ID dato: Es el identificador del elemento bajo el cual se puede evaluar si la clasificación del elemento (documento reducido a datos) es correcta. También representa la cantidad de datos o elementos que conforman la data set.

Datos de entrada: Son el conjunto de características extraídas de los documentos, los cuales conforman los datos de entrada para el árbol de decisión J48 que se usó en el trabajo de investigación. Los datos de entrada son

Datos de salida: Es la oficina a la cual el documento debería ser enviado, también sirve como data test para el árbol de decisión J48.

Figura A. 14: Descripción formato data set: datos de entrada y salida

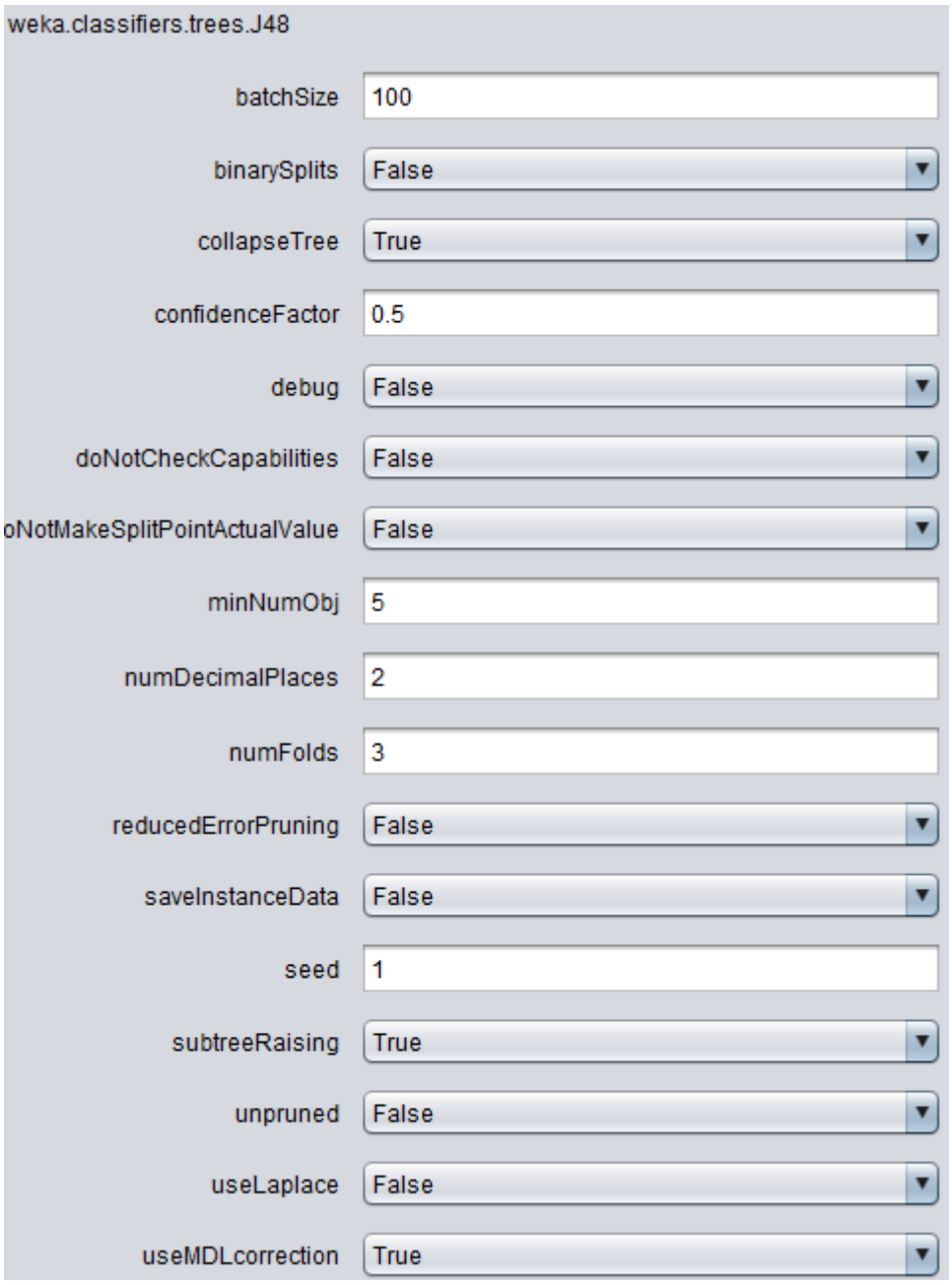


Elaboración propia

Nota: La data set se almacenó en un archivo csv.

Anexo 5: Configuración en WEKA del árbol J48.

Figura A. 15: Configuración del árbol J48 en el software WEKA



Número de documentos usados para la validación cruzada: 90

Anexo 6: Clasificación de los documentos sin el uso del árbol J48.

Figura A. 16: Resumen de clasificación de documentos sin el uso del árbol J48

Oficina receptora final	Total documentos atendidos	Total Documentos atendidos 1° clasificación	Total Documentos enviados a otra oficina
Adm. Personal	101	70	31
Tesorería	93	90	3
Almacén	70	53	17
Contabilidad	42	31	11
Patrimonio	28	24	4

Elaboración propia

En la anterior figura se observa el resumen por oficina de la clasificación de los documentos sin el uso del árbol de decisión J48. Para obtener el porcentaje o precisión en la clasificación de documentos sin el uso del árbol J48 usaremos la siguiente formula.

$$Precisión\ clasificación \text{ sin } el\ árbol\ J48 =$$

$$((Total\ Documentos\ clasificados\ correctamente)/$$

$$(Documentos\ totales\ atentidos\ ))\ *100$$

$$Total\ Documentos\ clasificados\ correctamente = 70 + 90 + 53 + 31 + 24$$

$$Documentos\ totales\ atentidos\ = 101 + 93 + 70 + 42 + 28$$

$$Precisión\ clasificación \text{ sin } el\ árbol\ J48 = \left(\frac{268}{334}\right) * 100$$

$$Precisión\ clasificación \text{ sin } el\ árbol\ J48 = \ 80.24$$

Anexo 7: Documentos de la UGEL Formulario Único de Tramite (FUT)

Figura A. 17: Ejemplo Formato Documento Presentado