Universidad Católica de Salta

Minería de imágenes aplicada a digitalizaciones de legajos de clientes.

Gabriela Moya

Carrera: Ingeniería en Informática

Facultad de Ingeniería

2021

Título

Ingeniera en Informática

Profesor guía

Nombre: Firma:_____

Tribunal evaluador

Nombre: Firma:_____

Nombre: Firma:_____

Nombre: Firma:_____

Alumna

Nombre: Moya, Gabriela Jorgelina Firma:_____

Fecha de exposición

Dedicatorias

A mi papá, en su memoria.

A mi mamá, quien me dio todo.

Agradecimientos

A la Lic. Carolina Cardozo quien fue mi profesora guía. Valoro inmensamente la ayuda y paciencia durante la ejecución y corrección del proyecto.

A la Lic. Nancy Lucero, quien tuvo las palabras justas para motivarme cuando lo necesitaba.

A Paula Castellanos quien voluntariamente me asistió en la revisión de la redacción de este trabajo.

A mi familia y amigos, quienes siempre me acompañaron y alentaron durante la carrera y en especial en la etapa final.

Contenido

1. Introducción 6
2. Estado de la cuestión 8
2.1 Minería de datos 8
2.2 Herramientas de minería de datos 10
2.3 Metodologías de minería de datos 11
2.4 Procesamiento de imágenes 14
2.5 Minería de imágenes 14
2.6 Aplicación de minería de imágenes 15
3. Definición del problema 16
3.1 Descripción de la empresa 16
3.2 Objetivos del negocio 19
3.3 Objetivo general de la solución 19
3.4 Objetivos específicos de la solución 19
3.5 Objetivos de la minería de datos 19
3.6 Criterio de éxito de la minería de datos 19
3.7 Alcance 19
4. Solución propuesta 20
4.1 Metodología seleccionada 20
4.2 Recursos necesarios 20
4.3 Selección de herramientas y técnicas 21
4.4 Supuestos 23
4.5 Restricciones 23
4.6 Riesgos y contingencias 23
4.7 Planificación del Proyecto 26
4.8 Costos y Beneficios 33
4.9 Factibilidad del Proyecto 35
4.9.1 Factibilidad legal 35
4.9.2 Factibilidad Técnica - Operativa 36
4.10 Entregables 36
5. Implementación de la metodología CRISP-DM 38
5.1 Comprensión del negocio 38
5.2 Comprensión de los datos 38
5.2.1 Recolección inicial de imágenes y propiedades de las imágenes 38
5.2.2 Descripción de los datos 43

5.2.3 Exploración de los datos 46
5.2.4 Verificación de la calidad de los datos 53
5.3 Preparación de los datos 53
5.3.1 Selección de los datos 53
5.3.2 Integración de los datos 54
5.3.3 Limpieza de datos 55
5.3.4 Vista minable 59
5.4 Modelado 59
5.4.1 Selección de técnicas de modelado 60
5.4.2 Diseño de las pruebas 60
5.4.3 Desarrollo del modelo 62
5.5 Evaluación 68
5.5.1 Evaluación de los resultados de la minería de datos con el criterio de éxito 68
5.5.2 Determinación de las siguientes acciones 69
5.6 Desarrollo 69
5.6.1 Plan de desarrollo 70
5.6.2 Control y mantenimiento 71
5.6.3 Prototipo 71
6. Conclusiones y futuras líneas de investigación 72
7. Bibliografía 74
Anexo 75
A. Metodología CRISP-DM 75
B. Clases seleccionadas 79
C. Vista minable 81
D. *Class recall* y *class presicion* para los modelos obtenidos 82
E. Árbol de decisión 86
F. Prototipo 92

Índice de imágenes

Imagen 2-1 Proceso de extracción de conocimiento. 9
Imagen 2-2 Encuesta sobre el uso de metodologías para proyectos de minería datos. 11
Imagen 2-3 Diagrama del proceso SEMMA. 12
Imagen 2-4 Ciclo de vida del modelo CRISP-DM. 13
Imagen 3-1 Diagrama de la estructura organizacional de la empresa 16
Imagen 3-2 Proceso de carga de digitalizaciones de legajos de clientes 17
Imagen 3-3 Interfaz de aplicación de consulta de documentos de legajos de clientes 18
Imagen 4-1 Estructura de desglose de trabajo 26
Imagen 4-2 Diagrama Gantt de la planificación del proyecto 32
Imagen 5-1 Imagen original y resultantes de la aplicación de transformaciones 42
Imagen 5-2 Diagrama de base de datos para el almacenamiento de los datos obtenidos 44
Imagen 5-3 Diagrama de caja para los atributos Int_0 e Int_255 de la tabla Sobel 47
Imagen 5-4 Diagrama de caja para los atributos Int_0 e Int_255 de la tabla DTF 48
Imagen 5-5 Diagrama de caja para el atributo Int_0 de la tabla Integral 49
Imagen 5-6 Diagrama de caja para el atributo Int_0 de la tabla Integral 50
Imagen 5-7 Diagrama de caja para el atributo Int_0 de la tabla Smooth 51
Imagen 5-8 Diagrama de caja para el atributo Int_255 de la tabla *Smooth* 52
Imagen 5-9 Operador *K-NN Global Anomaly* para detección de anomalías 55
Imagen 5-10 Parametrización del operador KNN Global Anomaly Score 56
Imagen 5-11 Diagrama de caja para puntajes de anamalía de ejemplos 57
Imagen 5-12 Operador *Filter Examples* para exclusión de anomalías 58
Imagen 5-13 Parametrización del módulo *Filter Examples* 58
Imagen 5-14 Operador *Cross Validation* en RapidMiner 60
Imagen 5-15 Parametrización del operador *Cross Validation* 61
Imagen 5-16 Proceso *Cross Validation* en RapidMiner 62
Imagen 5-17 Proceso de preparación de datos, modelado y validación en RapidMiner 63
Imagen 5-18 Proceso *Cross Validation* en RapidMiner 63
Imagen 5-19 Parametrización del operador *Naive Bayes* en RapidMiner 64
Imagen 5-20 Parámetros para el operador Decision Tree en RapidMiner 65
Imagen 5-21 Parámetros para el operador *Random Tree* en RapidMiner 66
Imagen 5-22 Parámetros para el operador *Random Tree* en RapidMiner 67
Imagen 5-23 Parámetros para el operador k-NN en RapidMiner 67
Imagen 5-24 Ejemplo de reglas del modelo de clasificación de imágenes 70
Imagen 5-25 Proceso de importación de documentos 70
Imagen F-1 Interfaz web del prototipo de clasificación de imágenes 92
Imagen F-4 Resultados obtenidos del proceso de clasificación 95
Imagen F-5 Resultados obtenidos del proceso de clasificación 95
Imagen F-6 Resultados obtenidos del proceso de clasificación 96
Imagen F-7 Resultados obtenidos del proceso de clasificación 97
Imagen F-8 Resultados obtenidos del proceso de clasificación 98

Índice de tablas

Tabla 4-1 Riesgos asociados al proyecto....23
Tabla 4-2 Estructura de desglose de trabajo....31
Tabla 4-3 Costos de licencias de software....33
Tabla 4-4 Costo de hardware....34
Tabla 4-5 Costos de recursos humanos....34
Tabla 4-6 Costo total de la ejecución del proyecto....34
Tabla 5-1 Cantidad de imágenes por clases....40
Tabla 5-2 Descripción de atributos de la tabla Altoancho....44
Tabla 5-3 Descripción de atributos de la tabla Sobel....45
Tabla 5-4 Descripción de atributos de la tabla DTF....45
Tabla 5-5 Descripción de atributos de la tabla Integral....45
Tabla 5-6 Descripción de los atributos de la tabla Smooth....45
Tabla 5-7 Atributos de tabla correspondiente a los datos integrados.....54
Tabla 5-8 Valores de parámetros del diagrama de cajas....57
Tabla 5-9 Atributos de la vista minable....59
Tabla 5-10 Niveles de precisión de los modelos obtenidos....68
Tabla B-1 Clases conservadas luego del proceso de selección y limpieza de datos....80
Tabla B-2 Fragmento de la vista minable resultante de la fase de preparación de los datos.....81
Tabla D-3 Valor de *class precision* para cada clase de cada uno de los modelos obtenidos....83
Tabla D-4 Valor de *class recall* para cada clase de cada uno de los modelos obtenidos....85

Abstract

Se realiza la aplicación de minería de imágenes a un conjunto de imágenes con el objetivo de obtener un modelo que permita realizar su clasificación de forma automática. Para esto se combinan técnicas de minería de datos y procesamiento de imágenes. La clasificación se basa en agruparlas según características similares obtenidas a partir del análisis y del procesamiento de imágenes. Al tratase de un conjunto de digitalizaciones de legajos de clientes, aplicar este proceso permite conocer el tipo de documento asociado a la imagen, lo que agiliza el proceso de búsqueda. Poder conocer las cantidades por tipos de documentos permite analizar la necesidad de conservar o descartar un documento físico, y así disminuir su volumen lo que implica una disminución de costos de almacenamiento.

1. Introducción

Los diversos procesos que se llevan a cabo en las empresas generan una gran cantidad de documentos que los respaldan, por ejemplo, el alta o baja de un cliente, la entrega de un equipo, la recepción de materiales, contratos, etc. Estos documentos contienen información histórica y los mismos son resguardados por diversos motivos, que pueden ser requerimientos legales, de auditoría, necesidad de consultar la información histórica, entre otros.

Una compañía de seguros genera diariamente un gran volumen de documentos físicos derivados de los distintos procesos de negocio. Estos son conservados para poder acceder a la información que ellos contienen y por exigencias legales. En la empresa, cada cliente posee un legajo, el cual contiene documentos de distintos tipos, por ejemplo, formularios de solicitudes, notas, documentos de identidad, acuses de recibos.

Debido a que la cantidad de documentos es significativa y son consultados frecuentemente por distintas personas, se buscó la forma de mejorar este proceso. El objetivo era evitar que los empleados se desplazasen físicamente al lugar donde se encontraban almacenados todos los legajos de los clientes para realizar la búsqueda y consulta de forma manual. Además, un problema que se presentaba es que, al retirar uno o varios de ellos, quedaban en posesión de la persona que los consultaba y no podían ser consultados por nadie más. La solución se basó en la digitalización de los documentos en guarda y de los que ingresan diariamente. Las digitalizaciones son almacenadas en un repositorio y se cuenta con un sistema informático propio que permite a los usuarios realizar la consulta de las mismas, sin la necesidad de desplazarse de sus puestos de trabajo y contando con la posibilidad realizar esta tarea simultáneamente. Debido a la gran cantidad de clientes, la cual se encuentra en constante crecimiento, el repositorio cuenta actualmente con más de dos millones de digitalizaciones.

Cuando un usuario realiza la consulta de los documentos del legajo de un cliente, obtiene como resultado un listado de todas las digitalizaciones asociadas al cliente. La única forma de poder identificar el tipo, es abrir la imagen y realizar su inspección visual. Disponer de la información que identifique el tipo de documento al cual corresponde, permitiría agilizar el proceso de búsqueda, ya que, si buscamos un documento en particular, nuestra búsqueda estará acotada al conjunto de imágenes que pertenecen a ese tipo. Adicionalmente, conocer qué tipos de documentos se tienen actualmente en guarda, resultará de utilidad para desechar aquellos que no necesitan ser conservados, lo que permitirá reducir el costo de almacenamiento físico. Debido a que los documentos reflejan los procesos de negocio, el análisis de las cantidades y tipos de los mismos, podrían aportar información de utilidad.

Para contar con una etiqueta o campo que indique el tipo de documento de cada imagen es necesario realizar el proceso de clasificación de las mismas. La tarea de clasificación consiste en agrupar elementos que poseen características similares y asignarlos a una determinada clase. Clasificar de forma manual todas las digitalizaciones de los legajos de los clientes existentes y de las que ingresan diariamente consumiría una gran cantidad de tiempo y recursos. Es por este motivo que se busca una solución a este problema, mediante el uso de herramientas proporcionadas por las áreas de minería de datos y el procesamiento de imágenes. El objetivo

será obtener un modelo clasificador de imágenes que luego pueda ser empleado para el desarrollo de una herramienta de clasificación automática que se integre a los sistemas existentes en la empresa.

Las características de una imagen pueden expresarse mediante atributos y sus valores. Por ejemplo, el ancho, el alto, la cantidad de píxeles de un determinado color, entre otros. Mediante el procesamiento de imágenes es posible transformarlas o modificarlas para ocultar o resaltar determinadas características. Si realizamos la extracción de los valores de los atributos que describen a las imágenes (tanto originales como procesadas) obtendremos una gran cantidad de datos. Para analizar estos datos en búsqueda de conocimientos empleamos las herramientas que proporciona la minería de datos. En este caso en particular, el objetivo es la obtención de un modelo que permita predecir a qué clase pertenece una determinada imagen según sus características.

El proceso presenta varios desafíos, uno de ellos es determinar qué atributos son los que se seleccionarán para describir a las imágenes. Es necesario analizar qué transformaciones son las adecuadas para aplicar a las imágenes resultantes de digitalizaciones de los documentos. También se deben definir las propiedades necesarias para diferenciar las imágenes y poder agrupar aquellas que son similares. Imágenes que poseen características similares, pertenecerán a la misma clase. Otro desafío es que los valores de las propiedades definidas se puedan obtener de forma automatizada.

La clasificación de la totalidad de las imágenes permitirá obtener el atributo que identifica a qué categoría pertenece cada uno de los documentos. Esta información deberá ser integrada al sistema de consultas de legajos de clientes que posee la empresa para que pueda ser empleada por los usuarios. El proceso de clasificación deberá realizarse en dos etapas. La primera consistiría en clasificar todas las digitalizaciones existentes en el repositorio de legajos de clientes. Una segunda etapa debe contemplar la integración del proceso de clasificación con el proceso de importación de las digitaciones al sistema de consulta de legajos. De esta forma la información de la clase de una digitalización estará disponible tanto para las históricas como para las que ingresen diariamente.

2. Estado de la cuestión

Construir un modelo clasificador de imágenes requiere del uso en conjunto de conocimientos y recursos proporcionados por las áreas de minería de datos y del procesamiento de imágenes. Esta última aporta las herramientas necesarias para realizar transformaciones que permitan obtener información adicional que no se muestra en las imágenes originales. La minería de datos permite realizar la obtención del conocimiento del gran volumen de datos resultante de la extracción de valores de atributos que describen a las imágenes originales y procesadas.

2.1 Minería de datos

"La minería de datos es un proceso de negocio para explorar un gran volumen de datos para descubrir patrones y reglas significativas." (Linoff & Berry, 2011) La tecnología facilita la recolección y almacenamiento de grandes volúmenes de datos. Estos conjuntos de datos pueden contener información valiosa, pero para obtenerla es necesario realizar el análisis y procesamiento de ellos, de modo que sea posible obtener conocimiento.

La extracción del conocimiento es un proceso que consta de cinco fases:

- Integración y recopilación
- Selección
- Limpieza y transformación
- Minería de datos, evaluación e interpretación
- Difusión y uso

En la Imagen 2-1 se muestra un diagrama del proceso. La fase de **integración y recopilación** es la fase inicial, en ella se reúnen todos los datos necesarios. Si los datos provienen de más de una fuente, estos deberán ser unificados. También, se debe contemplar la posibilidad de que los datos sobre los cuales se necesita trabajar deban ser generados ya que no fueron recopilados previamente. La fase siguiente es la de **selección, limpieza y transformación**, donde se seleccionan los datos sobre los cuales se trabajará descartando aquellos que no sean útiles. También durante esta etapa se analizan los datos en búsqueda de valores erróneos o faltantes y se toman decisiones sobre las acciones a seguir en estos casos. Puede ser necesario transformar los datos, por ejemplo, si tienen atributos cuyos valores son continuos, pueden ser convertidos a discretos asignándoles un valor según el rango al que pertenecen. Una vez finalizadas estas tareas, obtenemos una vista minable, es decir, los datos se encuentran preparados para ser usados en la siguiente fase. En la fase de **minería de datos** buscamos obtener un modelo a partir del conjunto de datos previamente definido. En este punto se determina qué tipo de tarea de minería de datos, modelo y algoritmo se emplearán. (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004)

Imagen 2-1 Proceso de extracción de conocimiento.

Existen diversos tipos de tareas de minería de datos. Estas se agrupan dependiendo del tipo de problema que tratan de resolver. Entre ellas encontramos tareas de clasificación, segmentación, asociación, regresión, pronóstico, análisis de secuencia, análisis de desviación, entre otras. (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004)

Las tareas de minería de datos pueden tener como objetivo predecir un valor dependiendo de las características de los mismos. Entre estos tipos de tareas se encuentran el agrupamiento, la regresión y la clasificación.

El agrupamiento consiste en definir grupos de acuerdo a características similares de los datos. En la tarea de clasificación se define un atributo o conjunto de atributos predecibles que permiten identificar un caso en particular con una categoría. La segmentación consiste en separar el conjunto de datos en subconjuntos según atributos similares. La tarea de asociación pretende descubrir los conjuntos de atributos que se repiten frecuentemente y encontrar las reglas para esas asociaciones. La regresión "consiste en aprender una función real que asigna a cada instancia un valor real" (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004), es decir que esta tarea intenta determinar una función que modele el conjunto de datos estudiado.

El resultado de aplicar los procesos de minería de datos a un conjunto de datos, es la obtención de un modelo que permite dar solución al problema planteado inicialmente. Durante el proceso de minería de datos es posible obtener varios modelos, de los cuáles se deberá elegir uno teniendo en cuenta el grado de precisión que ofrece, el costo de procesamiento y la complejidad de implementación. Esta tarea se lleva a cabo en la siguiente fase, **evaluación e interpretación**. El objetivo es analizar los modelos obtenidos y seleccionar aquel que resuelva de la mejor forma el problema planteado inicialmente. Y por último se realiza la fase de **difusión y uso**, que consiste en dar a conocer el modelo a los interesados, a aquellos que evaluarán el modelo y a quienes se beneficiarán con el mismo. (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004)

2.2 Herramientas de minería de datos

Actualmente existen distintas herramientas software dedicadas a la minería de datos. Entre ellas se encuentran Weka[1], RapidMiner Studio[2], SPSS Modeler (IBM)[3], JMP[4] y SAS Visual Data Mining and Machine Learning (VDMML)[5]. A continuación, se detallan las principales características de cada una:

- **Weka:** es una herramienta libre para minería de datos. Posee una interfaz de usuario gráfica, así como también puede ser accedida mediante una interfaz de línea de comandos o una interfaz de programación de aplicaciones de Java. Weka permite realizar la importación de datos, el entrenamiento del modelo y su evaluación, sin necesidad de codificar. También se provee documentación sobre los fundamentos de minería de datos vinculados a la herramienta y capacitaciones gratuitas sobre su uso.

- **RapidMiner** Studio: su principal característica es que posee una interfaz gráfica intuitiva para realizar minería de datos. RapidMiner brinda las herramientas necesarias para cubrir todo el proceso. Permite el acceso a los datos, explorarlos, prepararlos, realizar el modelado y la validación del modelo obtenido. Los procesos son implementados mediante bloques que realizan tareas específicas. Los bloques son parametrizados según las necesidades del usuario.

- **SPSS Modeler:** esta herramienta no solo se enfoca en la minería de datos, si no que ofrece otras herramientas interesantes como ser minería de texto y análisis de redes sociales. El diseño de la herramienta sigue los lineamientos de la metodología de minería de datos CRISP-DM. La interfaz gráfica se basa en el uso de bloques configurables y flujos que conectan los bloques. Esta característica la hace accesible para usuarios que no poseen conocimientos en el área de la programación.

- **JMP:** este software está centrado principalmente en el análisis estadístico de los datos, pero también ofrece la posibilidad de crear modelos predictivos. Presenta un interfaz de usuario gráfica y ofrece la posibilidad de generar automáticamente el código de los modelos obtenidos, en el lenguaje de programación deseado para ser incorporado a soluciones existentes.

- **SAS Visual Data Mining and Machine Learning (VDMML):** esta herramienta software provee asistencia para todo el proceso de minería de datos mediante una interfaz visual y da la posibilidad de usar programación. Permite la generación automática del código para cada paso del proceso. Posee herramientas integradas para realizar el análisis de texto en diversos idiomas. Da la posibilidad a los usuarios de trabajar de forma colaborativa.

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] https://rapidminer.com/
[3] http://www-03.ibm.com/software/products/es/spss-modeler
[4] http://www.jmp.com/es/
[5] https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html

2.3 Metodologías de minería de datos

CRISP-DM y SEMMA son metodologías para realizar el proceso de minería de datos. Según una encuesta realizada por KDnuggets[6] en octubre de 2014, se encuentran entre las tres primeras metodologías más usadas para aplicar el proceso de descubrimiento de conocimiento. La más empleada es CRISP-DM, seguida de la elección de aplicar una metodología propia no estandarizada (en inglés *my own*) y en tercer lugar se encuentra SEMMA. En la Imagen 2-2 se puede observar el resultado de la encuesta realizada para determinar cuáles son las más usadas en proyectos de minería de datos (Piatetsky-Shapiro & Mayo, 2014).

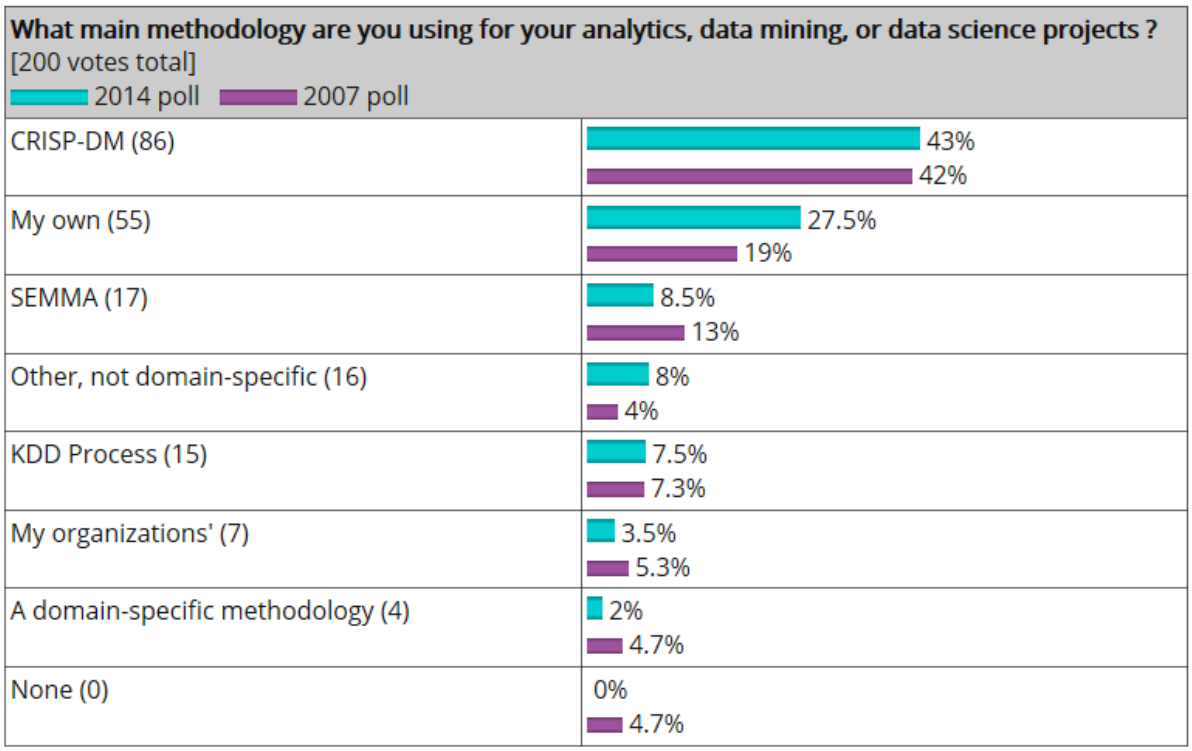


Imagen 2-2 Encuesta sobre el uso de metodologías para proyectos de minería datos. Fuente: https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

SEMMA es una metodología propuesta por la empresa SAS[7]. SEMMA son las siglas para *Sample Explore Modify Model Asses*, en español, muestreo, exploración, modificación, modelado y evaluación. Estas son cada una de las fases que se ejecutan en el proceso del descubrimiento del conocimiento. En la Imagen 2-3 se puede observar un diagrama del proceso.

[6] https://www.kdnuggets.com/about/index.html. KDNuggets es un sitio líder reconocido por la comunidad científica en los siguientes temas: inteligencia artificial, big data, minería de datos, ciencia de datos y *machine learning*,

[7] https://www.sas.com/en_us/home.html

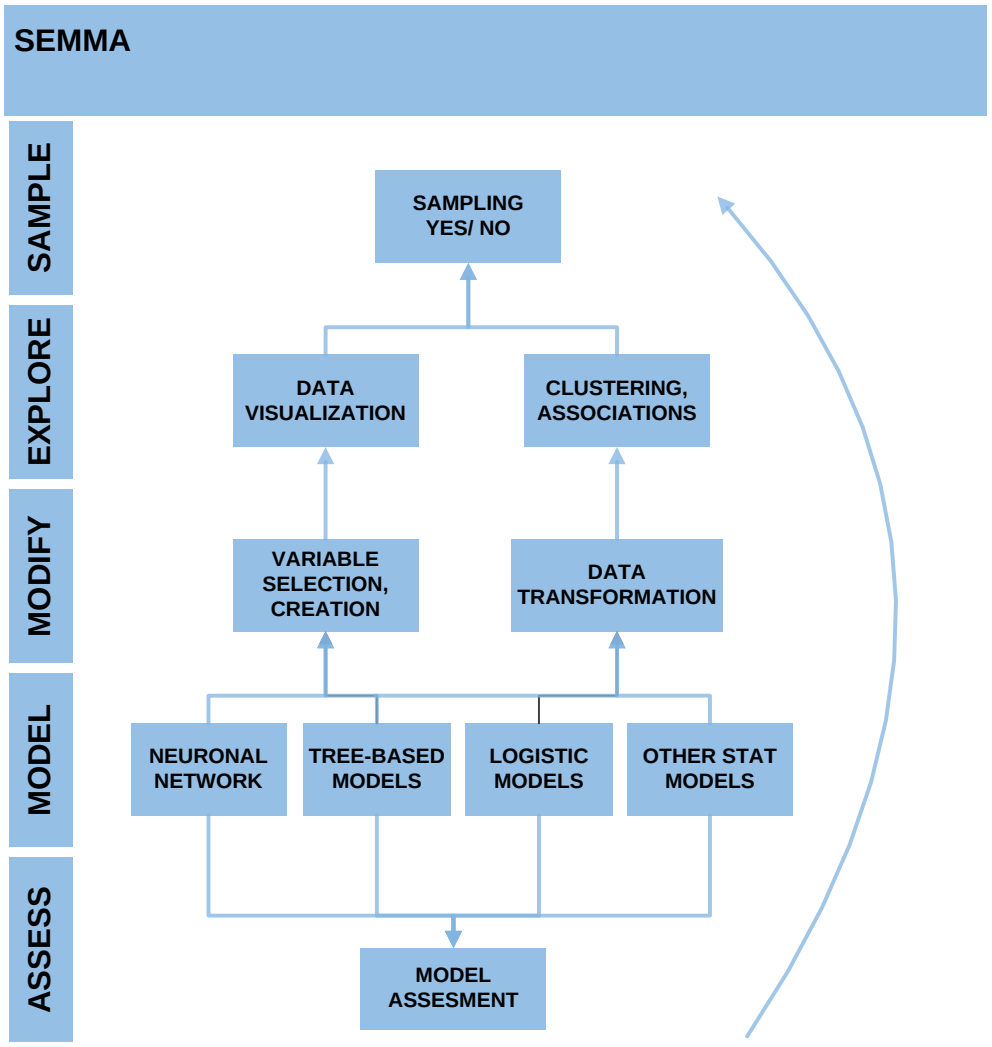


Imagen 2-3 Diagrama del proceso SEMMA. Fuente: https://documentation.sas.com

La primera fase, consiste en seleccionar una muestra de forma estadística que logre la mejor representación de la totalidad de los datos. Trabajar con una muestra representativa y no con todo el conjunto de datos permitirá agilizar el tiempo de procesamiento. El objetivo de la fase de exploración es, mediante el análisis de los datos, determinar si existen patrones que puedan generar reglas. La fase de exploración permite conocer los datos con los que se cuentan. Es posible que surja la necesidad de eliminar, agregar o cambiar datos, esto se realiza en la etapa de Modificación. Una vez concluida la etapa de modificación se cuenta con un conjunto de datos limpio y completo para pasar al Modelado. En el modelado se realiza la aplicación de las distintas técnicas de minería de datos y se busca definir un modelo que resuelva el problema planteado inicialmente. Una vez obtenido el modelo, es necesario evaluarlo para determinar la precisión de sus resultados (Olson & Delen, 2008). Esta metodología se encuentra estrechamente ligada a la herramienta SAS Enterprise Miner, que permite la implementación de cada fase de SEMMA mediante nodos.

El modelo de referencia CRISP-DM contempla el entendimiento del negocio, el entendimiento y preparación de los datos, el modelado, la evaluación y por último el desarrollo. Estas son cada una de las fases que componen el ciclo de vida de un proyecto de minería de

datos (Imagen 2-4). Estas fases se encuentran relacionadas entre sí y es posible que durante el desarrollo del proyecto sea necesario retroceder a una fase para brindar retroalimentación entre las mismas. Por cada una de las fases se definen tareas genéricas que guían los pasos a seguir durante el proyecto.

La primera fase, el **entendimiento del negocio**, consiste en definir objetivos, alcances, requerimientos dentro de la perspectiva del negocio.

Las fases de **entendimiento y preparación de los datos**, abarcan las actividades iniciales de recolección y análisis de los datos y el armado del conjunto de datos sobre el cual se trabajará.

En la **fase de modelado**, se aplican los modelos seleccionados de acuerdo a las características particulares del problema.

Una vez obtenidos los modelos, se pasa a la **fase de evaluación** para asegurar que se cumplen los objetivos planteados inicialmente.

Y, por último, se lleva a cabo la **etapa de desarrollo** que consiste la aplicación del conocimiento obtenido, ya sea informando los resultados o implementando un proceso repetible de minería. (Chapman, y otros, 2000)

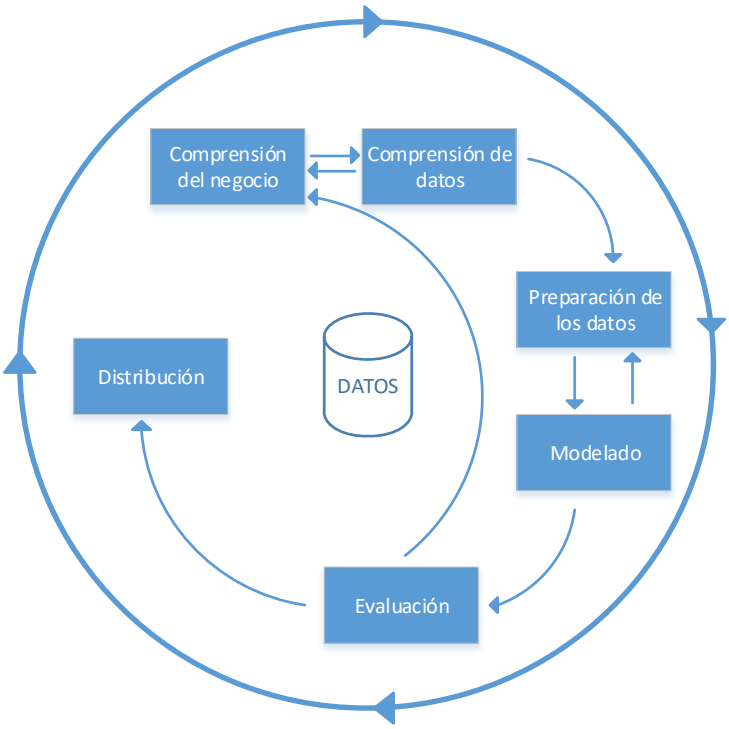


Imagen 2-4 Ciclo de vida del modelo CRISP-DM. Fuente: CRISP-DM 1.0 Step-by-step data mining guide

2.4 Procesamiento de imágenes

Una imagen digital puede ser considerada como una representación discreta de datos que contiene información espacial (distribución) y de intensidad (color). (Solomon & Breckon, 2011) Una imagen digital de dos dimensiones es la representación de la respuesta de un sensor (por ejemplo, un scanner) expresada en un eje de coordenadas cartesianas bidimensional.

Una imagen bidimensional está conformada por píxeles dispuestos en forma de una matriz, los cuales al variar su intensidad componen la imagen. Cada pixel tiene una ubicación definida dentro de la matriz que puede denotarse por un par de coordenadas cartesianas. Además, los mismos tienen información asociada sobre su intensidad representada por un valor numérico.

Un espacio de color es la representación usada para almacenar la información de color. Este puede estar constituido por más de un canal que indica cuánto de un determinado color se aporta para obtener el color de un pixel. El espacio RGB permite representar mediante un vector de tres dimensiones el color de un pixel. RGB son las siglas para Red (rojo), Green (verde) y Blue (azul). Si los tres valores del vector son iguales, se obtiene un color en la escala de grises.

El procesamiento de imágenes consiste en la modificación de la imagen mediante operadores matemáticos. Modificar una imagen permite eliminar datos porque generan ruido, es decir, distorsión de la imagen, también es posible acentuar determinados detalles, combinar imágenes, agregar filtros. El objetivo de procesar las imágenes es ocultar determinadas características y destacar otras, lo que permite analizar si las imágenes que poseen similares características, pertenecen a una misma categoría. Los dos principales operadores morfológicos empleados para realizar el procesamiento de las imágenes son la dilatación y erosión. La primera consiste en el proceso de ampliar y engrosar regiones angostas, aumentando los bordes. La erosión, produce el efecto contrario, reduce pequeñas características aisladas. Combinando y aplicando sucesivamente estos operadores se pueden definir otros operadores (Solomon & Breckon, 2011). Otra técnica que puede ser usada es calcular la integral de una imagen. Como resultado obtenemos una nueva imagen, en la cual el valor de la intensidad de cada píxel está determinado por la suma de las intensidades de los píxeles en línea vertical por arriba y en línea horizontal por la izquierda. (Viola & Jones, 2001). El cálculo de la transformada de Fourier de una imagen da como resultado una imagen en el dominio de las frecuencias.

Una imagen digital puede contener otro tipo de información asociada a ella que puede ser útil para realizar la clasificación de la misma, como ser dimensión, tamaño, dispositivo con la cual fue capturada o fecha de creación.

2.5 Minería de imágenes

La aplicación en conjunto de los conocimientos de las áreas de minería de datos y el procesamiento de imágenes nos permite la realización de tareas de clasificación de imágenes. La minería de datos permite realizar el análisis de grandes volúmenes de datos con el objetivo de encontrar patrones. En este caso los datos a ser analizados son aquellos que describen las características de las imágenes. Estos datos deben ser extraídos de las imágenes, tanto de las originales como de aquellas resultantes de la aplicación de técnicas de procesamiento de imágenes. También es importante elegir adecuadamente las características de las imágenes que

permitan asociarlas a una clase o tipo y que las diferencie de aquellas pertenecientes a otras clases. (Yousofi, Esmaeili, & Sada, 2016)

2.6 Aplicación de minería de imágenes

En la actualidad la minería de imágenes es aplicada en áreas como la medicina, la agricultura, o los procesos industriales. En la medicina, se realizan análisis sobre imágenes como ser radiografías, ecografías, tomografías con el objeto de poder realizar diagnósticos. Un caso en particular es el uso de imágenes de lesiones en la piel o melanomas de seres humanos con el objeto de predecir tipos de melanomas para evitar realizar biopsias (Coelho, Gonçalves, Portela, & Santos, 2019). Otro caso es la aplicación de minería de imágenes en el control de los sistemas de fijación de rieles en vías de trenes, para optimizar el proceso de inspección de seguridad de los mismos. (Song, Guo, Jiang, & Li, 2019)

Existen programas como perClass[8] que fue diseñado para realizar la clasificación de imágenes basándose en imágenes de muestras que utiliza para aprender. Este programa permite ser usado para el análisis y la clasificación de defectos en la comida, para detectar presencia de minerales en rocas, el reconocimiento de género en imágenes de rostros, entre otras aplicaciones. CVB Manto[9], es otro programa que ofrece la posibilidad de realizar la clasificación de imágenes sin la necesidad de indicarle cuál es el patrón de clasificación. Durante una fase de entrenamiento el programa determina los criterios para la definición de las clases. IBM propone una herramienta que "reconoce automáticamente categorías semánticas para contenido visual diverso". Esta herramienta pretende clasificar a las imágenes según una temática determinada, por ejemplo, identificar categorías como rostros de personas, personas con anteojos, parejas, equipos o soldados.

[8] http://perclass.com/
[9] http://www.stemmer-imaging.co.uk/en/products/series/cvb-manto/

3. Definición del problema

La digitalización masiva de documentos, permite obtener la imagen digital de documentos físicos para posibilitar su consulta de forma ágil y sencilla, mediante una aplicación que brinda una interfaz para tal fin. Cada una de las digitalizaciones contiene información asociada de utilidad para su identificación, pero carece del tipo de documento del que se trata. Debido a esto, se cuenta con una gran base de imágenes en continuo crecimiento, sin poder determinar a qué tipo pertenece cada documento, sin previamente visualizar el mismo.

3.1 Descripción de la empresa

La empresa sobre la cual se trabajará es una compañía dedicada al rubro de los seguros. Se encuentra en la categoría de medianas y pequeñas empresas. Está constituida por 5 gerencias, la gerencia general, de la cual dependen la gerencia de sistemas, gerencia técnica y de operaciones, la gerencia de administración y finanzas y la gerencia comercial (Imagen 3-1). El desarrollo del proyecto involucrará la gerencia de sistemas, en particular el área de desarrollo y el área de suscripción perteneciente a la gerencia técnica y de operaciones.

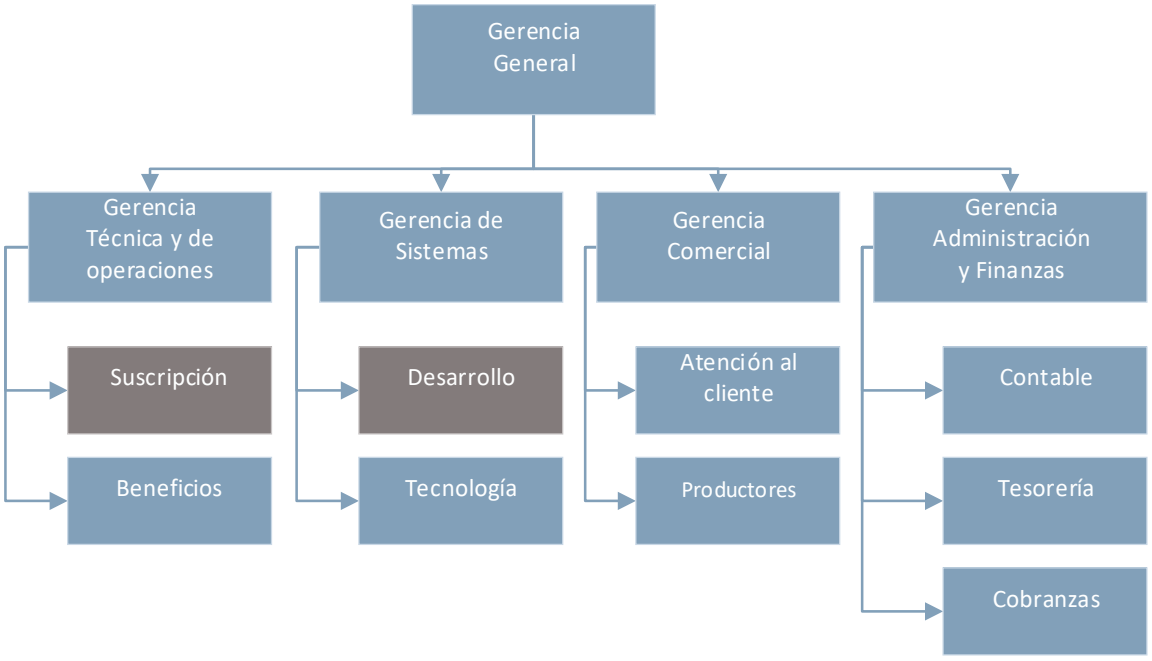


Imagen 3-1 Diagrama de la estructura organizacional de la empresa

El área de suscripción genera diariamente documentos físicos. Cada día estos documentos son enviados al proveedor encargado de prestar el servicio de guarda y digitalización. Como resultado se obtienen imágenes en escalas de grises que son guardadas como PDF. Estos documentos PDF son subidos a un servidor de archivos. Luego de ser escaneados, los documentos físicos son almacenados por el proveedor del servicio de guarda. Un proceso automático, que se ejecuta diariamente, copia los archivos desde el servidor del proveedor a un servidor local. Luego, se inicia la siguiente etapa del proceso que consiste en mover cada uno de los archivos a una ubicación determinada dentro del servidor de archivos. Por cada archivo se genera un registro en una base de datos que permite asociar su ubicación con un identificador que indica a qué cliente pertenece cada documento. Una vez finalizado este proceso, las imágenes se encuentran disponibles para su consulta a través de una aplicación propia. El esquema de este proceso se encuentra en la Imagen 3-2.

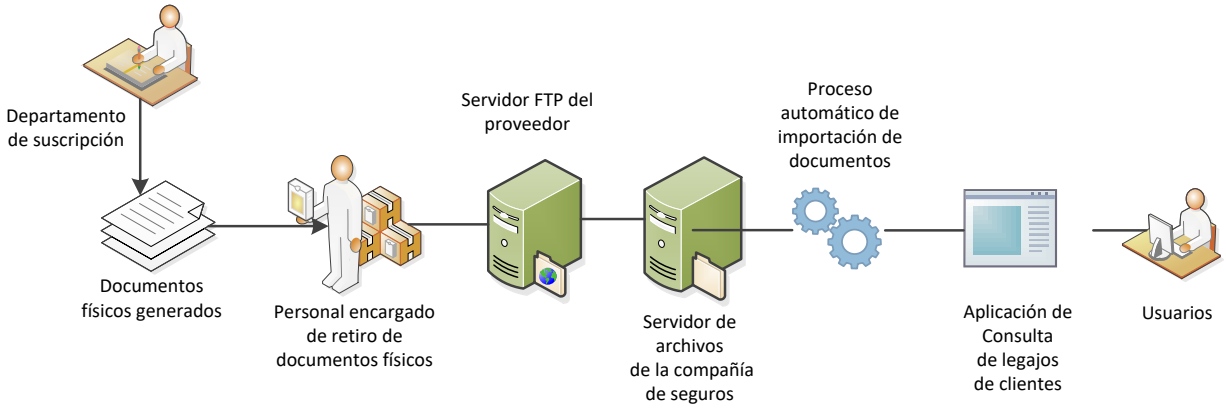


Imagen 3-2 Proceso de carga de digitalizaciones de legajos de clientes

Para realizar la búsqueda de una imagen en esta aplicación, el usuario debe ingresar el número de cliente. Este dato es validado y, en caso de ser correcto, se muestra la información del cliente junto con el listado de documentos PDF que tiene asociados. Cada PDF puede contener una o más imágenes (una imagen por página). Al hacer clic en el nombre de cada documento, el contenido de este se muestra en el panel de la derecha. En la Imagen 3-3 se muestra un esquema de la interfaz de consulta de documentos de los legajos de clientes (los nombres de las imágenes y el contenido de la imagen fueron distorsionados, los datos del cliente son ficticios a modo de mantener la confidencialidad).

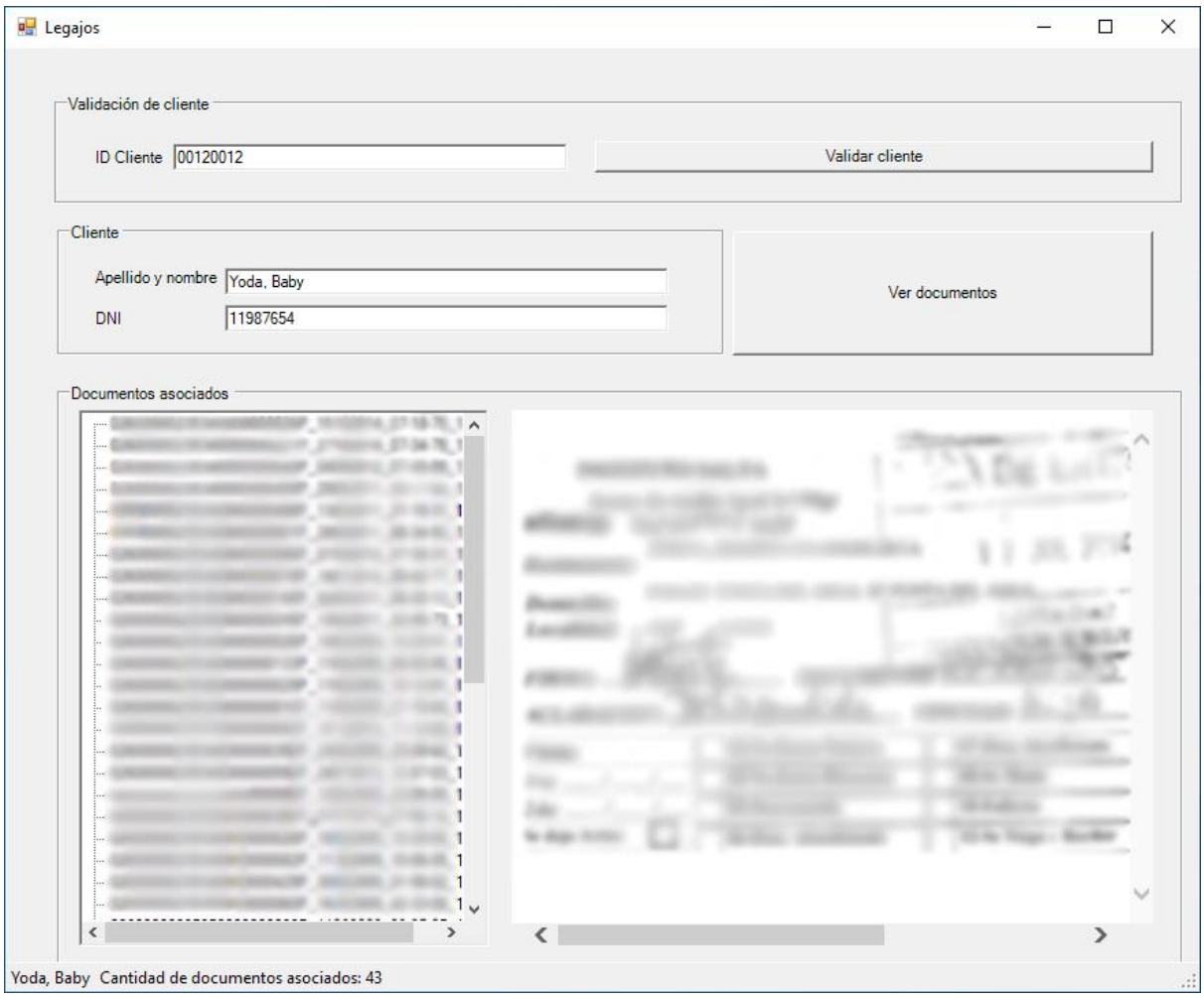


Imagen 3-3 Interfaz de aplicación de consulta de documentos de legajos de clientes

El listado solo muestra el nombre con el que fue guardado el archivo, que consiste en una secuencia de caracteres alfanuméricos. Este nombre no da ninguna información que permita conocer de qué clase es el documento. Cuando es necesario buscar un tipo de documento en particular, se debe recorrer la lista y realizar la inspección visual de cada uno de ellos.

La búsqueda de los documentos podría agilizarse si se contara con la información de la clase a la que pertenece cada uno de ellos. Esto permitiría añadir un filtro por clase a la búsqueda, lo cual llevaría a la reducción de la cantidad de documentos devueltos por la interfaz y solo se contaría con los documentos de interés.

3.2 Objetivos del negocio

El objetivo principal es reducir el tiempo empleado en búsquedas de documentos. Esto permitirá agilizar las tareas operativas pudiendo dedicar el tiempo a tareas de análisis. Conocer la clase de cada documento permitirá realizar análisis sobre el tipo de documentos con el que se cuenta y los volúmenes del mismo. Esta información será de utilidad para tomar decisiones a nivel estratégico.

3.3 Objetivo general de la solución

Obtener un modelo clasificador capaz de clasificar imágenes correspondientes a digitalizaciones de documentos de legajos de clientes de una compañía.

3.4 Objetivos específicos de la solución

Definir las propiedades que se emplearán en el análisis de las imágenes que permitan distinguir una imagen de otra.

Categorizar las imágenes pertenecientes al conjunto de imágenes tomado como muestra.

Obtener un modelo que permita la clasificación de imágenes según las clases definidas con al menos un 80% de precisión.

Clasificar un lote de imágenes de acuerdo a las clases definidas.

3.5 Objetivos de la minería de datos

Obtener un modelo que permita realizar la clasificación de un lote de imágenes correspondientes a digitalizaciones de legajos de clientes.

3.6 Criterio de éxito de la minería de datos

El modelo propuesto debe realizar la clasificación de las imágenes con al menos un 80% de certeza.

3.7 Alcance

El desarrollo del proyecto consiste en la obtención de un modelo que permita la clasificación de imágenes obtenidas del proceso de digitalización de legajos de clientes. El modelo será entregado al área de desarrollo, la cual quedará a cargo del proceso de implementación e integración del modelo al software ya existente, quedando estos pasos fuera del alcance de este proyecto.

Con el objetivo de mostrar el funcionamiento del modelo obtenido se desarrollará un prototipo que permita mostrar cómo se realiza la tarea de clasificación. El prototipo tomará un subconjunto ejemplos correspondientes a imágenes seleccionadas aleatoriamente sin etiquetar y devolverá como resultado la etiqueta que identifica la clase a la cual pertenece la imagen. El objetivo del prototipo está acotado a mostrar el funcionamiento del modelo obtenido, pero no la integración al sistema de consulta de legajos existentes.

4. Solución propuesta

4.1 Metodología seleccionada

La metodología que guiará el desarrollo del proyecto será CRISP-DM. Debido a que la metodología SEMMA se encuentra asociada estrechamente a una herramienta software, se prefiere CRISP-DM que es independiente del software empleado para realizar el proceso de obtención del conocimiento. CRISP-DM abarca desde el planteo de objetivos hasta la implementación de la solución. Contempla el entendimiento del negocio, el entendimiento y preparación de los datos, el modelado, la evaluación y por último el desarrollo (Chapman, y otros, 2000). Si bien la metodología se aplica a minería de datos, la misma puede ser adaptada a minería de imágenes, ya que se trabajará con los datos extraídos de las imágenes, que reflejan sus características y estos conformarán una gran base de datos. Un resumen de la metodología se encuentra en el anexo A.

4.2 Recursos necesarios

Los recursos necesarios para llevar a cabo el proyecto son los detallados a continuación:

Datos

- **Conjunto de imágenes:** Las imágenes con las cuales se trabaja son una muestra obtenida del repositorio de digitalizaciones de legajos de clientes que cuenta con más de un millón y medio de imágenes. Las imágenes se procesan para obtener los atributos que permiten su clasificación, ya que esos datos no fueron recopilados previamente.

Software

- **Sistema administrador de base de datos:** El DBMS es de utilidad para la carga, organización y extracción de los datos, es decir, sirve de estructura de soporte para los datos.

- **Software para minería de datos:** Esta herramienta da soporte a los procesos de preparación de los datos, obtención y evaluación del modelo.

- **Software de procesamiento de imágenes:** Es necesario realizar el procesamiento de las imágenes para poder extraer datos de ellas, sobre los cuales se aplicarán técnicas de minería con el objeto de encontrar patrones.

Hardware

- **PC de escritorio:** sus características estarán definidas por los requerimientos del software con el cual se trabajará.

Recursos Humanos

- **Líder del proyecto**: Es la persona responsable de la planificación del proyecto, asignación de recursos, control del avance del proyecto. Es el encargado de ser el nexo entre el cliente y el analista. Debe coordinar los esfuerzos de los miembros del equipo de modo que se cumplan los objetivos del proyecto en el tiempo previsto y con los recursos asignados.

- **Analista:** Su función es realizar el análisis de la situación y el diseño de la solución. Participa en tareas de soporte al analista de datos. Será el encargado de realizar el procesamiento de las imágenes, tareas de extracción de datos, realización de copias de respaldo y ejecutar tareas definidas por el analista de datos.

- **Analista de datos:** Es el encargado de realizar el análisis e interpretación de los datos, aplicar técnicas y herramientas de minería de datos y realizar el seguimiento y evaluación del proceso de minería de datos.

4.3 Selección de herramientas y técnicas

Para llevar a cabo el proyecto es fundamental contar con el lote de imágenes a procesar, el software que permitirá el procesamiento de las imágenes y el hardware sobre el cual se trabajará.

Imágenes

Se debe contar con las imágenes seleccionadas como muestra, que se emplearán para realizar el entrenamiento de los modelos de minería de datos. Se debe tener en cuenta que las imágenes se encuentran contenidas en documentos PDF, por lo cual deberán ser extraídas para poder realizar su análisis y procesamiento.

Software

El software que se debe encontrar instalado y funcionando en el equipo en el cual se trabajará es el siguiente:

Sistema operativo: El sistema operativo sobre el cual se trabajará es Windows 10. Se ha probado la compatibilidad de las distintas herramientas con las cuales se trabajará durante el desarrollo del proyecto.

Librería para extracción de imágenes: Como se mencionó anteriormente las imágenes se encuentran dentro de archivos PDF por lo cual deben ser extraídas. La mejor herramienta encontrada para extraer elementos de un PDF fue JPEDAL[10] (Java Pdf Extraction Decoding Access Library). Esta librería provee diversas funciones para la manipulación de archivos en formato PDF. La principal función de interés es la extracción de imágenes de los documentos manteniéndola en su formato original y sin ningún tipo de transformación. Esta librería se encuentra enfocada en el procesamiento rápido de los documentos, optimizando el desempeño y reduciendo el uso de memoria, lo cual es sumamente beneficioso considerando la elevada cantidad de documentos con la cual se trabajará.

Otra alternativa evaluada fue la librería Docotic.PDF[11]. Esta librería se encuentra disponible para C# y .Net. Permite lectura, edición y creación de documentos PDF, también permite

[10] https://www.idrsolutions.com/jpedal/
[11] https://bitmiracle.com/pdf-library/

realizar la extracción de texto y de imágenes contenidas en los documentos PDF. El proceso de extracción de imágenes se realiza sin realizarles ningún tipo de transformación.

Entorno de desarrollo: Se deberá realizar la instalación de dos entornos de desarrollo, Eclipse y Microsoft Visual Studio 2019. Debido a que la librería que permite a extracción de las imágenes de los archivos PDF se encuentra en JAVA se empelará Eclipse [12]para el desarrollo de la primera etapa del proyecto. Para la realización de las tareas de procesamiento de imágenes y obtención de los valores de los atributos de las imágenes se desarrollarán scripts en el lenguaje vb.net. El entorno de desarrollo a emplear será Microsoft Visual Studio 2019.

Librería para procesamiento de imágenes: El procesamiento de imágenes se implementará mediante la librería Emgu CV[13]. Permite el acceso a la librería OpenCV[14] desde plataformas .Net. En este caso el lenguaje de programación que será usado es VB.net. Es compatible con la versión 2012 de Visual Studio en adelante. La librería contiene algoritmos que permite realizar tareas de procesamiento de imágenes, reconocimiento facial, identificación de objetos, análisis de videos, entre otras. Inicialmente se consideró el uso de la herramienta Image Processing Toolbox[15] de Matlab[16], que permite emplear el uso de operadores morfológicos para modificar las imágenes logrando así poder destacar ciertas características de interés u ocultar aquellas no sean de utilidad. Debido a que la librería EmguCV permite su uso en vb.net siendo este el principal lenguaje de programación en el área de desarrollo y que cubre las mismas funcionalidades que la herramienta Image Processing Toolbox, se opta por EmguCV como mejor alternativa.

Administrador de base de datos: Debido a que los motores de base de datos con los que se trabaja actualmente en la empresa son Microsoft SQL Server, se deberá instalar Microsoft SQL Server Management Studio 2017 para el acceso a la base de datos.

Herramienta de minería de datos: De las herramientas para realizar minería de datos analizadas en la sección 2.2, se optó por RapidMiner. La elección de este software se basó en que es una herramienta dedicada exclusivamente a la tarea de la minería de datos, proporcionando los elementos necesarios para cubrir todo el proceso, desde la preparación de los datos hasta la obtención del modelo y su evaluación. Ofrece una interfaz intuitiva basada en el uso de módulos, dónde cada módulo realiza una función particular y permite su parametrización. RapidMiner es de utilidad tanto, para personas con amplios conocimientos en el área de análisis de datos como así también para aquellas que se inician. También cuenta con documentación sobre el uso de la herramienta, capacitaciones y foros de ayuda de la comunidad. La versión de RapidMiner seleccionada es la Professional. Esta versión permite

[12] https://www.eclipse.org/
[13] http://www.emgu.com/wiki/index.php/Main_Page
[14] https://opencv.org/
[15] https://la.mathworks.com/products/image.html
[16] https://la.mathworks.com/products/matlab.html

trabajar con hasta 100.000 ejemplos o filas y hacer uso de hasta dos procesadores lógicos. No es necesaria la instalación de complementos adicionales.

Hardware

Los requerimientos del hardware están determinados por el software que se usará. Debido al consumo de tiempo y recursos para llevar a cabo el proyecto se prefiere disponer con una computadora de escritorio dedicada exclusivamente a estas tareas. El equipo deber tener un procesador Core i7, su equivalente o superior. Se requieren 16 GB de memoria RAM y una capacidad de disco duro de 500 GB. Estos requerimientos son los mínimos necesarios para el desarrollo del proyecto.

4.4 Supuestos

Se da por hecho que todas las imágenes contenidas en los documentos PDF se encuentran en escalas de grises y que las mismas al ser extraídas serán guardadas en formato .png conservando su color en escala de grises.

También se supone que los documentos PDF con las digitalizaciones pertenecientes al departamento de suscripción no se encuentran encriptados. En el caso de que algún documento haya sido encriptado por error, se deberá informar al equipo de desarrollo y solicitar su desencriptado.

4.5 Restricciones

El resultado esperado del desarrollo del proyecto es la obtención de un modelo que permita la clasificación de un conjunto de imágenes digitales. El propósito del proyecto no está centrado en el desarrollo de software, si no en la obtención del modelo.

Por cuestiones de confidencialidad las imágenes no podrán ser extraídas de los servidores de la empresa, por lo que se restringe el uso de servicios en la nube para su procesamiento. Una vez extraídos los datos que describen a las imágenes, cualquier propiedad que revele algún tipo de información confidencial debe ser anonimizado.

4.6 Riesgos y contingencias

A continuación, en la Tabla 4-1 se listan los riesgos asociados al proyecto. Se les asignaron valores según la probabilidad de ocurrencia y el impacto que tienen sobre el proyecto. Los valores se asignan según una escala del 1 al 5 en la cual el número uno representa la menor probabilidad de ocurrencia o menor impacto sobre el proyecto y el cinco la mayor probabilidad de ocurrencia o impacto.

N°	Riesgo	Probabilidad de Ocurrencia	Impacto	Valoración
1	Pérdida de datos obtenidos	3	5	15
2	Interrupción en el procesamiento de imágenes	3	4	12
3	Pérdida de imágenes	2	5	10
4	Pérdida de un integrante del equipo	2	4	8
5	Aumento en los costos del proyecto	5	5	25

Tabla 4-1 Riesgos asociados al proyecto

Una vez identificados y analizados los riesgos se elaboran los planes de contingencia para cada caso en particular.

Pérdida de datos

Al trabajar con información que se encuentra en formato digital existe el riesgo de pérdida de información, ya sea por falla en el soporte en el que se encuentra, borrado accidental o escritura sobre la información existente. En este proyecto se cuenta con dos grupos de información, las imágenes sobre las cuales se realiza el análisis y los datos obtenidos como resultado del análisis y procesamiento de imágenes. Con el objeto de mitigar este riesgo, se detallan a continuación las medidas a tomar para resguardar la información perteneciente a cada uno de los grupos de datos mencionados.

- **Imágenes:** Las imágenes originales pueden ser extraídas cuantas veces sea necesario desde el repositorio de donde fueron obtenidas. Debido a que las imágenes no pueden ser extraídas a un medio de almacenamiento externo, se guarda una copia del listado de los identificadores de las mismas en una unidad de red sobre la cual se realizan copias de respaldos diarios en cintas, según el procedimiento de backup del área de Tecnología. Esto permitirá, en caso de ser necesario, poder recuperar las imágenes desde su repositorio haciendo uso del listado de identificadores de las imágenes.

- **Datos obtenidos:** Los datos obtenidos son almacenados en primera instancia en archivos de formato.CSV (valores separados por coma). A medida que se obtengan estos archivos, además de ser almacenados localmente para su tratamiento, se almacenarán en la nube identificándolos con la fecha de creación y datos que contiene el archivo. Debido que los datos obtenidos solo guardan información respecto a la composición de la imagen y no contienen información de los clientes, es posible realizar copias de respaldos en la nube. Una vez que los datos son analizados se los almacenará en una base de datos para facilitar el manejo de la información. La copia de respaldo de la base de datos se realizará diariamente y antes de realizar cualquier modificación de importancia sobre los datos, de modo que si llegara a fallar es posible volver al punto más próximo de estabilidad.

Interrupción en el procesamiento de imágenes

El procesamiento de imágenes implica la modificación de las mismas con el objetivo de ocultar o resaltar determinadas características. Estos procesos generarán nuevas imágenes a partir de las originales. Esta tarea puede consumir una gran cantidad de tiempo y en caso de producirse una interrupción durante el procesamiento de las imágenes, podría perderse todo el trabajo realizado hasta ese punto. Con el objeto de minimizar pérdida de datos y tiempo, las imágenes serán procesadas en lotes de cantidades reducidas, de esta forma se tendrá más control sobre los procesos.

A medida que se vayan obteniendo los lotes de imágenes procesadas, serán almacenadas en una unidad de red destinada para tal fin, sobre la cual se realizará la copia de respaldo en cintas magnéticas según el procedimiento de backup del área de Tecnología.

Pérdida de un integrante del equipo

Si se produce la pérdida de un miembro del equipo durante el desarrollo del proyecto puede ocasionar un retraso en el cronograma y a su vez la sobrecarga de tareas para el resto de integrantes. Para mitigar este riesgo, al momento de realizar la selección del personal se conservará la información de aquellos recursos que no lograron formar parte del equipo, pero que presentan un perfil prometedor.

Si la salida de uno de los miembros del equipo se produce en las etapas iniciales del proyecto, se identificarán las tareas críticas que deberán ser redistribuidas entre los miembros restantes hasta la incorporación del nuevo integrante. Si la salida se produce en una etapa avanzada del proyecto se analizará el costo/beneficio de sumar una nueva persona en ese determinado momento, lo cual puede implicar capacitación previa sobre el proyecto en desarrollo y adaptación a la dinámica de trabajo.

Aumento en los costos del proyecto

Debido al contexto económico en el cual se ejecuta el proyecto, existe un elevado riesgo del incremento del costo de los recursos humanos, el software y el hardware, debido al incremento en la inflación. Para minimizar este riesgo se opta por cotizar todos los recursos en dólares de modo de minimizar el impacto del aumento de costos en la ejecución del proyecto.

4.7 Planificación del Proyecto

Las tareas que se llevan a cabo para cumplir con el objetivo del proyecto se muestran en el siguiente EDT (Estructura de Desglose de Trabajo):

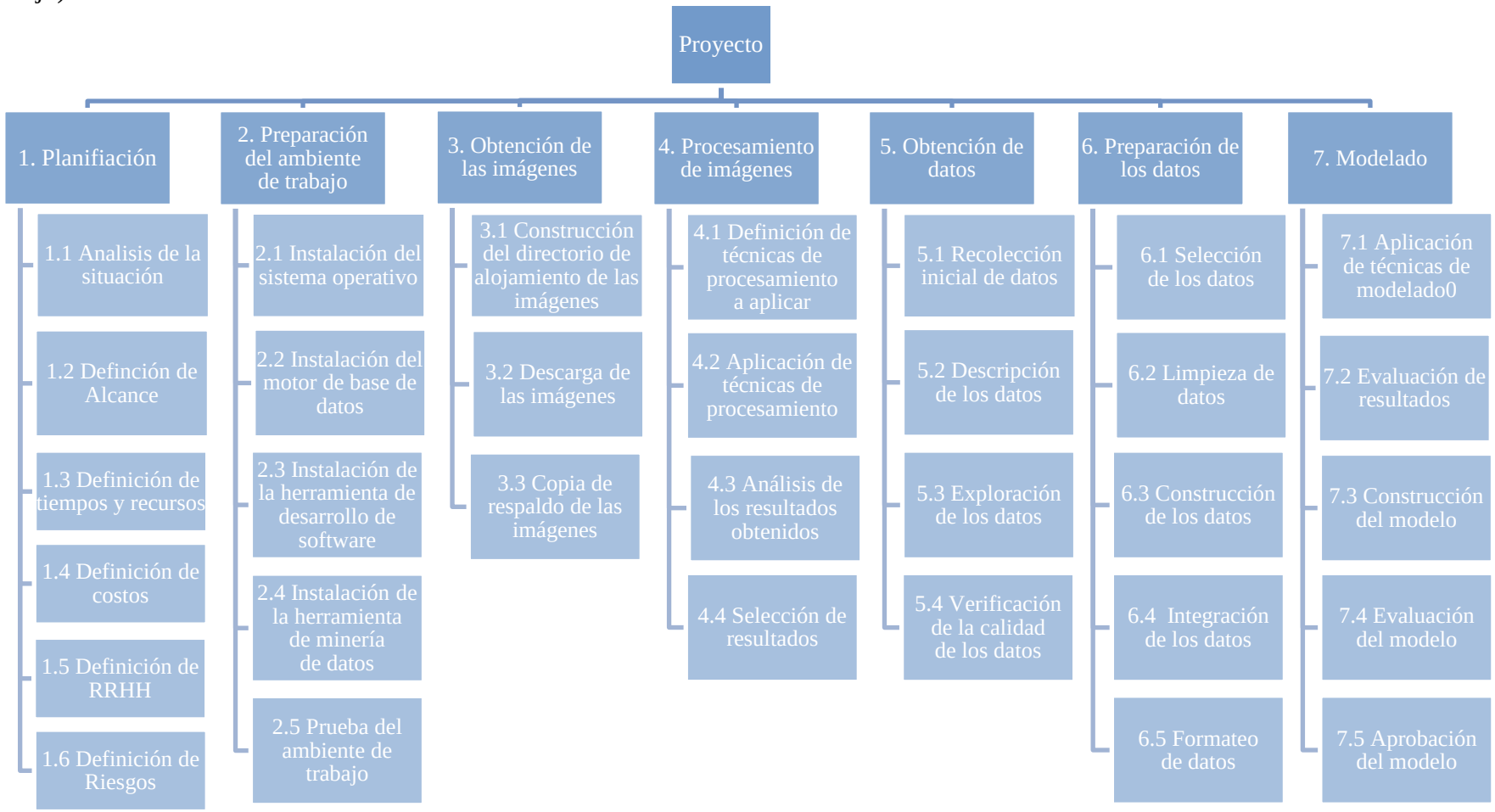


Imagen 4-1 Estructura de desglose de trabajo

A continuación, se describen cada una de las tareas de la EDT.

1. Planificación

1.1 Análisis de la situación: se realiza el análisis del problema evaluando el contexto del mismo, con el objetivo de plantar una solución. Se debe realizar el relevamiento de los requerimientos. En esta tarea se definen los objetivos y se identifican las partes interesadas.

1.2 Definición del alcance: Definir claramente en qué consistirá la solución, indicando qué tareas se realizarán y cuáles no se encuentran abarcadas por el proyecto.

1.3 Definición de tiempos y recursos: En base a los objetivos planteados se deben definir cuáles serán las tareas que se deben realizar para alcanzarlos. Por cada tarea se debe estimar la duración, fechas y dependencias entre ellas. Esto permitirá confeccionar el cronograma del proyecto. Se deben definir todos los recursos necesarios para la realización de las tareas, esto abarca al personal, *software*, *hardware* y cualquier otro elemento necesario.

1.4 Definición de costos: Una vez definidos los tiempos y recursos se debe confeccionar el presupuesto del proyecto teniendo en cuenta los recursos a emplease para la ejecución del mismo.

1.5 Definición de riesgos: Se deben identificar los riesgos que puedan afectar negativamente la ejecución del proyecto e idear los planes de contingencia necesarios para minimizar el impacto de los mismos.

1.6 Elaboración del plan de proyecto: Se deben documentar todas las tareas realizadas en la etapa de planificación y se debe comunicar a los interesados el plan del proyecto, que contiene la definición de cómo se llevará a cabo la ejecución del proyecto para obtener la solución al problema planteado inicialmente.

2. Preparación del ambiente de trabajo

2.1 Instalación del sistema operativo: consiste en la instalación del sistema operativo elegido sobre el cual se trabajará, además se deberán instalar todos los controladores y software básico para el correcto funcionamiento de la PC.

2.2 Instalación del motor de base de datos: Una vez instalado el sistema operativo y las herramientas básicas se procede a instalar y configurar el motor de base de datos elegido.

2.3 Instalación de la herramienta de desarrollo de software: Se instala el software de desarrollo elegido junto con las librerías necesarias para la ejecución del proyecto.

2.4 Instalación de la herramienta de minería de datos: Se instala la herramienta de minería de datos.

2.5 Prueba del ambiente de trabajo: Se procede a realizar pruebas sobre el software instalado con el objetivo de determinar el correcto funcionamiento del mismo y detectar errores o falta de algún componente, de modo de realizar las correcciones necesarias.

3. Obtención de las imágenes

3.1 Construcción del directorio de alojamiento de las imágenes: Se crea el directorio que aloja a las imágenes, de modo que la referencia a la ubicación de las imágenes se mantiene constante.

3.2 Descarga de las imágenes: En primera instancia se realiza la selección aleatoria de las imágenes que conformarán el conjunto de muestra con el cual se trabajará. Una vez identificadas las imágenes se procede a realizar la copia desde el repositorio al directorio creado para su alojamiento.

3.3 Copia de respaldo de las imágenes: Se realiza una copia de respaldo de las imágenes antes de comenzar a trabajar con ellas, de modo que si son modificadas o eliminadas accidentalmente puedan ser recuperadas. Adicionalmente se guardará una copia del listado de los identificadores de las imágenes tomadas como muestra.

4. Procesamiento de imágenes

4.1 Definición de técnicas de procesamiento a aplicar: Se analiza cuáles son las técnicas de procesamiento de imágenes que serán empleadas teniendo en cuenta que se trabajará con digitalizaciones de documentos pertenecientes a legajos de clientes. Las técnicas elegidas deben permitir mostrar información adicional que sea de utilidad para el proceso de clasificación.

4.2 Aplicación de técnicas de procesamiento: Se aplican las técnicas de procesamiento de imágenes definidas en el punto anterior. Se almacenan las imágenes resultantes para ser accedidas en pasos posteriores.

4.3 Análisis de los resultados obtenidos: Se analizan los resultados para identificar los datos obtenidos y determinar si existe la necesidad de aplicar otras técnicas de procesamiento de imágenes.

4.4 Selección de resultados: De todos los resultados obtenidos se conservan solo aquellos que brinden datos que permitan diferenciar las imágenes y se descartan los que no.

5. Obtención de datos

5.1 Recolección inicial de datos/imágenes: se realizan los procesos necesarios para obtener los valores de los atributos correspondientes a las imágenes originales y procesadas. Estos datos son almacenados en la base de datos creada para tal fin.

5.2 Descripción de los datos: Se analizan los conjuntos de datos obtenidos y se describe cada uno de ellos.

5.3 Exploración de los datos: en este punto se realiza un análisis más profundo de los datos. Se usan herramientas de minería de datos y estadísticas que permitan mostrar patrones en los datos, identificar rango de valores para los atributos, determinar si existen variaciones en el valor de determinados atributos.

5.4 Verificación de la calidad de los datos: se analizan los datos para determinar si existen valores faltantes, si es necesario realizar transformaciones para ser procesados, si existen valores anómalos.

6. Preparación de los datos

6.1 Selección de los datos: En base a los análisis realizados sobre los datos se decide cuáles serán los que se conservarán.

6.2 Limpieza de datos: En el caso de que se detecten datos erróneos, si es posible serán corregidos, caso contrario, serán descartados. También se realiza la detección de valores anómalos y se analiza si deben ser conservados o excluidos del conjunto de datos.

6.3 Construcción de los datos: En esta tarea se crean nuevos datos a partir de los existentes, se realizan transformaciones sobre los datos.

6.4 Integración de los datos: Se realiza la unificación de todos los datos obtenidos hasta el momento en una única vista o tabla.

6.5 Formateo de datos: Se realizan cambios en los datos para adaptarlos a los formatos admitidos por la herramienta de minería de datos.

7. Modelado

7.1 Aplicación de técnicas de modelado: Una vez definidos los conjuntos de datos se aplican las técnicas de modelado para obtener el modelo.

7.2 Evaluación de resultados: Se analizan los resultados obtenidos de la aplicación de las técnicas verificando si estos cumplen con los objetivos planteados.

7.3 Construcción del modelo: Se selecciona el modelo que cumple con los objetivos del proyecto.

7.4 Evaluación del modelo: Se evalúa el modelo elegido con los datos destinados a las pruebas.

7.5 Aprobación del modelo: Se aprueba el modelo en base a al resultado satisfactorio de la evaluación del mismo.

En la Tabla 4-2 se detallan las tareas junto con la duración y los recursos humanos asignados para cada una de ellas.

EDT	Tarea	Duración	Predecesora	Recurso
1	**Planificación**	**78 horas**		
1.1	Análisis de la situación	16 horas		Líder de Proyecto
1.2	Definición del alcance	8 horas		Líder de Proyecto
1.3	Definición de tiempos y recursos	12 horas		Líder de Proyecto
1.4	Definición de costos	12 horas		Líder de Proyecto
1.5	Definición de riesgos	8 horas		Líder de Proyecto
1.6	Elaboración del plan de proyecto	10 horas	1.1, 1.2, 1.3, 1.4, 1-5, 1.6	Líder de Proyecto
2	**Preparación del ambiente de trabajo**	**11 horas**		
2.1	Instalación del sistema operativo	4 horas	NA	Analista
2.2	Instalación del motor de base de datos	2 horas	2	Analista
2.3	Instalación de la herramienta de desarrollo de software	2 horas	3	Analista
2.4	Instalación de la herramienta de minería de datos	1 horas	4	Analista
2.5	Prueba del ambiente de trabajo	2 horas	5	Analista
3	**Obtención de las imágenes**	**3,5 horas**	**1**	
3.1	Construcción del directorio de alojamiento de las imágenes	0,5 horas	6	Analista
3.2	Descarga de las imágenes	2 horas	8	Analista
3.3	Copia de respaldo de las imágenes	1 hora	9	Analista
4	**Procesamiento de imágenes**	**32 horas**		
4.4	Definición de técnicas de procesamiento a aplicar	4 horas	7	Analista de datos
4.5	Aplicación de técnicas de procesamiento	24 horas	12	Analista de datos
4.6	Análisis de los resultados obtenidos	2 horas	13	Analista de datos
4.7	Selección de resultados	2 horas	14	Analista de datos

5	**Obtención de datos**	**88 horas**	**11**	
5.1	Recolección inicial de datos	40 horas	15	Analista de datos
5.2	Descripción de los datos	16 horas	17	Analista de datos
5.3	Exploración de los datos	16 horas	18	Analista de datos
5.4	Verificación de la calidad de los datos	16 horas	19	Analista de datos
6	**Preparación de los datos**	**48 horas**	**16**	
6.1	Selección de los datos	16 horas	20	Analista de datos
6.2	Limpieza de datos	16 horas	22	Analista de datos
6.3	Integración de los datos	8 horas	23	Analista de datos
6.4	Formateo de datos	8 horas	24	Analista de datos
7	**Modelado**	**48 horas**	**21**	
7.1	Aplicación de técnicas de modelado	24 horas	25	Analista de datos
7.2	Evaluación de resultados	8 horas	27	Analista de datos
7.3	Construcción del modelo	4 horas	28	Analista de datos
7.4	Evaluación del modelo	8 horas	29	Analista de datos
7.5	Aprobación del modelo	4 horas	30	Analista de datos

Tabla 4-2 Estructura de desglose de trabajo

A continuación, en la Imagen 4-2 se muestra el diagrama Gantt de la planificación del desarrollo de las tareas del proyecto.

EDT	Nombre de tarea	Duración	Comienzo	Fin
1	**Planificación**	**78 horas**	**lun 13/7/20**	**vie 24/7/20**
1.1	Análisis de la situación	16 horas	lun 13/7/20	mar 14/7/20
1.2	Definición del alcance	8 horas	mié 15/7/20	mié 15/7/20
1.3	Definición de tiempos y recursos	12 horas	jue 16/7/20	vie 17/7/20
1.4	Definición de costos	12 horas	lun 20/7/20	mar 21/7/20
1.5	Definición de riesgos	8 horas	mié 22/7/20	mié 22/7/20
1.6	Elaboración del plan de proyecto	10 horas	jue 23/7/20	vie 24/7/20
2	**Preparación del ambiente de trabajo**	**10,4 horas**	**lun 27/7/20**	**mar 28/7/20**
2.1	Instalación del sistema operativo	4 horas	lun 27/7/20	lun 27/7/20
2.2	Instalación del motor de base de datos	2 horas	lun 27/7/20	lun 27/7/20
2.3	Instalación de la herramienta de desarrollo de software	2 horas	lun 27/7/20	lun 27/7/20
2.4	Instalación de la herramienta de minería de datos	1 hora	mar 28/7/20	mar 28/7/20
2.5	Prueba del ambiente de trabajo	1 hora	mar 28/7/20	mar 28/7/20
3	**Obtención de las imágenes**	**3,5 horas**	**mar 28/7/20**	**mar 28/7/20**
3.1	Construcción del directorio de alojamiento de las imágenes	0,5 horas	mar 28/7/20	mar 28/7/20
3.2	Descarga de las imágenes	2 horas	mar 28/7/20	mar 28/7/20
3.3	Copia de respaldo de las imágenes	1 hora	mar 28/7/20	mar 28/7/20
4	**Procesamiento de imágenes**	**32 horas**	**mar 28/7/20**	**lun 3/8/20**
4.1	Definición de técnicas de procesamiento a aplicar	4 horas	mar 28/7/20	mié 29/7/20
4.2	Aplicación de técnicas de procesamiento	24 horas	mié 29/7/20	lun 3/8/20
4.3	Análisis de los resultados obtenidos	2 horas	lun 3/8/20	lun 3/8/20
4.4	Selección de resultados	2 horas	lun 3/8/20	lun 3/8/20
5	**Obtención de datos**	**88 horas**	**lun 3/8/20**	**mar 18/8/20**
5.1	Recolección inicial de datos	40 horas	lun 3/8/20	lun 10/8/20
5.2	Descripción de los datos	16 horas	lun 10/8/20	mié 12/8/20
5.3	Exploración de los datos	16 horas	mié 12/8/20	vie 14/8/20
5.4	Verificación de la calidad de los datos	16 horas	vie 14/8/20	mar 18/8/20
6	**Preparación de los datos**	**48 horas**	**mar 18/8/20**	**mié 26/8/20**
6.1	Selección de los datos	16 horas	mar 18/8/20	jue 20/8/20
6.2	Limpieza de datos	16 horas	jue 20/8/20	lun 24/8/20
6.3	Integración de los datos	8 horas	lun 24/8/20	mar 25/8/20
6.4	Formateo de datos	8 horas	mar 25/8/20	mié 26/8/20
7	**Modelado**	**44 horas**	**mié 26/8/20**	**jue 3/9/20**
7.1	Aplicación de técnicas de modelado	24 horas	mié 26/8/20	lun 31/8/20
7.2	Evaluación de resultados	8 horas	lun 31/8/20	mar 1/9/20
7.3	Construcción del modelo	4 horas	mar 1/9/20	mié 2/9/20
7.4	Evaluación del modelo	4 horas	mié 2/9/20	mié 2/9/20
7.5	Aprobación del modelo	4 horas	mié 2/9/20	jue 3/9/20

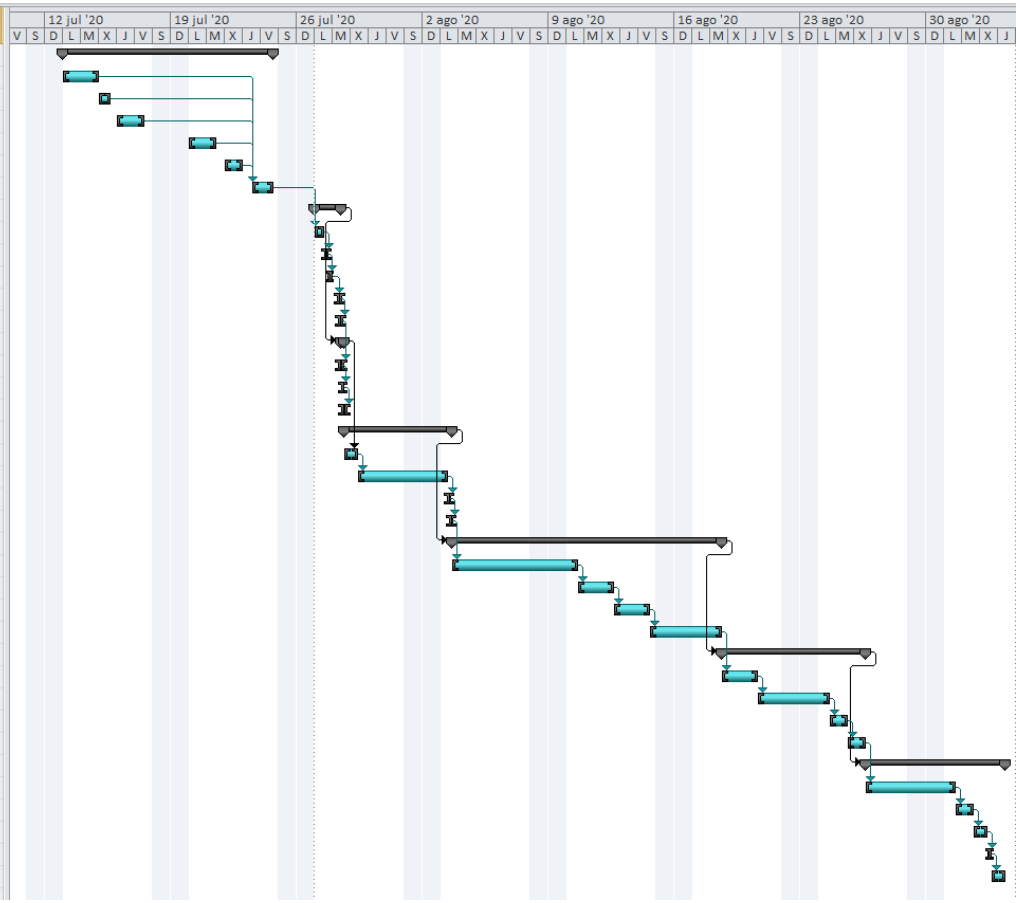


Imagen 4-2 Diagrama Gantt de la planificación del proyecto

4.8 Costos y Beneficios

Costos

Los costos de las licencias de software para el desarrollo del proyecto se encuentran detallados en la Tabla 4-3. Los precios de las licencias se encuentran detallados en dólares y la suscripción a las mismas es por el período de un año.

Producto	Cantidad	Precio Unitario (U$D)	Subtotal (U$D)
Licencia de RapidMiner Studio Professional[17]	1	7.500,00	7.500,00
Licencia de Visual Studio 2019[18]	1	1.199,00	1.199,00
Licencia de Windows 10[19]	1	199,99	199,99
Licencia de librería "Java Pdf Extraction Decoding Access Library"	1	1.000,00	1.000,00
Licencia de SQL Server 2017 Express	1	0,00	0,00
Licencia EMGU CV	1	0,00	0,00
Total	**4**	**9.898,99**	**9.898,99**

Tabla 4-3 Costos de licencias de software

Tanto la licencia de SQL Server 2017 Express como la licencia de la librería EMGU CV no tienen costo. La librería EMGU CV posee una licencia GNU lo cual implica que es de uso gratuito, pero no permite que sea modificada.

El costo del hardware necesario para el funcionamiento óptimo del software asociado se detalla a continuación en la Tabla 4-4. Los precios de los equipos se encuentran expresados en dólares y cumplen con los requerimientos establecidos en la sección 4.3.

[17] Precio consultado el 01/08/2020 en: https://rapidminer.com/pricing/
[18] Precio consultado el 01/08/2020 en: https://visualstudio.microsoft.com/es/vs/pricing-details/
[19] Precio consultado el 01/08/2020 en: https://www.microsoft.com/es-us/p/windows-10-pro/df77x4d43rkt/48DN

Producto	Cantidad	Precio Unitario (U$D)	Subtotal (U$D)
ThinkCentre M90n[20]	1	1.499,00	1.499,00
ThinkVision S24e 23.8 Inch LED Backlit LCD Monitor[21]	1	129,00	129,00
Total	**2**	**1.538,00**	**1.538,00**

Tabla 4-4 Costo de hardware

El costo de los recursos humanos involucrados en el proyecto es calculado en base a al tiempo asignado a cada uno de los recursos (en horas) y los honorarios por hora según COPAIPA[22]. Los valores fueron convertidos a dólares según la cotización [23]del Banco de la Nación Argentina[24]. Esta información se detalla en la Tabla 4-5, los costos se encuentran expresados en pesos argentinos.

Recurso	Cantidad de horas	Precio por hora (U$D)	Subtotal (U$D)
Líder de proyecto	78	11,02	859,28
Analista de minería de datos	215	7,87	1.691,80
Analista	110	6,56	721,31
Total			**3.272,39**

Tabla 4-5 Costos de recursos humanos

El costo total de la ejecución del proyecto, contemplando costos de software, hardware y recursos humanos necesarios es de 14.709,38 dólares estadounidenses, como se muestra en la tabla. Al cotizarse el presupuesto del proyecto en dólares se minimiza el impacto de un posible aumento del costo de los recursos.

Concepto	Subtotal (U$D)
Hardware	1.538,00
Software	9.898,99
Recursos Humanos	3.272,39
Total	**14.709,38**

Tabla 4-6 Costo total de la ejecución del proyecto

[20] Precio consultado el 01/08/2020 en: https://www.lenovo.com/us/en/desktops-and-all-in-ones/thinkcentre
[21] Precio consultado el 01/08/2020 en: https://www.lenovo.com/us/en/accessories-and-monitors/monitors/office
[22] http://www.copaipa.org.ar/informatica/
[23] Cotización consultada el 31/03/2021 en: https://www.bna.com.ar/Personas
[24] https://www.bna.com.ar/Personas

Beneficios

El análisis de los beneficios esperados al finalizar la ejecución del proyecto será realizado de forma cualitativa, ya que el beneficio inmediato que se pretende obtener es una optimización de las tareas de los analistas de emisión y de los usuarios del actual sistema de consulta de legajo digital de clientes. En este caso en particular, realizar supuestos para cuantificar los beneficios podría resultar en un análisis de costo-beneficio incorrecto.

Reducción del tiempo empleado en tareas operativas: La implementación del proyecto generará una disminución en los tiempos de búsqueda de documentos para los usuarios, lo cual reducirá el tiempo de las tareas operativas, pudiendo destinarlo a tareas de mayor importancia como las de análisis. Se estima que actualmente el tiempo invertido en la búsqueda de documentos es de 1 hora diaria para el personal del área de emisión. La implementación del proyecto evitará esta tarea por lo cual, considerando una jornada laboral de 8 horas, implica la reducción del 12% del tiempo invertido.

Agilización del proceso de atención al cliente: Durante la atención al cliente se realizan consultas a su legajo digital en búsqueda de determinados tipos de documentos dependiendo del tipo de consulta del cliente. Para esto, el responsable de la atención, debe recorrer uno por uno los documentos y realizar la previsualización de los mismos hasta encontrar el indicado. Una vez implementado el proyecto el empleado podrá realizar la búsqueda filtrando por tipo de documento lo cual disminuirá el tiempo de espera del cliente y por lo tanto aumentará su satisfacción.

Reducción de cantidad de documentos en guarda: Al conocer cuál es la documentación física que se tiene almacenada, es posible determinar cuál es aquella que no requiere mantenerse en guarda, de forma de poder destruirla, reduciendo costos de almacenamiento, liberando espacios físicos y tiempo de búsqueda de los documentos físicos. El costo que podrá ser reducido se conocerá una vez finalizado el proyecto, ya que obtendremos la clase de cada documento.

4.9 Factibilidad del Proyecto

4.9.1 Factibilidad legal

Las imágenes correspondientes a las digitalizaciones de los legajos de los clientes pueden contener información sensible de carácter confidencial. Debido a que la información que se extrae de las imágenes para el proceso de minería no revela este tipo de información, no existen restricciones en cuanto al uso de la misma. Sin embargo, durante el proceso de extracción de las imágenes y el procesamiento de las mismas, la información sensible se encuentra expuesta. Es por esto que la ubicación de red asignada para su almacenamiento solo será accesible por el usuario de la persona responsable de las tareas de extracción y procesamiento de imágenes. Adicionalmente quien tenga acceso a las imágenes firmará un contrato de confidencialidad. Estas medidas asegurarán el resguardo de la privacidad de la información sensible.

4.9.2 Factibilidad Técnica - Operativa

Los recursos necesarios (recursos humanos, software y hardware) para la ejecución del proyecto se encuentran disponibles para su adquisición.

En un análisis inicial se puede observar que no existe resistencia al cambio por parte del personal involucrado en el proyecto. Todos ellos comparten la necesidad, los requerimientos y los beneficios que obtendrán al finalizar el proyecto. Con el objeto de mantener esta situación, durante el desarrollo de las tareas se involucrarán a las partes interesadas y se mantendrá una comunicación fluida para informar sobre el estado del proyecto.

4.10 Entregables

A continuación, se describen los entregables que se esperan obtener como resultado de la ejecución de las tareas del proyecto:

Plan del proyecto: Una vez finalizada la etapa de planificación se obtiene como resultado el plan del proyecto. Este documento contiene la información respecto al alcance, objetivos, recursos, riesgos y contingencias, costos y beneficios, cronograma de actividades y toda información necesaria que guiará la ejecución del proyecto. El plan será comunicado a todas las partes interesadas.

Script de extracción de imágenes: Las imágenes de las digitalizaciones de los documentos de los legajos de los clientes se encuentran almacenadas en archivos con formato PDF. Un documento PDF puede estar constituido por una o más páginas. Cada página del documento posee una única imagen resultante del proceso de digitalización. En el caso de los documentos que poseen más de una página, la primera es la que define la clase a la cual pertenecen el resto. Para poder realizar el procesamiento de las imágenes es necesario extraerlas de los documentos PDF. Al finalizar el proyecto se entregará al departamento de Desarrollo el código del programa desarrollado para realizar esta tarea.

Script de procesamiento de imágenes: La tarea de procesamiento de imágenes permite generar nuevas a partir de las originales con el objetivo de acentuar o atenuar determinadas características. Esto permitirá obtener información adicional sobre cada imagen de modo de poder emplearla en el proceso de descubrimiento de conocimiento. El procesamiento de las imágenes se realizará mediante el uso de la librería EMGU CV (descripta en la sección 4.3), el código fuente resultante será entregado al departamento de Desarrollo.

Script de extracción de datos: Para realizar el análisis de las imágenes, tanto originales como procesadas, es necesario realizar la extracción de las características que poseen, en el formato admitido por la herramienta de minería de datos. El código fuente empleado para realizar esta tarea será entregado al departamento de Desarrollo.

Proceso de RapidMiner: se hará entrega del proceso de modelado creado con la herramienta RapidMiner. Si en el futuro se llegara a requerir un reentrenamiento del algoritmo con nuevos datos, el mismo proceso podrá ser usado con mínimas o ninguna modificación.

Modelo: Como resultado del proceso de minería de imágenes, en este proyecto en particular, se espera obtener un modelo que realice la clasificación de imágenes de digitalizaciones de legajos de clientes. El modelo obtenido será entregado al departamento de desarrollo para que una vez concluido el proyecto se realice la integración del mismo al software existente de consulta de legajos de clientes.

5. Implementación de la metodología CRISP-DM

5.1 Comprensión del negocio

La compañía de seguros produce diariamente grandes volúmenes de información en formato papel. Para facilitar el acceso a esta información se tomó la decisión de realizar la digitalización de los documentos físicos pertenecientes a los legajos de los clientes. También se implementó un software de consulta que permite realizar la visualización de los documentos en formato digital. Las digitalizaciones actualmente no poseen ningún dato que permita identificar el tipo de documento del cual se trata. Esto genera como consecuencia el enlentecimiento en el proceso de búsqueda de documentos y a su vez no se posee el conocimiento sobre con qué tipos de documentos se cuentan.

La ejecución de la primera fase de la metodología CRISP-DM consiste en lograr un conocimiento en detalle del negocio para poder alcanzar los objetivos planteados y satisfacer la necesidad planteada inicialmente. En el capítulo 3 se describe el negocio, se definen los sus objetivos y los objetivos de minería de datos y se define el alcance del proyecto. En el capítulo 4 se detallan los recursos necesarios para la ejecución del proyecto, los riesgos y contingencias, los costos y beneficios y el análisis de factibilidad. También se detalla el plan de tareas del proyecto y los entregables.

5.2 Comprensión de los datos

5.2.1 Recolección inicial de imágenes y propiedades de las imágenes

Las imágenes con las cuales se trabaja provienen de un repositorio de legajos de clientes. Este repositorio está constituido por más de un millón y medio de documentos en formato PDF. Estos documentos pueden contener una o más hojas. En cada hoja del documento se encuentra solo una imagen que se obtiene como resultado del escaneo de la documentación perteneciente a los legajos de los clientes.

Considerando que la cantidad de documentos PDF que se generan mensualmente como producto de los escaneos, se mantiene constante y que la cantidad de documentos promedio por mes, en un período de tres años, es de 15.481 documentos, se decide tomar una muestra igual a la media. En el caso de que el documento contenga más de una imagen, solo se recupera la primera, ya que las siguientes se tratan del dorso de la primera o imágenes adicionales que se anexan. Por lo tanto, el proceso de extracción dará como resultado 15.481 imágenes.

La muestra se toma de la base de datos de forma aleatoria usando la función random provista por vb.net. A partir de la muestra se procede a la extracción de las imágenes de cada uno de los archivos. Para lograr esto se hace uso de la librería “Java Pdf Extraction Decoding Access Library”. Esta librería permite acceder a cada objeto contenido en un documento PDF, inspeccionar el tipo y extraer solo aquellos que son de interés, en este caso, imágenes. Se extrae la primera imagen de cada archivo PDF y se almacena en un nuevo repositorio, asignándole el nombre original del PDF.

De este proceso se obtuvieron un total de 15.481 imágenes en formato PNG. El nombre de cada archivo es único, por lo tanto, este nombre es asignado a la imagen como identificador único. Este identificador permite asociar la imagen con un determinado cliente.

Debido a que no se cuenta con información de las imágenes, se deben definir los atributos que se usarán para describirlas y su clase. Los documentos pertenecientes a los legajos de los clientes pueden ser agrupados en clases. Para determinar la clase de cada imagen de forma certera, el proceso se realiza de forma manual, inspeccionando visualmente cada imagen y asignándole la respectiva clase. Es necesario que la asignación de la clase sea correcta, ya que este valor es uno de los más importantes en el proceso de construcción del modelo de clasificación; será el atributo a predecir.

De este proceso se obtuvieron un total de 73 posibles clases de documentos, lo cual implica que el atributo clase puede tomar cualquiera de los 73 valores nominales que representan cada clase. Con el objeto de mantener la confidencialidad de los datos, los nombres de las clases serán reemplazados por la letra C, seguida de dos dígitos, por ejemplo, C15. En la Tabla 5-1 se muestran las clases identificadas, la cantidad de imágenes por cada una de ellas y el porcentaje que representan del total:

Clase	Cantidad de Imágenes	Porcentaje del total de imágenes
C03	2019	13,04%
C80	1719	11,11%
C25	1613	10,42%
C61	546	3,53%
C59	525	3,39%
C49	395	2,55%
C65	380	2,45%
C50	373	2,41%
C39	354	2,29%
C01	315	2,03%
C15	308	1,99%
C76	297	1,92%
C57	285	1,84%
C71	282	1,82%
C26	275	1,78%
C46	234	1,51%
C48	230	1,49%
C37	209	1,35%
C84	195	1,26%
C72	187	1,21%
C41	148	0,96%
C06	141	0,91%
C27	125	0,81%
C44	124	0,80%
C24	121	0,78%
C70	116	0,75%

C08	115	0,74%
C69	108	0,70%
C62	101	0,65%
C43	100	0,65%
C68	98	0,63%
C74	95	0,61%
C73	93	0,60%
C17	91	0,59%
C64	89	0,57%
C58	88	0,57%
C66	81	0,52%
C38	75	0,48%
C54	71	0,46%
C34	67	0,43%
C11	60	0,39%
C75	57	0,37%
C05	52	0,34%
C07	50	0,32%
C55	48	0,31%
C56	48	0,31%
C16	47	0,30%
C28	47	0,30%
C77	43	0,28%
C13	36	0,23%
C29	34	0,22%
C36	34	0,22%
C04	32	0,21%
C12	32	0,21%
C51	29	0,19%
C18	28	0,18%
C31	27	0,17%
C02	24	0,16%
C67	22	0,14%
C33	18	0,12%
C60	17	0,11%
C19	14	0,09%
C47	14	0,09%
C30	13	0,08%
C42	13	0,08%
C52	13	0,08%
C22	12	0,08%
C53	12	0,08%
C32	11	0,07%
C35	10	0,06%
C45	10	0,06%
C09	9	0,06%
C10	5	0,03%

Tabla 5-1 Cantidad de imágenes por clases

Para poder realizar el proceso de clasificación de las imágenes es necesario conocer sus características. Debido a que no se cuenta con estos datos, es necesario definir los atributos que describen las imágenes y obtener los valores para esos atributos. A continuación, se detallan atributos que describen a las imágenes originales:

- **Alto:** atura de la imagen expresada en píxeles.
- **Ancho:** ancho de la imagen expresada en píxeles.
- **Cantidad de píxeles blancos:** se contabilizan la cantidad de píxeles blancos de la imagen original. Un píxel blanco tendrá los valores RGB (255, 255, 255).
- **Cantidad de píxeles negros:** cantidad de píxeles negros que poseen los valores RGB (0, 0, 0).

Los valores para estos atributos son obtenidos de las imágenes originales mediante un script desarrollado en vb.net. El resultado del script es almacenado en un archivo CSV.

Sobre la imagen original se aplican transformaciones mediante métodos provistos por la librería EMGU CV. Esta librería brinda herramientas para el procesamiento de imágenes. A cada imagen se le aplica un método que realiza su transformación, lo cual da como resultado una nueva imagen, visualmente distinta, que es almacenada. Como se menciona en la sección 2.4, el procesamiento de imágenes consiste en modificar una imagen mediante operadores matemáticos con el objetivo de destacar u ocultar determinadas características. Siendo la erosión y la dilatación las dos transformaciones principales de las cuales se derivan muchas otras, se seleccionaron los métodos en EMGU CV que realizan este tipo de tarea. Al erosionar una imagen, los bordes de las figuras son reducidos, convirtiéndose en parte del fondo. En cambio, al dilatar una imagen los bordes de las figuras son engrosados consumiendo parte del fondo. Otras técnicas como el cálculo de la transformada de Fourier y la integral nos permiten representar la información contenida en una imagen de una forma alternativa. Por ejemplo, en el caso aplicar la transformada de Fourier obtendremos una imagen expresada un dominio distinto, el dominio de las frecuencias. El objeto de aplicar estas dos técnicas es encontrar nuevas características de las imágenes que puedan servir para ser empleadas en el proceso de clasificación. En la Imagen 5-1 se muestra un ejemplo de las transformaciones elegidas aplicadas a las imágenes originales y sus respectivos resultados.

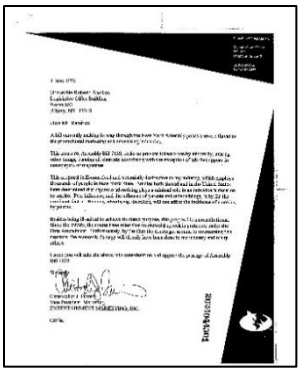
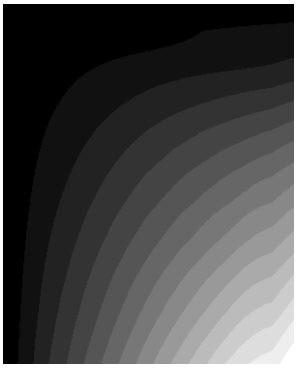
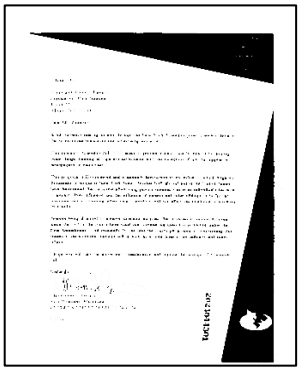


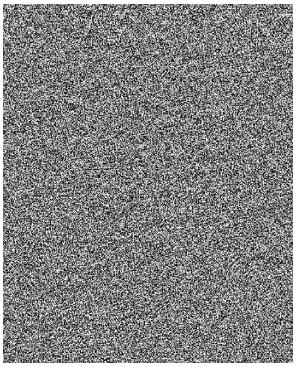
Imagen Original



Integral



Desenfoque



Transformada de Fourier

Imagen 5-1 Imagen original y resultantes de la aplicación de transformaciones

Las transformaciones empleadas son las siguientes:

- **Integral:** En la imagen obtenida como resultado, el valor de la intensidad de cada píxel está determinado por la suma de las intensidades de los píxeles en línea vertical por arriba y en línea horizontal por la izquierda.

- **Desenfoque y diferenciación gaussiana (sobel):** El operador sobel detecta los bordes en la imagen y como resultado devuelve una imagen con los bordes aclarados y el fondo oscuro.

- **Desenfoque (smoothmedian):** realiza el desenfoque de una imagen empleando el filtro median. Este filtro remueve el ruido en las imágenes segmentando la imagen en pequeños fragmentos, calcula el promedio de la intensidad de cada fragmento y el pixel que se encuentra en la parte central toma el valor calculado.

- **Transformada de Fourier (DTF):** El método DTF toma la imagen y devuelve como resultado su transformada de Fourier, es decir, obtenemos la información de la imagen original expresada en el dominio de las frecuencias.

Para cada una de las imágenes resultantes se determinan la cantidad de pixeles blancos y la cantidad de pixeles negros. Esto da como resultado los siguientes atributos:

- **Integral.Int_0:** cantidad de píxeles negros de una imagen que se obtuvo a partir de la aplicación del método Integral.
- **Integral.Int_255:** cantidad de píxeles blancos de una imagen que se obtuvo a partir de la aplicación del método Integral.
- **Sobel.Int_0:** cantidad de píxeles negros de una imagen que se obtuvo a partir de la aplicación del método Sobel.
- **Sobel.Int_255:** cantidad de píxeles blancos de una imagen que se obtuvo a partir de la aplicación del método Sobel.
- **Smoothmedian.Int_0:** cantidad de píxeles negros de una imagen que se obtuvo a partir de la aplicación del método Smoothmedian.
- **Smothmedian.Int_255:** cantidad de píxeles blancos de una imagen que se obtuvo a partir de la aplicación del método Smoothmedian.
- **DTF.Int_0:** cantidad de píxeles negros de una imagen que se obtuvo a partir de la aplicación de la transformada de Fourier a la imagen.
- **DTF.Int_255:** cantidad de píxeles blancos de una imagen que se obtuvo a partir de la aplicación de la transformada de Fourier a la imagen.

Los valores obtenidos para cada atributo por cada imagen son almacenados en archivos CSV.

5.2.2 Descripción de los datos

Para simplificar la manipulación de los datos obtenidos, son alojados en tablas en una base de datos con el esquema que se muestra en la Imagen 5-2. Cada tabla posee una cantidad de 15.481 registros, un registro por cada imagen procesada, que posee los valores de los atributos extraídos.

Todas las tablas poseen un identificador único, el nombre de la imagen, que es tomado del archivo PDF del que proviene la imagen. La tabla Imagen posee dos atributos, el nombre de la imagen y su ubicación física. La tabla Original almacena el nombre de la imagen junto con los atributos Int_0 e Int_255 que contabilizan la cantidad de píxeles blancos y píxeles negros respectivamente. La tabla Altoancho almacena, además del identificador único, la información del ancho y alto expresados en píxeles de las imágenes originales. Estos atributos permiten buscar

relaciones entre la clase de la imagen y las dimensiones de la misma. La información sobre la clase a la pertenece cada imagen se encuentra almacenada en la tabla Clase.

Por cada lote de nuevas imágenes obtenidas a partir de las transformaciones aplicadas a la imagen original, se crea una tabla. Estas tablas son Sobel, Integral, DTF y Smooth. Sus estructuras son similares, un atributo para el identificador único (nombre de la imagen) y dos atributos adicionales que indican la cantidad de píxeles blancos y negros de cada imagen.

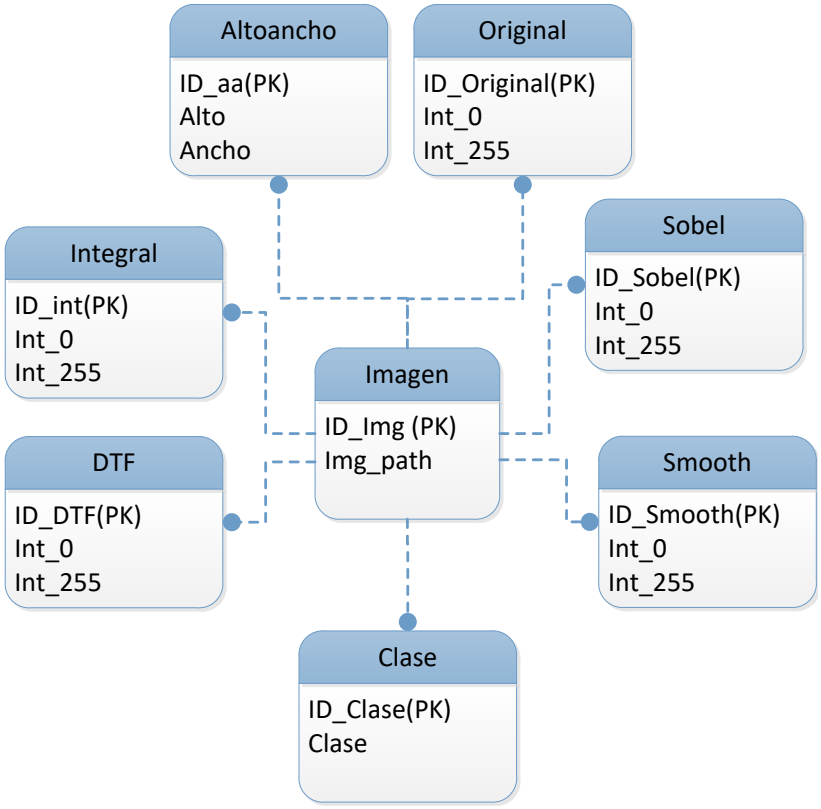


Imagen 5-2 Diagrama de base de datos para el almacenamiento de los datos obtenidos

La Tabla 5-2 muestra un resumen de características de los atributos de la tabla Altoancho que permiten realizar una primera exploración de los datos.

Nombre de la tabla: Altoancho							
Atributo	**Tipo de dato**	**Valor mínimo (px)**	**Valor máximo (px)**	**Cant. de valores únicos**	**Cant. de valores nulos**	**Media**	**Desvío estándar**
ID_aa	nvarchar	-	-	15841	0	-	-
Alto	Integer	228	4516	1416	0	1884	1169,27
Ancho	Integer	435	4324	633	0	1515	753,81

Tabla 5-2 Descripción de atributos de la tabla Altoancho

Se puede observar que los atributos Alto y Ancho no presentan valores nulos o faltantes. La cantidad de valores únicos para el atributo ID_aa indica que existe un registro para cada una de las imágenes pertenecientes al lote tomado como muestra.

Las tablas DTF, Integral, Sobel y Smooth contienen los datos de las imágenes que se obtuvieron como resultado del procesamiento de la imagen original. Las características de los datos de estas tablas se detallan a continuación:

Nombre de la tabla: Sobel							
Atributo	**Tipo de dato**	**Valor mínimo (px)**	**Valor máximo (px)**	**Cant. de valores únicos**	**Cant. de valores nulos**	**Media**	**Desvío estándar**
ID_Sob	nvarchar	-	-	15.481	0	-	-
Int_0	Integer	111.701	11.039.441	15.370	0	3.345.336,73	3.379.678,05
Int_255	Integer	112	11.039.441	14.638	0	143.035,70	120.712,33

Tabla 5-3 Descripción de atributos de la tabla Sobel

Nombre de la tabla: DTF							
Atributo	**Tipo de dato**	**Valor mínimo (px)**	**Valor máximo (px)**	**Cant. de valores únicos**	**Cant. de valores nulos**	**Media**	**Desvío estándar**
ID_DTF	nvarchar	-	-	15.481	0	-	-
Int_0	Integer	64.106	5.781.828	15.153	0	1.804.134,90	1.783.343,37
Int_255	Integer	62.880	5.684.757	15.154	0	1.766.695,90	1.743.303,46

Tabla 5-4 Descripción de atributos de la tabla DTF

Nombre de la tabla: Integral							
Atributo	**Tipo de dato**	**Valor mínimo (px)**	**Valor máximo (px)**	**Cant. de valores únicos**	**Cant. de valores nulos**	**Media**	**Desvío estándar**
ID_int	nvarchar	-	-	15481	0	-	-
Int_0	Integer	2037	1081912	11500	0	51976,86	57245,21
Int_255	Integer	1	2732	113	0	11,02	39,33

Tabla 5-5 Descripción de atributos de la tabla Integral

Nombre de la tabla: Smooth							
Atributo	**Tipo de dato**	**Valor mínimo (px)**	**Valor máximo (px)**	**Cant. de valores únicos**	**Cant. de valores nulos**	**Media**	**Desvío estándar**
ID_Smo	nvarchar	-	-	15481	0	-	-
Int_0	Integer	0	2813141	11693	0	67.406,21	197.699,85
Int_255	Integer	95.524	11.010.272	15360	0	3.267.057,79	3.185.694,80

Tabla 5-6 Descripción de los atributos de la tabla Smooth

Para las cuatro tablas el valor de cantidad de valores únicos es de 15481, lo cual implica que para todas las imágenes del lote de muestra existe un registro. El atributo cantidad de valores nulos, contabiliza aquellos registros que tienen valores faltantes. Como se puede observar, ninguna de las tablas posee valores faltantes.

La media indica el valor donde los datos tienden a estar más cerca uno de otros, mientras que el desvío estándar nos señala que tan dispersos se encuentran.

La cantidad de valores únicos para el atributo Int_255 en de la tabla Integral (Tabla 5-5), es el menor valor y difiero mucho del resto. Este dato debería ser analizado en mayor detalle para determinar si se encuentra agrupados en conjuntos definidos y tienen relación con la clase a la que pertenece cada imagen.

En la Tabla 5-6 se observa que 185 imágenes no poseen píxeles con intensidad igual a 0, es decir píxeles negros. Ese grupo de imágenes pueden ser consideradas como anómalas y requieren una inspección visual.

5.2.3 Exploración de los datos

Para realizar un análisis de la dispersión de los datos se emplearán diagramas de cajas. Este tipo de diagrama es una herramienta que nos permite observar la distribución de los elementos de la muestra en un gráfico. El rectángulo central representa el intervalo donde se agrupa el 50 por ciento del total de los datos. Las líneas en los extremos del gráfico representan el valor mínimo y máximo, estas se denominan colas. Cada una de ellas representan el 25% de los datos. En la Imagen 5-3 se muestra un diagrama de caja para los atributos Int_0 (cantidad de píxeles negros) e Int_255 (cantidad de píxeles blancos) de la tabla Sobel. Se observa que los valores para el atributo Int_255 se encuentran más concentrados que los valores del atributo Int_0, que se encuentran dispersos en un intervalo más amplio. Para el atributo int_255 se observan puntos que sobrepasan el valor máximo. Estos valores son considerados muy alejados y podrían tratarse de anomalías. Serán tenidos en cuenta para la fase de limpieza de los datos.

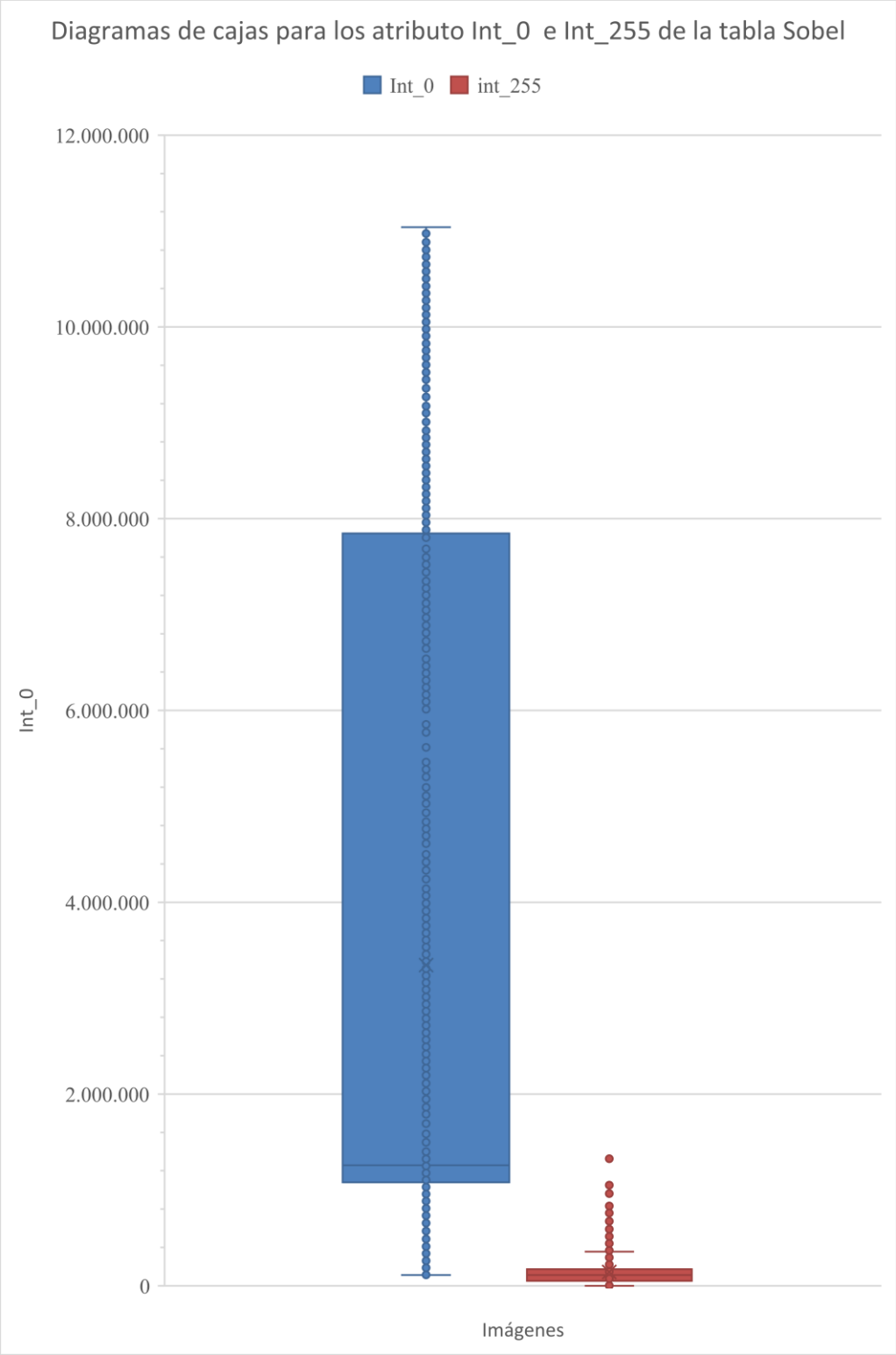


Imagen 5-3 Diagrama de caja para los atributos Int_0 e Int_255 de la tabla Sobel

En la Imagen 5-4 se puede apreciar la distribución de los datos para los valores de los atributos Int_0 e Int_255 de la tabla DTF. Para ninguno de los dos atributos existen valores muy alejados y los intervalos en los cuales se distribuyen los datos son similares.

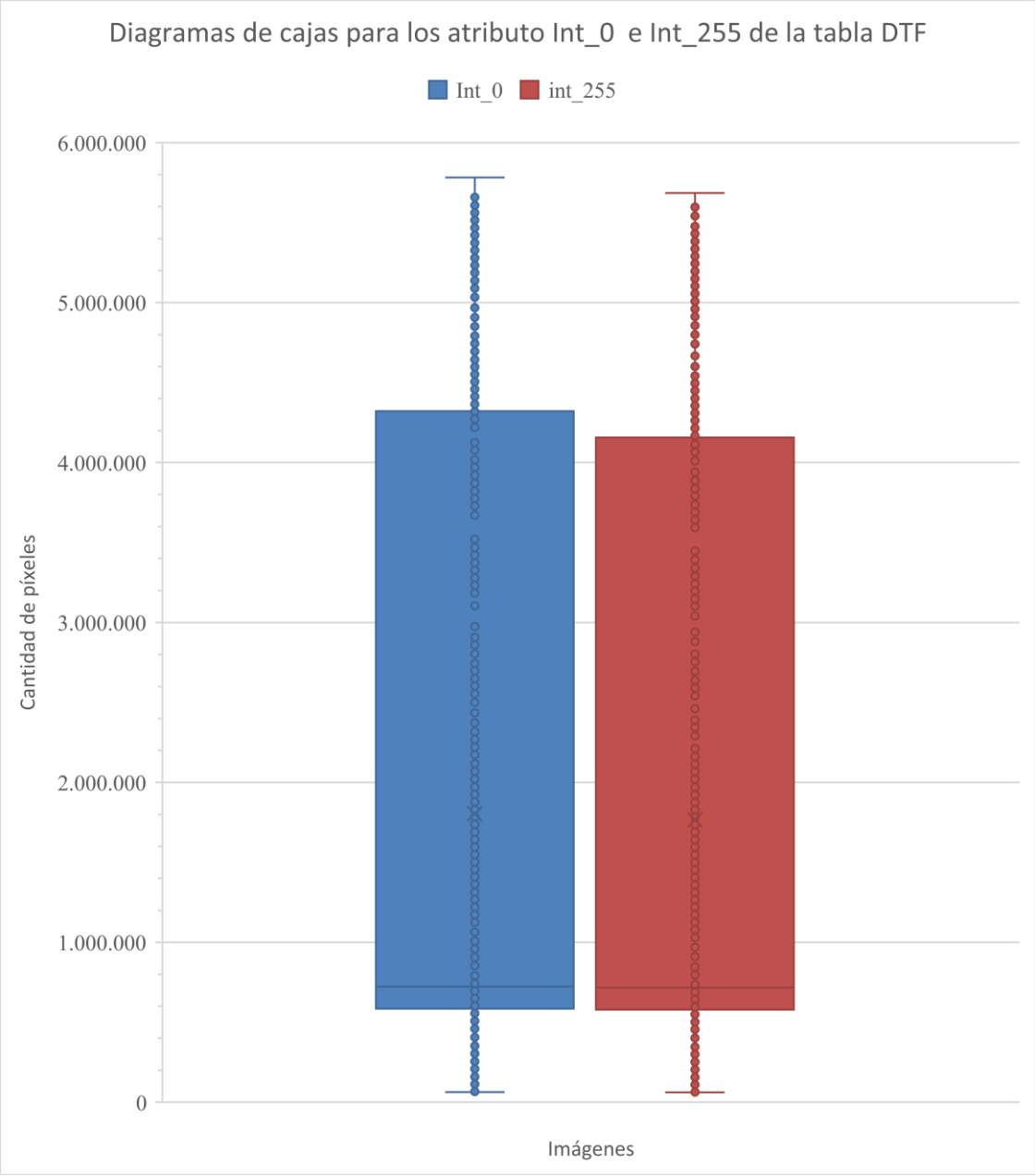


Imagen 5-4 Diagrama de caja para los atributos Int_0 e Int_255 de la tabla DTF

El diagrama de caja para la cantidad de píxeles negros (Int_0) de la tabla Integral muestra la presencia de valores muy alejados que deberán ser analizados para detectar anomalías. El resto de los valores se encuentran entre los valores 2.037 y 251.383.

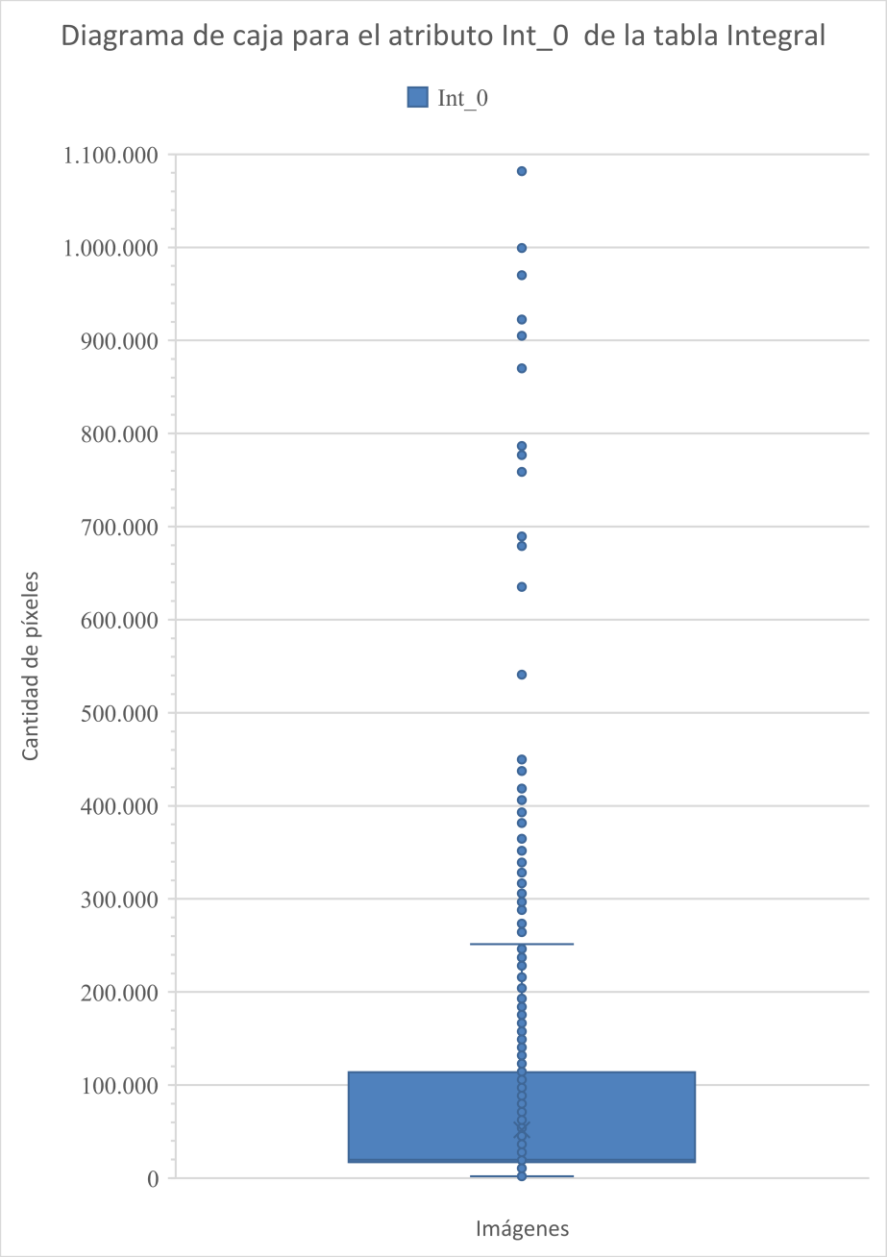


Imagen 5-5 Diagrama de caja para el atributo Int_0 de la tabla Integral

El diagrama de cajas para los valores que contabilizan la cantidad de píxeles blancos (Int_255) de la tabla Integral se encuentra representado en la Imagen 5-6. Los valores muy alejados se encuentran en el intervalo que va desde el valor 45 al 2732. El resto de los datos se encuentran distribuidos entre los valores 5 y 43. La mayor cantidad de datos se encuentran concentrados en un intervalo pequeño, existiendo poca variación de los mismos. Este atributo deberá ser analizado para determinar si es útil o no para el proceso de elaboración de un modelo de clasificación.

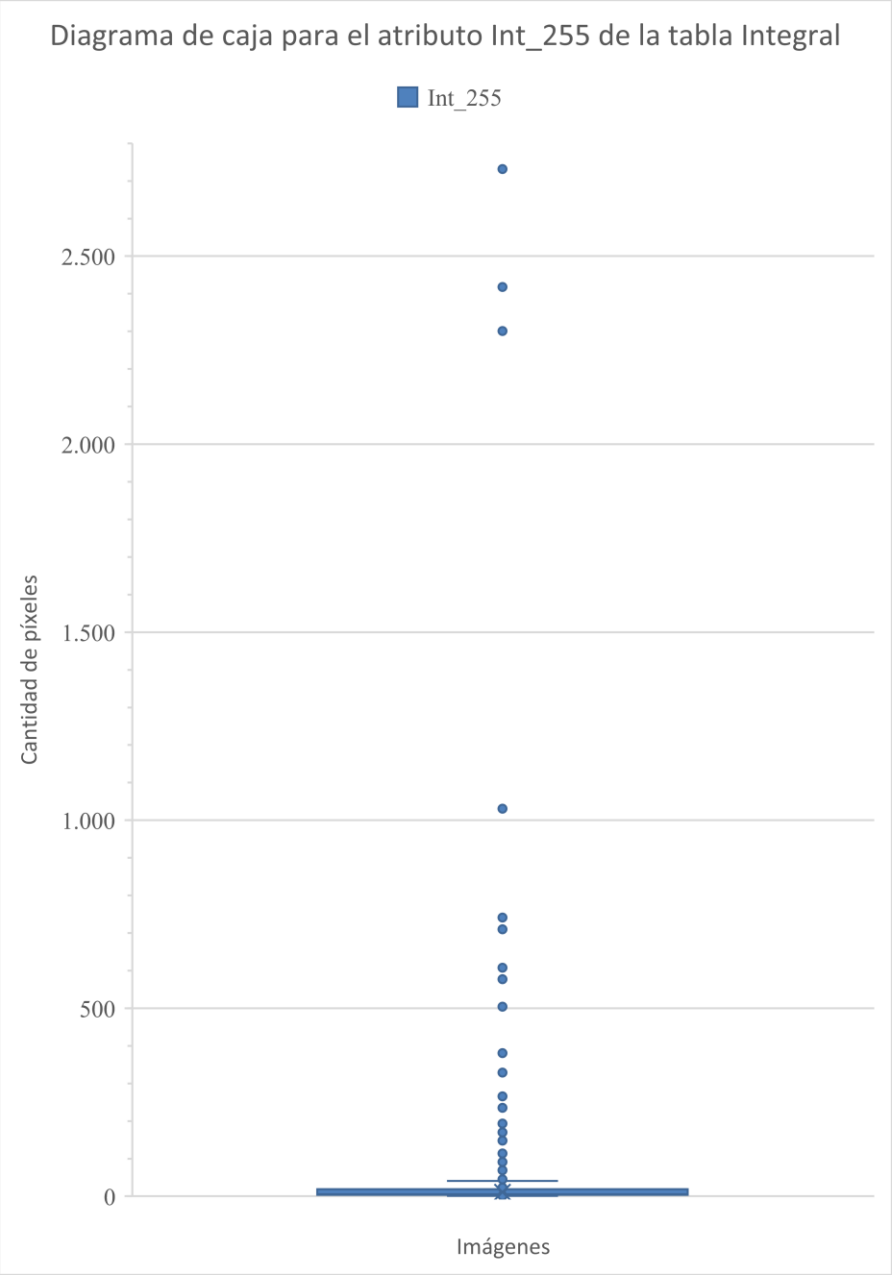


Imagen 5-6 Diagrama de caja para el atributo Int_0 de la tabla Integral

En la Imagen 5-7 se observa que los valores para la cantidad de píxeles negros se encuentran en su mayoría en el intervalo que va desde 0 a 140.422. En este caso se observa una concentración de valores muy alejados que van desde 140.422 hasta 1.510.002. A partir de este último valor existen puntos con una mayor distancia entre ellos. Estos valores deberán ser evaluados para detectar si se tratan de anomalías.

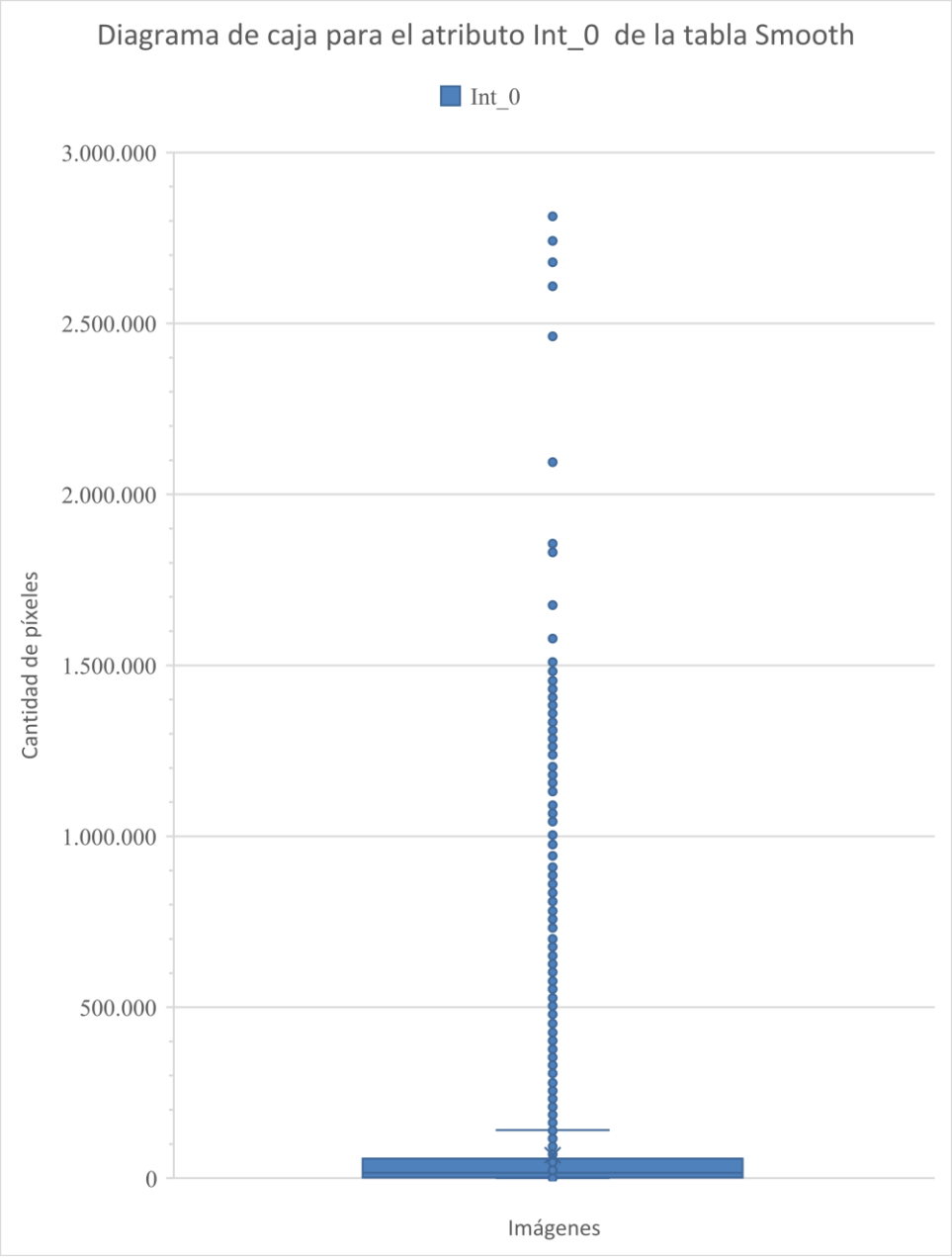


Imagen 5-7 Diagrama de caja para el atributo Int_0 de la tabla Smooth

El atributo que contabiliza la cantidad de píxeles blancos de la tabla *smooth* no presenta valores muy alejados (Imagen 5-8). Todos los valores se encuentran agrupados entre los valores 95.524 y 11.010.272.

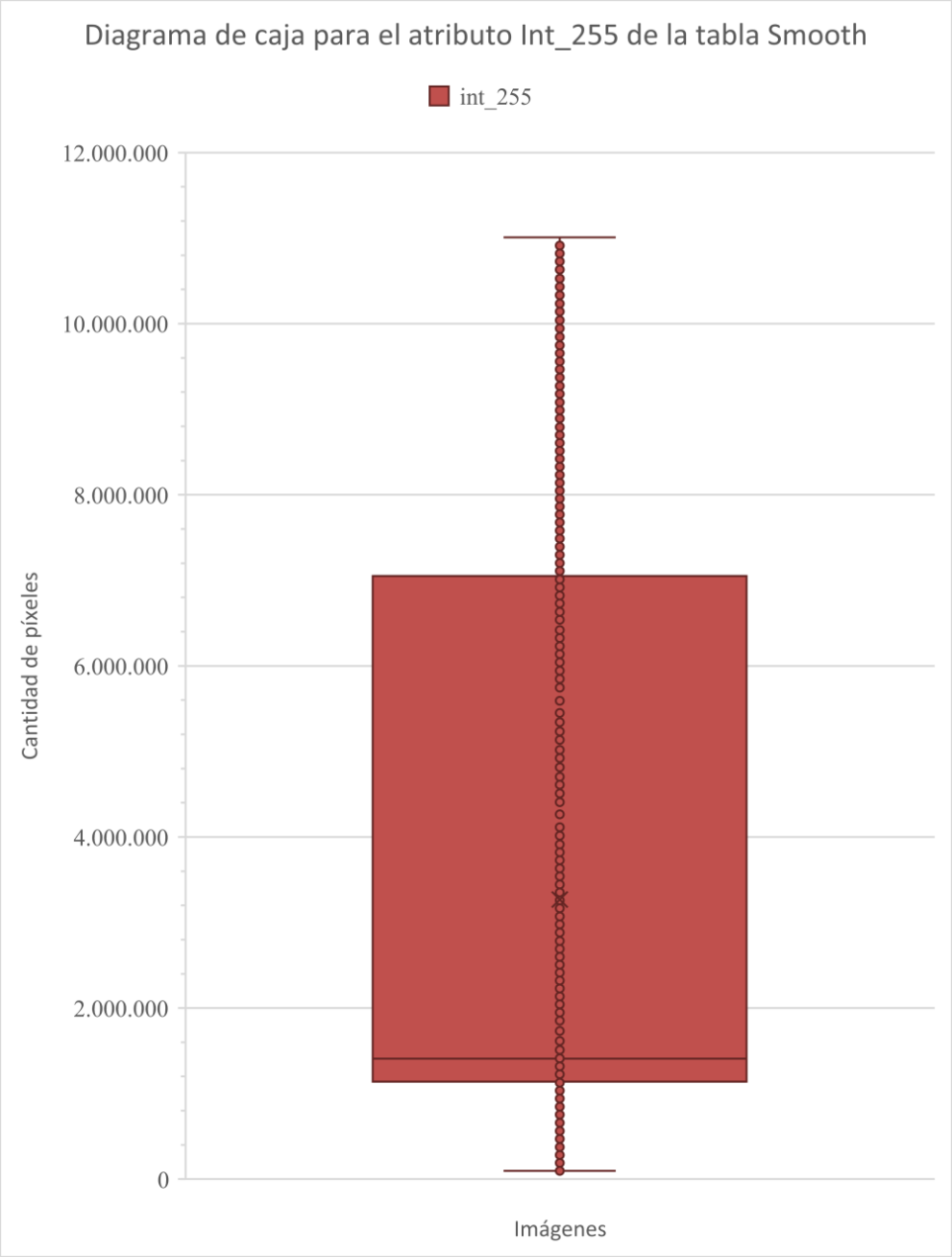


Imagen 5-8 Diagrama de caja para el atributo Int_255 de la tabla *Smooth*

Las diez clases con mayor cantidad de documentos corresponden al 57% del total. Estas imágenes corresponden a los documentos más utilizados. Aquellas que poseen cantidades menores están asociadas a procesos que se realizan con menor frecuencia o de forma esporádica.

5.2.4 Verificación de la calidad de los datos

Debido a que previamente no se contaba con información sobre las características de las imágenes, fue necesario realizar el proceso de extracción de la misma. Es por esto que, en el conjunto de datos obtenidos, no se presentan valores faltantes, inconsistencias de codificación o tipos de datos.

Los errores presentes en las imágenes se deben a un defecto en el proceso de digitalización de las mismas. Se analizan los valores de los atributos de las imágenes con el objetivo de detectar la presencia de valores anómalos (*outliers*). Estos valores podrían afectar negativamente los resultados de los procesos de minería de datos, por lo cual deberían ser excluidos e informados para su corrección.

Un error en el proceso de digitalización de los documentos puede generar imágenes en blanco o con una excesiva cantidad de píxeles negros que hace imposible reconocer el contenido de la imagen. Estos casos son detectados analizando los valores anómalos en los conjuntos de datos que contabilizan la cantidad de píxeles blanco y negros de cada imagen. Se realiza una inspección visual de las imágenes que presentan valores mínimos y máximos alejados de resto de los valores. Imágenes que carecen de píxeles negro y que poseen una cantidad elevada de píxeles blancos resultaron ser imágenes en blanco. Por el contrario, imágenes con una gran cantidad de píxeles negros y escasos o ningún píxel blanco, se tratan de imágenes ilegibles resultantes de un incorrecto proceso de digitalización.

Valores extremos en el alto y ancho de una imagen también pueden representar anomalías. Se analizaron las imágenes con valores más bajos y se encontraron tres, cuya digitalización se realizó de forma parcial, es decir no se digitalizó el documento en su totalidad. Estos casos son excluidos del lote de imágenes.

5.3 Preparación de los datos

Para proceder al modelado es necesario contar con una vista minable de los datos, es decir, disponer de los datos de tal forma que puedan ser aplicadas las técnicas de minerías de datos.

5.3.1 Selección de los datos

Durante el proceso de clasificación manual se encontraron documentos que no pertenecían a ninguna clase en particular. Estos documentos se fueron agrupando en la clase C25. Debido a que no guardan similitudes entre sí, son excluidos. También se encontraron imágenes que poseían alto contenido de tonos oscuros o tonos claros, lo cual no permitió identificar de qué documentos se trataban. Estas imágenes fueron agrupadas en la clase C14, las cuales serán excluidas.

5.3.2 Integración de los datos

Debido a que la información obtenida para cada imagen, se encuentra distribuida entre distintas tablas en la base de datos, se procede a realizar la integración de la misma para facilitar su análisis. Los datos serán unificados en una única tabla. Los atributos de la tabla resultante se muestran a continuación:

Atributo	Descripción
Imagen	Nombre la imagen
Alto	Alto de la imagen expresado en píxeles
Ancho	Ancho de la imagen expresado en píxeles
Ori0	Cantidad de píxeles negros de la imagen original
Ori255	Cantidad de píxeles blancos de la imagen original
DTF0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método DTF
DTF255	Cantidad de píxeles blancos de la imagen resultante de la aplicación del método DTF
INT0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Integral
INT255	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Integral
SMO0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Smoothmedian
SMO255	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Smoothmedian
SOB0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Sobel
SOB255	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Sobel
Clase	Nombre de la clase a la cual pertenece la imagen

Tabla 5-7 Atributos de tabla correspondiente a los datos integrados.

Los datos serán almacenados en archivos con formato CSV, lo cual será de utilidad para permitir importarlos a RapidMiner.

5.3.3 Limpieza de datos

Los datos recolectados pueden contener errores. La causa de los datos erróneos puede ser producto de una incorrecta digitalización de los documentos. Estos documentos pueden ser detectados mediante el análisis de los valores de los atributos que lo definen, realizando la búsqueda de valores anómalos. Es posible detectar anomalías identificando valores que se encuentran muy alejados del resto de los valores del conjunto al que pertenecen. Por ejemplo, si los valores del atributo que indica la cantidad de píxeles negros en una imagen se encuentran concentrados en un rango entre 4.000 y 14.000 píxeles, una imagen con 230 píxeles puede resultar una anomalía. Debido a que en este caso las anomalías corresponden a errores, estos valores pueden afectar negativamente al desarrollo del modelo. Es por eso que estos valores deberán ser excluidos.

Teniendo en cuenta los hallazgos de posibles anomalías realizados durante la fase de exploración de datos en la sección 5.2.3, se efectuó una inspección visual de aquellas imágenes. Las imágenes que resultaban ilegibles (por exceso de tonos claros u oscuros) fueron descartadas.

Debido a la elevada cantidad de registros que se poseen, para continuar con la tarea de detección de valores anómalos se emplearán las herramientas que ofrece RapidMiner.

El operador *KNN Global Anomaly Score* (Imagen 5-9) asigna un puntaje de anomalía a cada uno de los ejemplos basándose en el algoritmo de N vecinos más cercanos. Mientras más alto es el puntaje asignado, más anómalo es el ejemplo. En la Imagen 5-10 se puede observar la parametrización del operador.

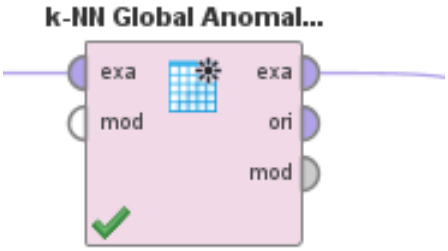


Imagen 5-9 Operador *K-NN Global Anomaly* para detección de anomalías

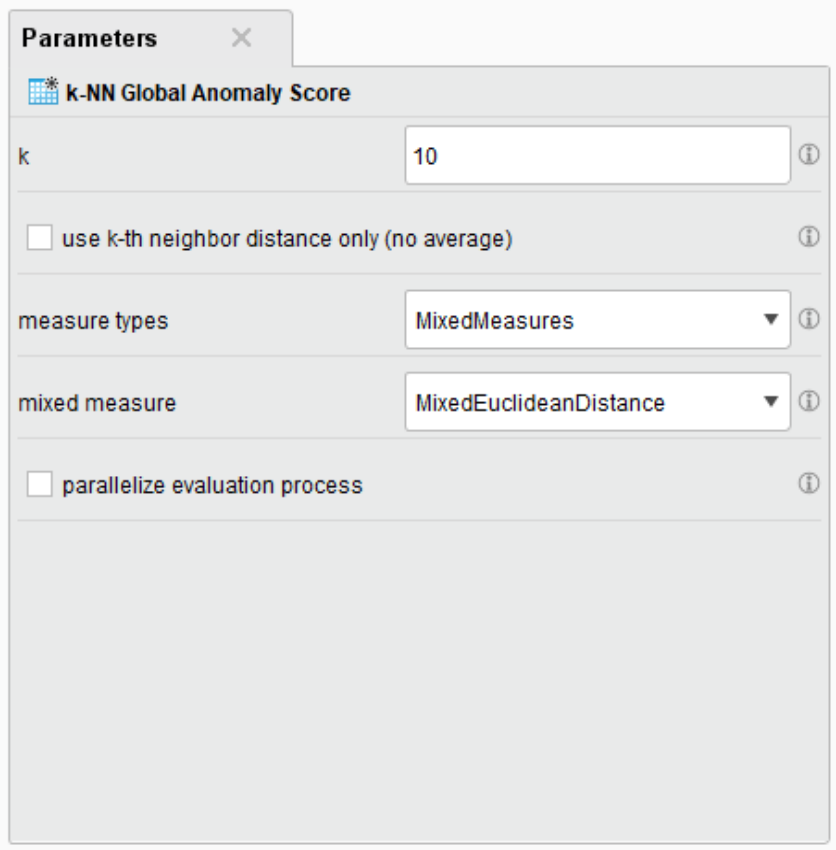


Imagen 5-10 Parametrización del operador KNN Global Anomaly Score

Como resultado se obtienen todos los ejemplos con su respectivo puntaje. El menor puntaje tiene un valor de 1.088,39 y el mayor 3.010.807,12. El promedio de estos valores es 44.052,14 y la desviación estándar tiene un valor de 87.468,26. Para realizar el análisis de los resultados obtenidos se hace uso de un diagrama de cajas. Esta herramienta permite realizar la exploración de los datos para conocer su estructura y dispersión. En la Imagen 5-11 se muestra el diagrama de cajas para el lote de datos y en la Tabla 5-8 se muestran los valores para cada una de las marcas del diagrama.

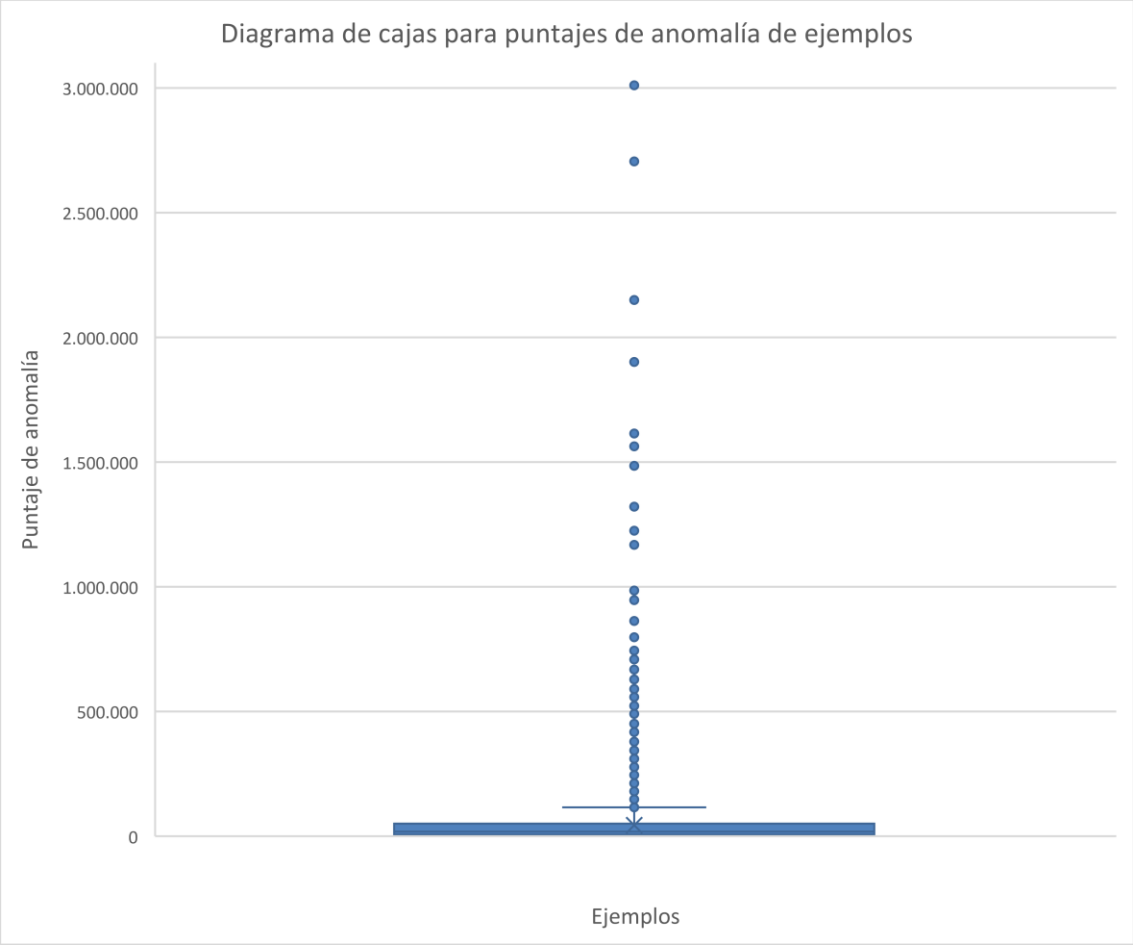


Imagen 5-11 Diagrama de caja para puntajes de anamalía de ejemplos

Parámetro	Valor
Valor alejado	3.010.807,11
Máximo	115.974,00
Tercer cuartil	50.653,25
Media	18.212,62
Primer Cuartil	7.090,23
Mínimo	1.088,39

Tabla 5-8 Valores de parámetros del diagrama de cajas

Los ejemplos comprendidos entre el primer cuartil y el tercer cuartil representan el 50 por ciento de los ejemplos. Los ejemplos que se encuentran entre los intervalos que van desde del mínimo y al primer cuartil y desde el tercer cuartil y hasta el máximo (las colas), representan cada uno de ellos el 25 por ciento de los datos. En este caso en particular se observan que existen valores alejados, es decir, valores que superan el valor máximo definido para el diagrama de cajas. Al encontrarse estos ejemplos muy alejados del resto de los valores son analizados para determinar si se tratan de anomalías. Para realizar una inspección visual de las imágenes correspondientes a estos ejemplos, se procede a agruparlas en un nuevo directorio. Analizando las imágenes a simple vista,

se puede apreciar que las anomalías se deben a imágenes que poseen gran cantidad de tonos oscuros o grandes cantidades de tonos claros, producto de errores cometidos en el proceso de digitalización. Por este motivo los ejemplos correspondientes a estas imágenes serán excluidos de la vista minable.

Para separar estos ejemplos se emplea el operador *Filter Examples* (Imagen 5-12), indicando que se omitan aquellos cuyo puntaje de anomalía sea mayor que 115.974, que es el valor correspondiente al máximo en el diagrama de cajas. En la Imagen 5-13 se muestra la parametrización del operador.

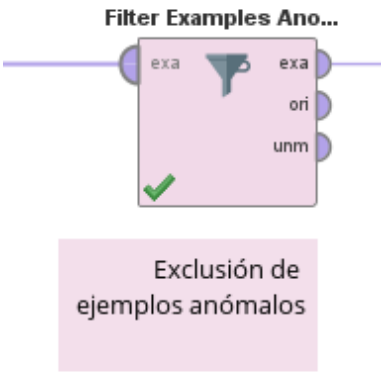


Imagen 5-12 Operador *Filter Examples* para exclusión de anomalías

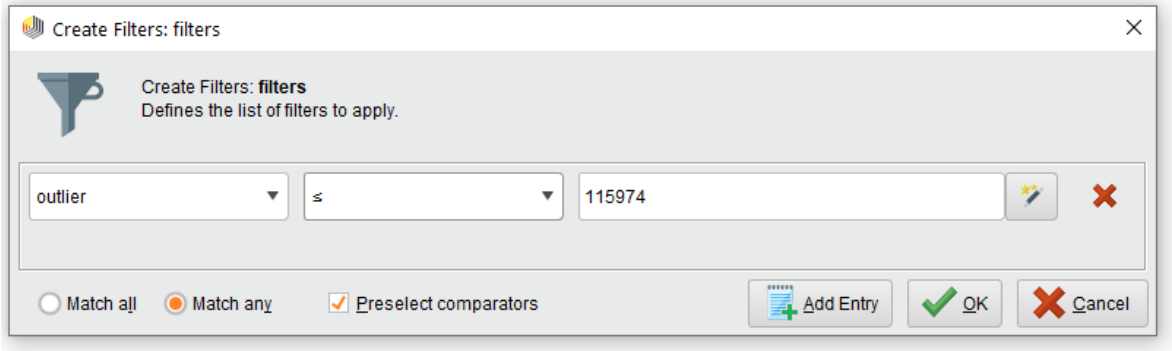


Imagen 5-13 Parametrización del módulo *Filter Examples*

Como resultado del proceso de limpieza, el conjunto de datos se redujo a una cantidad de 12.786 ejemplos. Al realizar nuevamente el conteo de elementos por clase se detectó que las clases C10, C35, C52 y C60 quedaron sin elementos. Estas clases deberían ser tomadas en cuenta para un futuro análisis de las imágenes. Esto nos deja con un total de 69 clases. El detalle de las mismas y la cantidad de elementos de cada una de ellas se muestran en el anexo B.

5.3.4 Vista minable

Al final de la fase de preparación de datos, tenemos como resultado una vista minable. Los ejemplos que conforman esta vista son aquellos resultantes de la unificación, limpieza y selección de datos. Estos serán los datos que se emplearán en siguiente fase para realizar el proceso de modelado. En el anexo C se muestra un fragmento de la tabla de los ejemplos que conforman la vista minable.

5.4 Modelado

En esta etapa del proceso de minería contamos con una vista minable, sobre la cual se aplicarán técnicas o métodos que permitirá la construcción del modelo que intenta resolver el problema planteado inicialmente. La vista minable cuenta con 12.786 ejemplos. Los atributos de estos ejemplos se detallan en la Tabla 5-9.

Atributo	Descripción	Tipo de dato
Imagen	Nombre la imagen	Texto
Alto	Alto de la imagen expresado en píxeles	Entero
Ancho	Ancho de la imagen expresado en píxeles	Entero
Ori0	Cantidad de píxeles negros de la imagen original	Entero
Ori255	Cantidad de píxeles blancos de la imagen original	Entero
DTF0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método DTF	Entero
DTF255	Cantidad de píxeles blancos de la imagen resultante de la aplicación del método DTF	Entero
INT0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Integral	Entero
INT255	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Integral	Entero
SMO0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Smoothmedian	Entero
SMO255	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Smoothmedian	Entero
SOB0	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Sobel	Entero
SOB255	Cantidad de píxeles negros de la imagen resultante de la aplicación del método Sobel	Entero
Clase	Nombre de la clase a la cual pertenece la imagen	Texto

Tabla 5-9 Atributos de la vista minable

5.4.1 Selección de técnicas de modelado

La elección de los métodos depende principalmente del tipo de tarea y de las características de los datos. El objetivo en este caso, es que dado un ejemplo X obtengamos su clase Y. Por lo cual se trata de una tarea predictiva. Al elegir las técnicas de modelado, se debe considerar que el atributo a predecir es un atributo nominal. Los métodos que aplican a los problemas de clasificación y en particular al problema que se está tratando, considerando el tipo de datos de los atributos y de la clase, son:

- ***Naive Bayes:*** es empleado para tareas de categorización. Este clasificador asume que, dada una clase de un ejemplo, los valores de los atributos son independientes entre sí.
- ***Decision tree***: esta técnica puede ser empleada para clasificación y regresión. El resultado es un árbol de decisión donde cada nodo representa una regla de decisión.
- ***Random Tree*****:** es similar a la técnica *Decision Tree*, pero en cada bifurcación solo analiza un subconjunto de datos tomado de forma aleatoria.
- ***Rule Induction*****:** Crea reglas por cada clase las cuales se van ajustando hasta cumplir un determinado criterio.
- ***K Nierest Neighbors*****:** es usado para tareas de clasificación o regresión. El algoritmo se basa en comparar un ejemplo sin etiquetar con sus k vecinos más cercanos. La clase del ejemplo será la misma que la de sus vecinos.

5.4.2 Diseño de las pruebas

Para determinar cuál es la técnica de modelado más adecuada, se deben evaluar los modelos que se obtendrán y comparar los resultados. Para realizar esta tarea se aplicará el método de evaluación validación cruzada. Este método consiste en dividir el conjunto de ejemplos en K subconjuntos. Los ejemplos pertenecientes a k-1 conjuntos se emplean para entrenar el modelo y el conjunto restante es usado para prueba. Este proceso se repite k veces y los indicadores de calidad del modelo son el promedio de las k ejecuciones.

La evaluación se implementará mediante el operador *Cross Validation* en RapidMiner (Imagen 5-14). Este operador estima cuán preciso es un modelo. En la Imagen 5-15 se muestra la parametrización del operador.

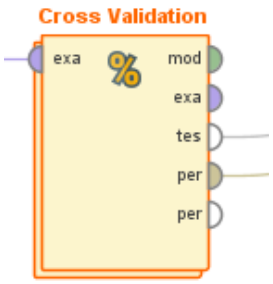


Imagen 5-14 Operador *Cross Validation* en RapidMiner

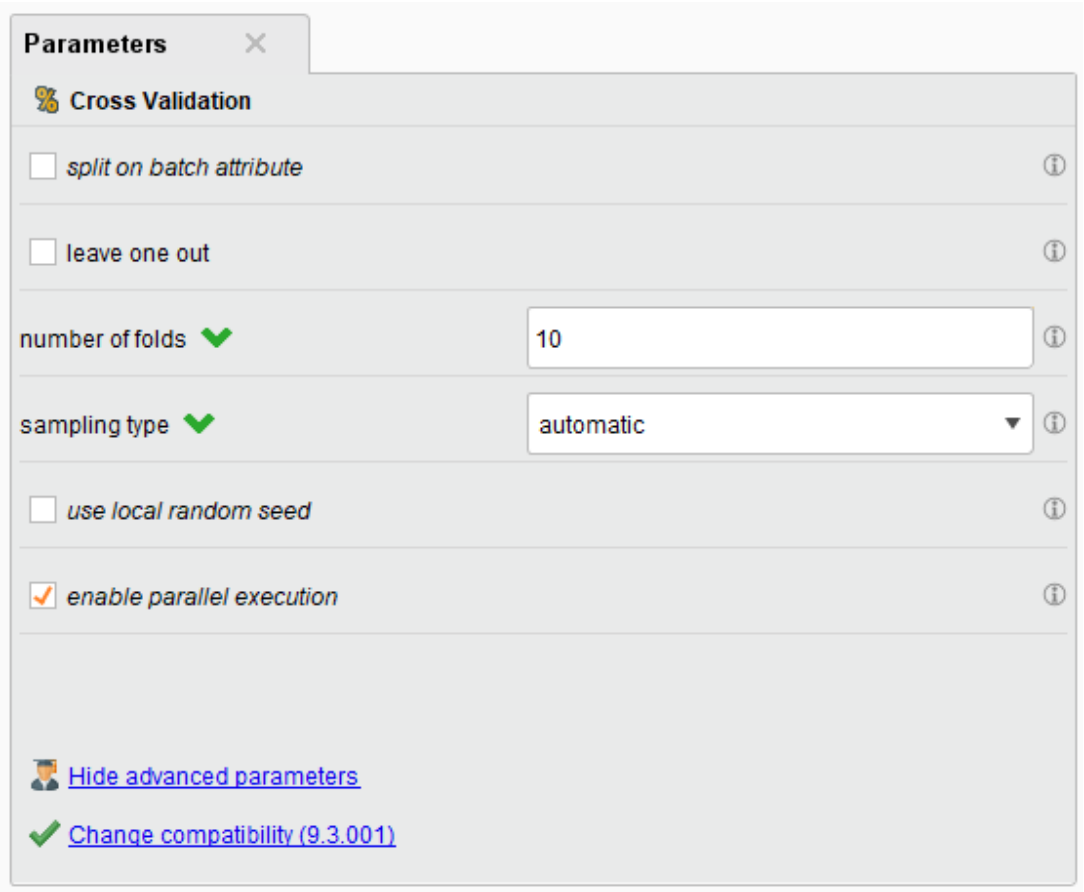


Imagen 5-15 Parametrización del operador *Cross Validation*

Split on batch attribute: esta opción desmarcada indica que la partición de los ejemplos en los subconjuntos se realiza en forma aleatoria.

Number of folds (número de subconjuntos): determina la cantidad de subconjuntos en los que se agruparán los ejemplos, en este caso 10.

Sampling type (tipo de muestreo): Automático. Selecciona por defecto el muestreo estratificado. Asegura que la distribución de la clase del subconjunto sea la misma que la de la totalidad de los ejemplos.

El operador cuenta con dos subprocesos; uno de entrenamiento del modelo y uno de prueba. En la primera etapa el algoritmo elegido es entrenado y obtenemos como resultado un modelo, que se pasa al proceso de evaluación. Para ser evaluado, el modelo obtenido debe ser aplicado a un conjunto de datos no etiquetados y medir el grado de precisión de las predicciones realizadas. El resultado obtenido es aplicado al operador Performance, que en este caso nos devuelve los indicadores *accuracy*, *class prescision* y *class recall*. Ver Imagen 5-16.

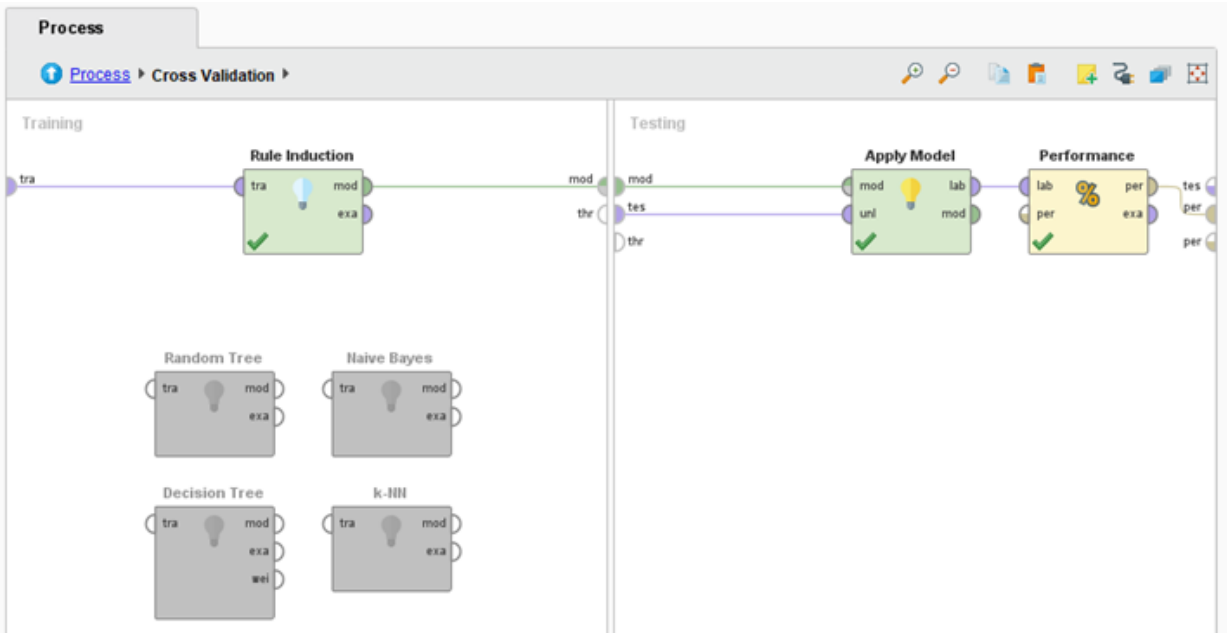


Imagen 5-16 Proceso *Cross Validation* en RapidMiner

Accuracy o exactitud es la relación entre la cantidad de predicciones que el modelo realizó correctamente y la cantidad total de predicciones. Este indicador nos da una visión global de qué tan preciso es el modelo, pero presenta un problema, no permite identificar si existen desbalances en la precisión de predicción de cada clase. Esto puede ser observado mediante los indicadores *class recall* y *class prescision*. *Class recall* o exhaustividad la cual indica la cantidad de casos en los cuales se predijo correctamente la clase en relación con la cantidad de ejemplos que pertenecen a dicha clase. Y por último el indicador *Class prescision* o precisión, que representa la cantidad de aciertos de predicciones para una clase sobre el total de predicciones realizadas para esa clase. (Google Developers, 2020)

5.4.3 Desarrollo del modelo

Para la construcción del modelo se empleó la herramienta de minería de datos RapidMiner. Los procesos de exploración y preparación de datos, modelado y validación de los modelos obtenidos se implementan a través de operadores (módulos) que realizan una tarea específica. Los operadores generalmente tienen entradas y salidas, parámetros para ser configurados y pueden ser conectados entre sí. Ver Imagen 5-17.

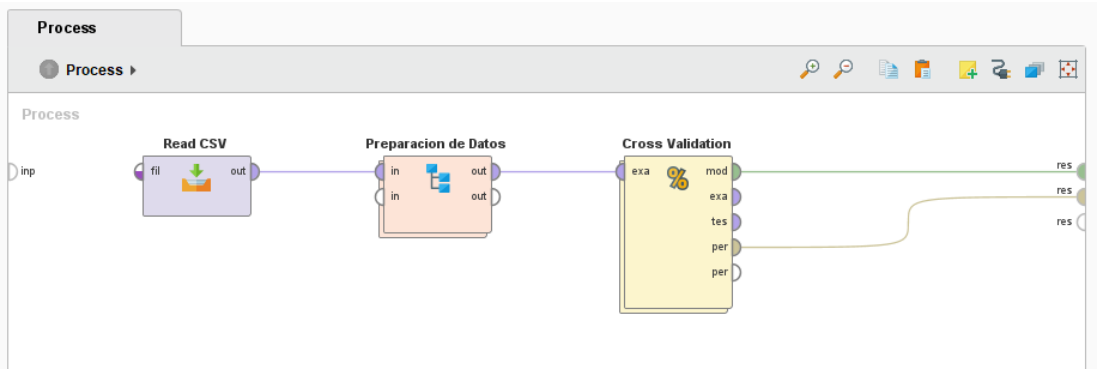


Imagen 5-17 Proceso de preparación de datos, modelado y validación en RapidMiner

El operador *Read CSV* obtiene los datos desde un archivo .csv. El siguiente módulo, Preparación de los datos, realiza el proceso de limpieza y selección de los datos y su salida es la vista minable. Este proceso se encuentra detallado en la sección 5.3. Los ejemplos obtenidos se envían como entrada al operador *Cross Validation*.

El operador *Cross Validation* (Imagen 5-18) cuenta con dos subprocesos. El subproceso de entrenamiento (*Training*) y el subproceso de prueba (*Testing*). En el primero se pueden observar los módulos para cada una de las técnicas de modelado definidas en la sección 5.4.1. Cada uno de estos módulos recibe como entrada los ejemplos de la vista minable y devuelve como resultado un modelo. El modelo obtenido se convierte en entrada del subproceso de prueba. El operador *Apply Model* recibe como entrada el modelo y el grupo de ejemplos no etiquetados (sin clase). El modelo es aplicado a los datos entrantes para predecir la clase de cada ejemplo, es decir, obtenemos los ejemplos con las clases determinadas por el modelo. A su vez, este resultado se convierte en entrada del operador *Performance*. Este operador determina los indicadores adecuados para cada modelo. Estos valores serán usados en la etapa de evaluación de modelos.

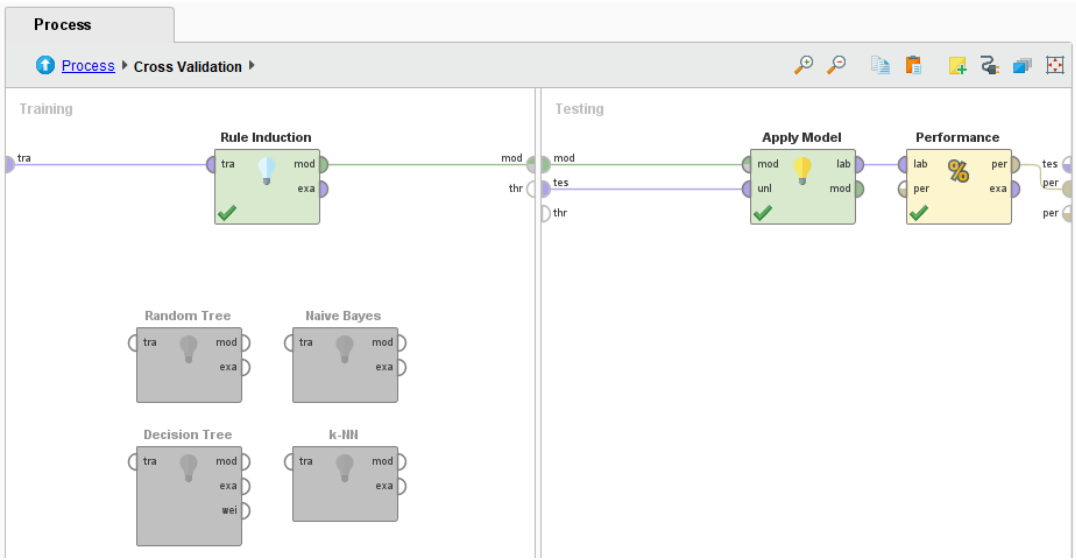


Imagen 5-18 Proceso *Cross Validation* en RapidMiner

A continuación, se describen cada uno de los operadores correspondientes a las técnicas de modelado seleccionadas, la parametrización, la descripción del modelo y los resultados de la evaluación del desempeño de cada modelo obtenido. El detalle de los resultados obtenidos para los indicadores *class precision* y *class recall* de cada modelo se encuentran en el anexo D.

- ***Naive Bayes***

Este operador genera un modelo de clasificación *Naive Bayes*. Su principal característica es que asume independencia entre los atributos. Es un modelo con bajo costo computacional. La parametrización del operador se muestra en la Imagen 5-19. Se tilda la opción "*Laplace correction*" que evita la ocurrencia de probabilidades con valor cero cuando un valor de un atributo no tiene ocurrencias en una determinada clase.

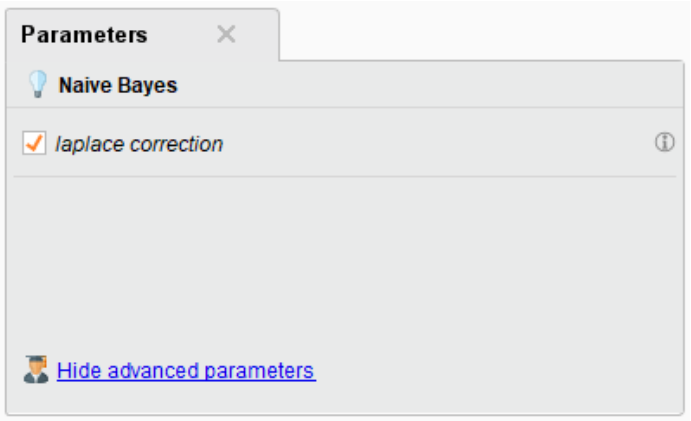


Imagen 5-19 Parametrización del operador *Naive Bayes* en RapidMiner

El modelo obtenido presenta un grado de exactitud del 54.92%con un error +/- 2.09%. Se puede observar que para los valores de *class recall* obtenidos, 19 clases tienen un valor mayor o igual a 80% y 17 clases con valor mayor a 50% pero menor a 80%. Para las restantes 34 clases el valor de *class recall* menor o igual al 50%. Para el valor de *class precision*, solo 12 clases poseen un valor mayor o igual a 80%. Del resto de las clases, 14 tienen un valor mayor al 50% y menor al 80% y 43 tiene un valor menor o igual al 50%.

- ***Decision tree***

El resultado de este operador es un árbol que puede emplearse para la clasificación. Cada nodo del árbol representa una regla de decisión para un atributo en particular. En la Imagen 5-20 se muestra la parametrización del operador.

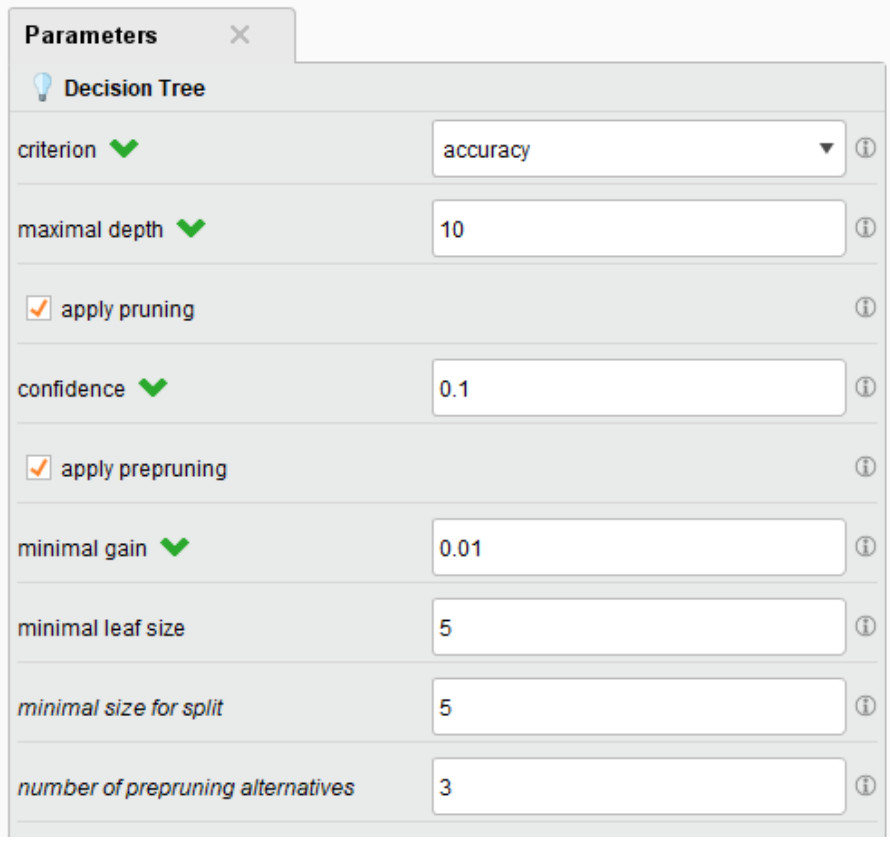


Imagen 5-20 Parámetros para el operador Decision Tree en RapidMiner

Como criterio (*criterion*) se selecciona la opción *accuracy*, que elige el atributo que maximiza la precisión total del árbol. La profundidad máxima del árbol se establece en 10. En este caso se especifica que realizará poda y pre poda. Los valores para el tamaño mínimo de hoja y tamaño mínimo para división se establecen según la menor cantidad de elementos que posee una clase, que es 5.

El resultado que se obtuvo es un árbol con 342 reglas y un nivel de exactitud de 81.56% con un error de +/- 0.89%. Las reglas generadas se encuentran detalladas en el anexo E. Al analizar el valor *class recall* se observa que 26 clases poseen un valor mayor o igual al 80%, 15 clases tienen un valor mayor a 50% y menor al 80% y las restantes 28 tienen un valor menor o igual al 50%. Realizando el análisis del valor *class precision*, se observan que 24 clases poseen un valor mayor o igual al 80%. Del resto de las clases 23 poseen un valor mayor a 50% y menor que 80%. Las restantes 22 clases tiene un valor menor al 50%.

- ***Random Tree***

Este operador construye un árbol de decisión al igual que el operador "*Decision Tree*", pero se diferencia en que al momento de realizar una bifurcación solo toma en cuenta un grupo de atributos seleccionados aleatoriamente. En la Imagen 5-21 se muestran los parámetros seleccionados.

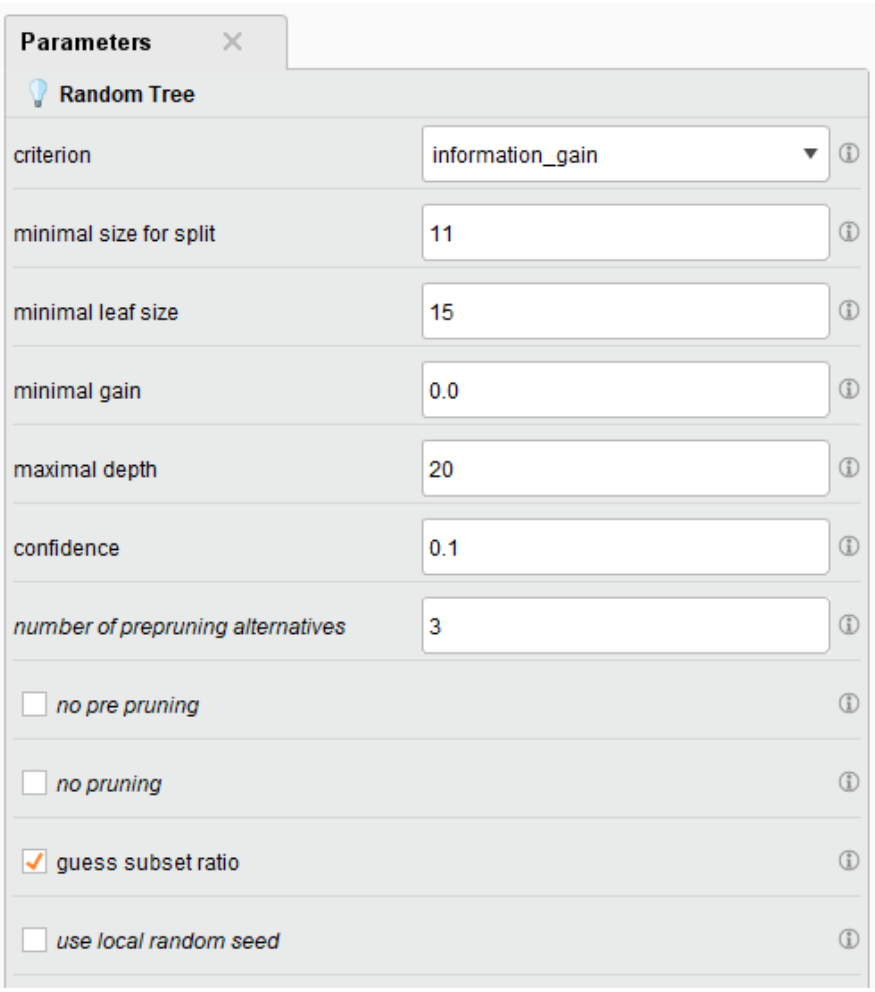


Imagen 5-21 Parámetros para el operador *Random Tree* en RapidMiner

La mejor exactitud obtenida para este modelo fue del 62.51% con un error de +/- 5.29%. La precisión del modelo en este caso es muy baja con un valor de error alto. Del total de clases, solo 5 tienen un valor mayor o igual al 80% para el indicador *class recall* y solo 3 para el indicador *class precision*. Considerando que se trabaja con un total de 69 clases, el modelo generado por el operador *Radom Tree* no es bueno realizando predicciones.

- ***Rule Induction***

El algoritmo que ejecuta este operador (RIPPER) ordena las clases de menor a mayor según la cantidad de elementos que tiene cada clase. Selecciona en ese orden cada una de las clases. Para cada clase crea una regla que se va ajustando al conjunto de ejemplos. En la Imagen 5-22 se muestra la parametrización del operador.

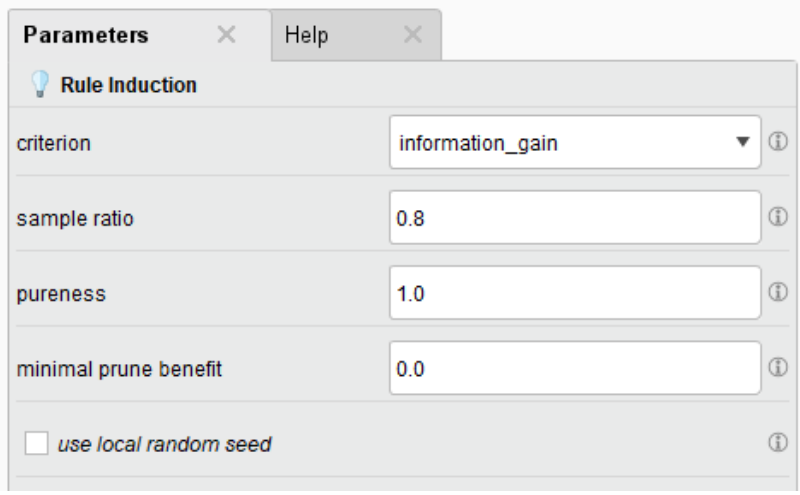


Imagen 5-22 Parámetros para el operador *Random Tree* en RapidMiner

El nivel de exactitud obtenido es del 80.38% con un error de +/- 0.97%. El modelo obtenido por este operador es un conjunto de 208 reglas. El detalle de las mismas se encuentra en el anexo D. Analizando los valores de *class recall* y *class precision*, se observa que 19 clases poseen un valor mayor o igual al 80% de *class recall*, 16 clases tienen un valor mayor a 50% y menor a 80% y 34 poseen un valor menor o igual al 50%. Respecto del indicador *class precision*, se observan que 19 clases poseen un valor mayor o igual al 80%, 23 clases poseen un valor mayor a 50% y menor que 80% y las 27 clases restantes tiene un valor menor al 50%.

- ***K Nierest Neighbors***

El operador k-NN ejecuta el algoritmo *K-Nierest Neighbors* (k vecinos más cercanos). Este algotirmo toma un ejemplo sin etiquetar y lo compara con los ejemplos etiquetados más cercanos para determinar su clase. La distancia del ejemplo se calcula en base a la distancia de los valores de cada uno de los atributos. En la Imagen 5-23 se muestra la parametrización del operador.

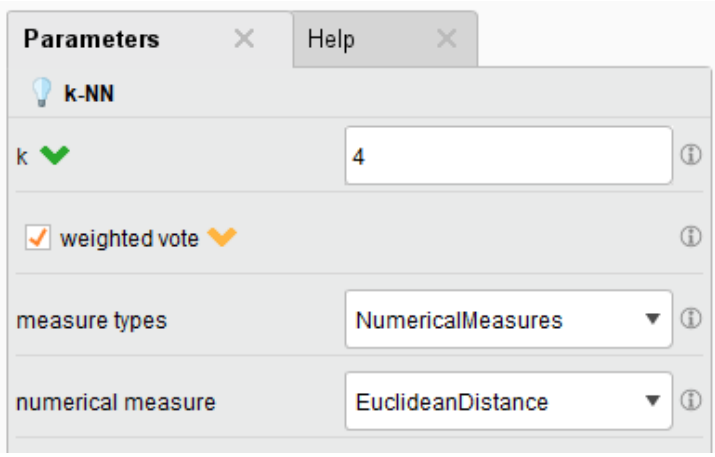


Imagen 5-23 Parámetros para el operador k-NN en RapidMiner

Los parámetros fueron seleccionados acorde al tipo de datos. El valor de K representa la cantidad de vecinos más cercanos que se deben tener en cuenta para asignar la clase al ejemplo sin etiquetar. En este caso el valor de K es 4, ya que es valor que maximiza la precisión del modelo. Debido a que los valores de los atributos son numéricos, al parámetro *measure types* (tipo de medidas) se le asigna el valor *NumericalMeasures* (medidas numéricas).

El valor de la exactitud del modelo obtenido es de 84.47%, con un error +/- 0.87%. Del total de 69 clases, 30 de ellas cuentan con un valor de *class recall* mayor o igual al 80%. Para el resto de las clases, 16 de ellas tienen un valor mayor al 50% y las 23 clases restantes tienen un valor menor o igual al 50%. Al analizar los valores de *class precision* observa que para 27 clases la precisión es igual o mayor a 80%, 27 clases tienen una precisión menor a 80% y mayor a 50%. Las restantes 15 clases tienen una precisión menor o igual al 50%.

5.5 Evaluación

5.5.1 Evaluación de los resultados de la minería de datos con el criterio de éxito

Por cada uno de los modelos obtenidos durante la fase de modelado (sección 5.4) se obtuvieron parámetros que permiten determinar el nivel de precisión. En la Tabla 5-10 se puede observar un resumen de los valores para los parámetros.

1Modelos	*Accuracy*	Error	Cantidad de clases *class recall* > 80%	Cantidad de clases *class presicion* > 80%
K Nierest Neighbors	84.47%	+/- 0.87%	30	27
Decision tree	81.56%	+/- 0.89%	26	23
Rule Induction	80.38%	+/- 0.97%	19	19
Naive Bayes	54.92%	+/- 2.09%	19	14
Random Tree	62.51%	+/- 5.29%	5	3

Tabla 5-10 Niveles de precisión de los modelos obtenidos

Tres de los modelos obtenidos cumplen con el objetivo de minería de datos, que es obtener un modelo que permita realizar la clasificación de las imágenes de las digitalizaciones de los legajos de los clientes con una precisión mayor al 80%. El modelo que presenta el mayor nivel de *accuracy* (relación entre la cantidad de predicciones correctas y la cantidad total de ejemplos) es el obtenido del algoritmo *K Nierest Neighbors*. Adicionalmente, este modelo tiene la mayor cantidad de clases con valores mayores al 80% para los indicadores *class recall* y *class precision*. En este caso, el modelo no genera un conjunto de reglas. La predicción de la clase se realiza cada vez que se ejecuta el proceso evaluando la distancia de cada ejemplo a sus vecinos más cercanos. En segundo lugar, se encuentra el modelo obtenido de *Decision tree*. El error en ambos casos no supera un 1 % por lo cual la variación de la precisión es mínima para ambos modelos. La ventaja que presenta este modelo es que como resultado obtenemos un conjunto de reglas del tipo si-entonces en forma de árbol, las cuales pueden ser codificadas de forma simple e integradas como una nueva funcionalidad al sistema ya existente. Y el tercer modelo que supera el 80% de precisión es el

generado por el algoritmo *Rule Induction.* Al igual que *Decision Tree* el modelo consiste en un conjunto de resultados del tipo si-entonces.

Analizando los tres modelos mencionados se selecciona como mejor modelo el obtenido de *Decision Tree*. Los aspectos tenidos en cuenta para elegir este modelo son:

- Cumplimiento con el objetivo de minería de datos ya que posee una precisión mayor al 80%.
- Al poder expresarse como un conjunto de reglas del tipo si-entonces simplifica la tarea de codificación para ser integrado al sistema existente.
- El modelo se genera una única vez, no como en el caso de *K Nierest Neighbors* que se genera en cada ejecución de clasificación de las imágenes. El único motivo por el cual debería crearse nuevamente el modelo de *Decision Tree* es ante el ingreso de documentos de una nueva clase. Esto solo ocurriría si se produjera un cambio en un proceso de negocio que obligue a cambiar el tipo de documentación solicitada a los clientes o la modificación de un formulario para cumplimentar con nuevos requerimientos. La probabilidad de ocurrencia de ambos casos es muy baja y ocurre muy esporádicamente.

5.5.2 Determinación de las siguientes acciones

Una vez obtenidos los modelos y realizada la evaluación de los mismos, se toma la decisión de implementar la solución, considerando que se encontró un modelo (*Decision tree*) que cumple con el objetivo de minería de datos y al expresarse como un conjunto de reglas la integración al software existente resultará simple de implementar.

El siguiente paso es la planificación de la implementación de un nuevo módulo que realice la clasificación de las imágenes históricas y de las que ingresan diariamente. Este módulo deberá integrarse al software existente de consulta de legajos. Además, se deberán realizar los cambios necesarios para que la información sobre la clase a la que pertenece una imagen se encuentre visible y pueda ser empleada por los usuarios. Esto implicará cambios en la base de datos para adicionar la información de la clase de cada documento y cambios en la interfaz en software de consulta de legajo de clientes para poder visualizar la clase. La realización de estos cambios queda fuera del alcance de proyecto.

5.6 Desarrollo

En esta última etapa del proceso contamos con un modelo que cumple con los objetivos de minería de datos. Las siguientes acciones serán determinar la forma de distribución, en este caso, los pasos a seguir para su implementación y definir cómo se controlará y mantendrá el resultado obtenido.

5.6.1 Plan de desarrollo

El modelo obtenido del algoritmo *Decision Tree* será entregado al departamento de desarrollo como un conjunto de reglas si-entonces. RapidMiner permite convertir el modelo obtenido con formato de árbol a reglas si-entonces. En la Imagen 5-24 se muestra un conjunto de las reglas obtenidas en formato de texto plano. Esta forma de presentar las reglas simplificará la tarea a los programadores para traducirlas a un lenguaje de programación.

```
if Alto > 427.500 and Alto > 1687.500 and Alto > 2332 and SOB255 > 291640 and Ori0 > 791060.500 and Alto > 4116.500 then C84
if Alto > 427.500 and Alto > 1687.500 and Alto > 2332 and SOB255 > 291640 and Ori0 > 791060.500 and Alto ≤ 4116.500 then C69
if Alto > 427.500 and Alto > 1687.500 and Alto > 2332 and SOB255 ≤ 291640 and Alto ≤ 2677 and SMO0 > 73028 and Ancho > 2569 and Alto ≤ 2547 then C15
if Alto > 427.500 and Alto > 1687.500 and Alto > 2332 and SOB255 ≤ 291640 and Alto ≤ 2677 and SMO0 > 73028 and Ancho ≤ 2569 then C57
if Alto > 427.500 and Alto > 1687.500 and Alto > 2332 and SOB255 ≤ 291640 and Alto ≤ 2677 and SMO0 ≤ 73028 and Alto > 2504 then C54
if Alto > 427.500 and Alto > 1687.500 and Alto > 2332 and SOB255 ≤ 291640 and Alto ≤ 2677 and SMO0 ≤ 73028 and Alto ≤ 2504 then C15
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 > 3308324.500 and Ancho > 2029.500 and Alto > 1987.500 then C61
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 > 3308324.500 and Ancho > 2029.500 and Alto ≤ 1987.500 then C18
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 > 3308324.500 and Ancho ≤ 2029.500 and INT255 > 15.500 and Ori0 ≤ 213433 then C17
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 > 3308324.500 and Ancho ≤ 2029.500 and INT255 ≤ 15.500 and Ori0 > 137215.500 then C61
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 > 3308324.500 and Ancho ≤ 2029.500 and INT255 ≤ 15.500 and Ori0 ≤ 137215.500 then C16
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 ≤ 3308324.500 and Ancho > 1938 and Alto > 1807 and INT0 > 61719.500 and INT0 > 73789 then C61
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 ≤ 3308324.500 and Ancho > 1938 and Alto > 1807 and INT0 > 61719.500 and INT0 ≤ 73789 then C17
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 ≤ 3308324.500 and Ancho > 1938 and Alto > 1807 and INT0 ≤ 61719.500 then C61
if Alto > 427.500 and Alto > 1687.500 and Alto ≤ 2332 and SOB0 ≤ 3308324.500 and Ancho > 1938 and Alto ≤ 1807 then C71
```

Imagen 5-24 Ejemplo de reglas del modelo de clasificación de imágenes

A continuación, se detallan los puntos a tener en cuenta para la implementación e integración del modelo de clasificación de imágenes.

- **Clasificación de imágenes históricas:** Se deberá realizar un proceso inicial y masivo de clasificación de todas las imágenes históricas existentes con el modelo de clasificación obtenido para luego integrar esa información al sistema existente.

- **Clasificación de imágenes que ingresan diariamente:** El modelo de clasificación deberá ser integrado al proceso de carga diaria de documentos de legajos de clientes. Una vez que los documentos se encuentran en el servidor de la compañía, se ejecuta el proceso automático de importación de los documentos (Imagen 5-25). Dentro de este proceso deberá ser integrado el algoritmo de clasificación de documentos, de modo que al importarlos al repositorio correspondiente e insertar los datos en la base de datos, se importe la información de la clase a la cual pertenecen.

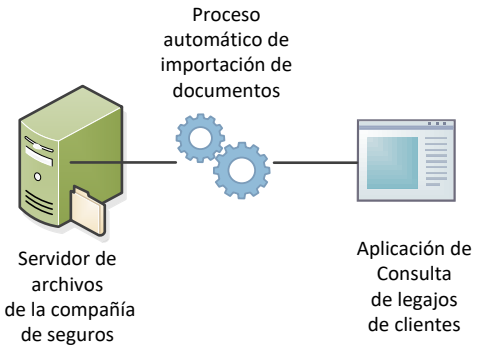


Imagen 5-25 Proceso de importación de documentos

- **Base de datos:** Se deberá agregar un campo que identifique la clase a la cual pertenece un documento. Se deberá analizar la necesidad de modificación de procedimientos almacenados que realizan la carga, modificación o eliminación de un documento en la base de datos.

- **Interfaz del software existente:** En la interfaz se deberá agregar el campo que muestre la información de la clase a la cual pertenece un documento. También se deberá agregar la funcionalidad de filtro de documentos por clase.

5.6.2 Control y mantenimiento

Al definir como objetivo de minería obtener un modelo que permita clasificar los documentos con un 80% de precisión, se acepta el riesgo de que la clasificación no sea siempre precisa. Si existen errores los mismo serán detectados durante la consulta de los documentos. Actualmente en la compañía existe una regla de cultura organizacional que indica que cualquier error encontrado en los sistemas debe ser reportado inmediatamente al área de desarrollo. Si existiesen errores los reportes de los mismos serán analizados con el objetivo de mejorar el algoritmo de clasificación.

Puede darse la situación de que ocurra el ingreso de un nuevo tipo de documento, ya sea por requerimiento legal que exija solicitar una nueva clase de documento al cliente o por necesidad de modificación de alguno existente. En este caso se deberá hacer uso de los scripts de extracción de imágenes, script de procesamiento de imágenes, script de extracción de datos y el proceso de RapidMiner. Todos estos elementos forman parte de los entregables descriptos en la sección 4.10. Con los mismos será posible crear una nueva vista minable que permitirá reentrenar el modelo de clasificación para que contemple nuevos tipos de documentos.

5.6.3 Prototipo

Con el objetivo de mostrar el funcionamiento del modelo obtenido se realizó el desarrollo de un prototipo implementado como una aplicación web. En el anexo F se detalla toda la información relacionada con el prototipo, desde su interfaz, forma de implementación y funcionamiento. Es importante destacar que el objetivo del prototipo está acotado a mostrar el funcionamiento del modelo de minería obtenido, por lo tanto, no se contempla el proceso de integración a los sistemas existentes.

6. Conclusiones y futuras líneas de investigación

La aplicación de minería de datos sobre conjuntos de imágenes para lograr su clasificación, permite realizar este proceso de manera mucho más eficiente, disminuyendo los tiempos y recursos necesarios. Contar con las imágenes clasificadas permite, a quienes las consultan, poder identificarlas rápidamente. Además, es posible conocer con mayor certeza los tipos de documentos que se tienen en guarda y sus cantidades.

El objetivo propuesto fue alcanzado al haber obtenido tres modelos que son capaces de realizar la clasificación de las imágenes con al menos un 80% de certeza. El modelo obtenido por el algoritmo *Decision Tree*, alcanza el 81% y al tratarse de un conjunto de reglas, permite una implementación simple y una rápida integración al sistema existente de consulta de legajos de clientes, por lo cual se considera la mejor opción. Este modelo permitirá realizar la clasificación de las imágenes con un alto nivel de precisión y en un tiempo significativamente menor que el que emplearía una persona.

Como trabajo futuro, se podría considerar emplear otras técnicas para el análisis de las imágenes, como ser la segmentación, es decir, dividir las imágenes en regiones para su análisis. Esto permitiría obtener atributos que las describan mejor e intentar aumentar la precisión del modelo obtenido por el algoritmo *Decision Tree*.

Otro enfoque para mejorar la precisión de las predicciones es emplear métodos multiclasificadores. Esto consiste en unir dos o más modelos para crear un nuevo modelo de clasificación (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004). En este caso se podrían considerar los dos modelos que obtuvieron el mayor valor de precisión, *K-Nearest Neighbors* y *Decision Tree*.

Debido a que uno de los temas centrales es el procesamiento de imágenes, sería altamente recomendable incluir en la etapa de recolección de datos a un experto en el tema. De este modo se podrían determinar las mejores formas de procesar las imágenes de modo que se obtenga información de mejor calidad de las mismas.

Emplear atributos como el histograma de una imagen podría aportar información que logre un mayor grado de precisión al clasificar las imágenes. El histograma muestra la cantidad de píxeles por cada intensidad (color) que posee una imagen. En este caso en particular, se trabaja con imágenes en escalas de grises, por lo tanto, existen 256 intensidades. Esto implicaría sumar 256 atributos si solo se trabaja con el histograma de las imágenes originales. Se debe evaluar el costo de adicionar estos datos en relación al beneficio del incremento en la precisión de predicción de la clase a la cual pertenecen las imágenes.

Mediante el empleo de una herramienta de reconocimiento óptico de caracteres se podría realizar la extracción del texto que contiene cada imagen. El texto podría ser de utilidad para detectar de forma más precisa a qué clase pertenece cada digitalización de los documentos.

Durante todo el proceso se observó que la mayor parte del tiempo invertido se encuentra en el proceso de extracción, selección y limpieza de los datos. Esto hace que sea un proceso al cual prestarle especial atención y buscar formas más eficientes de realizarlo.

Basarse exclusivamente en la detección de *outliers* antes de que la imagen ingrese al sistema de modo que se pueda corregir errores de forma temprana y tener evidencia para exigir una mejor calidad de servicio. Al revisar el proceso completo de extracción del conocimiento, se encontraron algunos puntos en los cuales podrían implementarse mejoras.

Durante el desarrollo del proyecto se observó que algunas tareas de procesamiento de imágenes y tareas de minería de datos consumen una gran cantidad de recursos, procesador, memoria y espacio en disco. Para estas tareas se podría contar con más de una PC para poder ejecutarlas de forma paralela. Otra alternativa es ejecutar este tipo de tareas en un servidor, para lo cual se deberían analizar los permisos necesarios, licencias, entre otros.

Al realizar la clasificación manual de los documentos se encontraron 73 distintas clases. Debido a que antes de la ejecución del proyecto no se tenía demasiada información sobre las características de las imágenes de las distintas clases se decidió seguir adelante con el análisis de todas las clases identificadas. Esto permitió tener mayor conocimiento sobre las imágenes y documentos en guarda. Una cantidad tan elevada de clases aumenta significativamente la complejidad de todo el proceso. Ahora que se posee un mejor conocimiento sobre las imágenes, podría evaluarse la posibilidad de solo considerar las clases que poseen mayor cantidad de documentos o determinar cuáles son las clases más relevantes según la experiencia y conocimiento de los usuarios. Reduciendo la cantidad de clases, permitiría enforcase en aquellas de mayor interés e intentar elevar la precisión en la predicción de la clase. Adicionalmente, se podría disminuir significativamente el tiempo y recursos invertidos en el proceso, sin que esto signifique una reducción en el beneficio.

Teniendo en consideración estos puntos y actuando en consecuencia, se podría lograr una mejor utilidad de los recursos y obtener un resultado de mayor calidad.

7. Bibliografía

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.* The CRISP-DM consortium.

Coelho, P., Gonçalves, J., Portela, F., & Santos, M. (2019). Towards to Use Image Mining to Classified Skin Problems - A Melanoma Case Study. *IEEE Access, 7*, 90132-90144. doi:10.1109/ACCESS.2019.2926837

De Bonet, J. S. (1997). Multiresolution Sampling Procedure for Analysis and Synthesis of Texture Images. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 361–368. doi:10.1145/258734.258882

De Bonet, J. S. (1997). *Structure Driven Image Database Retrieval.* Massachusetts Institute of Technology.

Google Developers. (28 de 03 de 2020). Obtenido de Google Developers: https://developers.google.com/machine-learning/crash-course/classification

Hernández Orallo, J., Ramírez Quintana, M., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos.* Madrid: Person Educación S.A.

Linoff, G., & Berry, M. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* Indianápolis: Wiley Publishing Inc.

Mahalle, A. y. (2015). *Mining and Clustering Based Image Segmentation* (3 ed., Vol. 3). International Journal of Advance Research in Computer Science and Management Studies.

Olson, D. L., & Delen, D. (2008). *Advance Data Mining Techniques.* Berlin: Springer.

Panda, M., Hassanien, A., & Abraham, A. (2016). Hybrid Data Mining Appproach for Image Segmentation Based Clasification. *International Journal of Rough Sets and Data Analysis, 3*(5), 65-81.

Piatetsky-Shapiro, G., & Mayo, M. (10 de 2014). *KDnuggets*. Recuperado el 21 de 06 de 2020, de KDnuggets:https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Solomon, C., & Breckon, T. (2011). *Fundamentals of Digital Image Procesing.* Oxford: Wiley – Blackwell.

Song, Q., Guo, Y., Jiang, J., & Li, C. (2019). High-speed Railway Fastener Detection and Localization Method based on convolutional neural network. *ArXiv, abs/1907.01141*.

Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, I - I. doi:10.1109/CVPR.2001.990517

Yousofi, M. H., Esmaeili, M., & Sada, M. (24 de Junio de 2016). A Study on Image Mining; Its Importance and Challenges. *American Journal of Software Engineering and Applications. Special Issue: Academic Research for Multidisciplinary, 5*(3-1), pp. 5-9. doi:10.11648/j.ajsea.s.2016050301.12

Anexo

A. Metodología CRISP-DM

A continuación, se describe en forma sintética la metodología CRISP-DM, que es la metodología seleccionada para guiar el desarrollo del presente proyecto.

CRISP-DM surge ante la necesidad de estandarizar el proceso de minería de datos. El objetivo era crear un modelo estándar, no propietario y disponible para quien lo necesitara. Esta metodología surge de la experiencia de proyectos de minería de datos reales.

El modelo de referencia cuenta con seis fases que se encuentran relacionadas entre sí y es posible que durante el desarrollo del proyecto sea necesario retroceder a una fase para brindar retroalimentación entre las mismas. Por cada una de las fases se definen tareas genéricas que guían los pasos a seguir durante el proyecto. A continuación, se muestra una breve descripción de las tareas:

Entender el negocio

Esta fase tiene como objetivo definir objetivos, alcances, requerimientos del proyecto dentro de la perspectiva del negocio. Las tareas a realizar son las siguientes:

- **Determinar los objetivos del negocio**: esta tarea tiene como objetivo lograr comprender la necesidad del cliente desde una perspectiva de negocio. En esta instancia se deben identificar factores que puedan afectar el resultado final del proyecto. Es necesario tener en claro, cuál es la estructura organizacional, quiénes son las personas claves, patrocinadores y unidades del negocio afectadas. Se deben establecer cuáles son los objetivos desde el punto de vista del negocio y cuáles serán los criterios de éxito del proyecto.
- **Evaluación de la situación:** Se deben identificar los recursos necesarios para la ejecución del proyecto (software, hardware, datos, recursos humanos, etc.), y su disponibilidad. Debe definirse los requerimientos del proyecto, supuestos y restricciones. Se deben identificar los riesgos y desarrollar planes de contingencia para mitigarlos. De debe realizar un análisis de los costos de la ejecución del proyecto y de los beneficios que generará.
- **Determinar los objetivos de la minería de datos:** se debe definir el objetivo desde el punto de vista de la minería de datos, especificando el resultado que se espera obtener. También se deberán especificar los criterios para determinar si el resultado es satisfactorio.
- **Generar el plan del proyecto:** determinar cuáles serán las actividades a realizar para lograr el objetivo del proyecto, junto con la duración, los recursos involucrados y las dependencias entre las tareas. Esta información, además de guiar el proyecto nos permitirá realizar una estimación de tiempos y recursos necesarios. En esta etapa, también se debe realizar una evaluación inicial de las herramientas y técnicas a emplear.

Entender los datos

La comprensión o entendimiento de los datos abarca las actividades iniciales de recolección, análisis de los datos y el armado del conjunto de datos sobre el cual se trabajará.

- **Recolección inicial de datos:** esta tarea consiste en realizar el proceso de adquisición de los datos. Se determinar qué información es necesaria, si la misma se encuentra disponible, cuáles serán los criterios para su selección y cómo será el proceso de obtención de los datos.
- **Describir los datos:** se realiza un análisis superficial de los datos para determinar, la cantidad, el formato, el tipo y complejidad. Por cada atributo se debe determinar el rango, su significado y relevancia que poseen para lograr el objetivo de minería de datos.
- **Explorar los datos:** realizar un análisis más profundo de los datos para corroborar hipótesis iniciales y, si es necesario, ajustar el objetivo de la minería de datos.
- **Verificar la calidad de los datos:** se debe verificar si los datos son completos, si presentan errores o valores faltantes. Si se detectan problemas de calidad, plantear posibles soluciones. Analizar los datos en búsqueda de valores anómalos y determinar si aportan o no valor. Se debe decidir qué datos conservar y cuáles descartar de acuerdo al análisis realizado.

Preparación de los datos

- **Seleccionar los datos:** se debe determinar qué datos serán usados en las fases posteriores del proyecto. Para realizar la selección de datos se deben tener en cuenta los objetivos de la minería de datos, los resultados obtenidos en la fase de entendimiento y los datos que serán necesarios para la fase de modelado.
- **Limpieza de los datos:** se deben tomar acciones respecto al análisis realizado en la tarea de verificación de calidad de los datos y las mismas deben ser documentadas indicando la forma en la cual se procedió.
- **Construcción de datos:** puede ser necesario generar nuevos datos a partir de los existentes, completar valores faltantes o realizar transformaciones de los mismos.
- **Integración de datos:** si los datos se encuentran distribuidos en distintas fuentes deberán unificarse para conformar una sola.
- **Dar formato a los datos:** se debe considerar la necesidad de realizar un cambio en el formato de los datos, orden de los atributos y de los registros según sea requerido por las herramientas a ser usadas en la fase de modelado.

Modelado

se aplican los modelos seleccionados de acuerdo a las características particulares del problema.

- **Selección de la técnica de modelado:** determinar cuáles son las técnicas de modelados adecuadas según el tipo de problema que se está tratando. Pueden existir restricciones adicionales que dependen de los datos o del tipo de herramienta elegida para asistir en el proceso de minería de datos.

- **Diseñar las pruebas:** Se debe definir un plan para realizar la evaluación de los modelos, incluyendo la descripción de proceso para realizar el entramiento, prueba y evaluación.
- **Construir el modelo:** se emplean las herramientas de modelado sobre los datos resultantes de la fase de preparación de datos. Se debe documentar el proceso de elección de parámetros, especificando valores y razones por las cuales se definieron. Por cada modelo obtenido se debe describir el resultado obtenido de los procesos de evaluación.
- **Valoración del modelo:** Los modelos obtenidos deben ser evaluados para determinar si cumplen o no con los criterios de éxito de la minería de datos establecidos previamente. De aquellos que cumplan con los criterios de éxito, se deben identificar cuáles son los mejores que ofrecen un mejor resultado.

Evaluación

Obtenidos los modelos es necesario pasar a la fase de evaluación para determinar si se cumplen los objetivos planteados inicialmente, tanto de minería de datos como de negocio.

- **Evaluación del resultado:** los modelos son evaluados respecto al cumplimiento de los objetivos de minería de datos y los objetivos del negocio planteados inicialmente. Se deben determinar si existen aspectos del negocio por los cuales un modelo no deba ser considerado. Como resultado final se deben determinar cuáles son los modelos aprobados luego de pasar por el proceso de evaluación.
- **Revisión del proceso:** se realiza una revisión desde el punto de vista de la calidad, identificando problemas y posibles mejoras, junto con sus posibles soluciones y ajustes a ser implementados.
- **Determinar los siguientes pasos:** en este punto se analizan los resultados obtenidos de la evaluación y de la revisión del proceso. Según esto se toma la decisión de realizar la implementación del proyecto, realizar más iteraciones o cambiar los objetivos de la minería de datos.

Desarrollo

Por último, se realiza la etapa de desarrollo que consiste en la aplicación del conocimiento obtenido, ya sea informando los resultados o implementando un proceso repetible de minería.

- **Plan de desarrollo:** elaborar el plan a ejecutar para implementar la solución obtenida en el proceso de minería de datos. Se deben especificar los pasos a realizar para el desarrollo de la solución, cómo se vinculará con los sistemas existentes y cómo será la forma en la cual se realizarán tareas de monitoreo con el objetivo de medir el beneficio.
- **Monitoreo y mantenimiento del plan:** el plan debe contemplar la forma en la cual se controlará el desempeño del modelo, si se producen cambios en el entorno o en los objetivos planteados inicialmente. Se deben determinar las acciones a seguir ante cada uno de estos escenarios.

- **Realizar el reporte final**: este reporte contiene el plan de ejecución de las tareas de minería de datos, los costos, los resultados obtenidos, el plan de implementación y recomendaciones para trabajos futuros.
- **Revisión del proyecto**: esta tarea final consiste en realizar una evaluación de la experiencia obtenida del proyecto. Se identifican y documentas puntos en los cuales se pueden hacer mejoras y lecciones aprendidas.

B. Clases seleccionadas

En la siguiente tabla se muestran las clases que fueron conservadas luego de realizar el proceso de selección y limpieza de los datos. Por cada una de ella se detalla la cantidad de elementos que contienen y el porcentaje que representar del total de imágenes que pasaron los criterios de selección.

Clase	Cantidad de Imágenes	Porcentaje del total de imágenes
C01	314	2,28%
C02	24	0,17%
C03	2019	14,69%
C04	32	0,23%
C06	141	1,03%
C07	50	0,36%
C08	115	0,84%
C11	60	0,44%
C12	32	0,23%
C13	36	0,26%
C15	307	2,23%
C16	47	0,34%
C17	91	0,66%
C18	28	0,20%
C19	14	0,10%
C21	212	1,54%
C22	12	0,09%
C23	760	5,53%
C24	107	0,78%
C26	275	2,00%
C27	125	0,91%
C28	47	0,34%
C29	34	0,25%
C30	13	0,09%
C31	27	0,20%
C32	11	0,08%
C33	18	0,13%
C34	67	0,49%
C36	34	0,25%
C37	209	1,52%
C38	75	0,55%
C39	354	2,57%
C41	148	1,08%
C42	13	0,09%

C43	100	0,73%
C44	124	0,90%
C45	10	0,07%
C47	14	0,10%
C48	230	1,67%
C49	768	5,59%
C51	29	0,21%
C53	12	0,09%
C54	71	0,52%
C55	48	0,35%
C56	48	0,35%
C57	285	2,07%
C58	88	0,64%
C59	525	3,82%
C61	546	3,97%
C62	102	0,74%
C63	800	5,82%
C64	111	0,81%
C65	380	2,76%
C66	81	0,59%
C68	98	0,71%
C69	109	0,79%
C70	116	0,84%
C71	282	2,05%
C72	280	2,04%
C74	95	0,69%
C75	57	0,41%
C76	297	2,16%
C77	43	0,31%
C80	1953	14,21%
C84	195	1,42%

Tabla B-1 Clases conservadas luego del proceso de selección y limpieza de datos

C. Vista minable

En la siguiente tabla se muestra un fragmento de los ejemplos que forman parte de la vista minable resultante de la fase de preparación de los datos:

Imagen	Alto	Ancho	Ori0	Ori255	DTF0	DTF255	INT0	INT255	SMO0	SMO255	SOB0	SOB255	Clase
044102X2DZ18062009	2516	3401	7700	8546363	4282103	3198771	124944	22	7244	8548530	8553943	1572	C04
232501X2BZ28052009	4267	2573	10810	10963597	5489970	4485028	159109	29	8188	10966091	10974062	2657	C04
223202X2DZ10062009	1425	2139	8489	3032328	1528650	1271945	44937	10	4017	3038113	3040909	3944	C04
111801X2DZ05062009	3598	2499	15446	8961507	4494821	3984027	131286	24	6462	8974214	8977582	7041	C04
072414X2DZ02102014	246	901	4330	208336	110855	107991	4224	2	225	220515	208519	7370	C02
072312X2BZ16092014	228	904	5671	191655	103254	100289	3725	2	122	205831	192448	8337	C02
105522X2DZ16072014	234	903	4395	196372	105463	103377	3410	2	0	211178	195079	8337	C02
080604X2BZ07042014	1289	903	7103	1149226	586382	559548	18212	5	247	1162865	1151108	8552	C23
080291X2DZ20092013	238	901	5784	197322	107110	104929	3425	2	8	214156	198325	8748	C02
085437X2BZ13012014	246	906	14274	197873	111610	108574	4198	3	9183	212618	206925	8804	C02
075279X2DZ05012016	249	903	7683	207065	112341	109736	4090	2	1113	222876	209404	9461	C02
071875X2DZ10102012	1297	903	5579	1153629	585500	573894	17542	5	52	1170870	1152985	9482	C23
111659X2BZ13032014	243	906	9521	200505	109517	107785	4816	4	2511	215725	205023	9680	C02
084359X2DZ15022012	1302	904	7241	1158786	588168	575982	17754	5	87	1176588	1159641	9727	C23
074155X2DZ13012014	237	901	6177	195164	106781	104483	3536	2	0	213101	195495	9960	C02
074393X2BZ23102015	228	901	7867	186975	102599	100395	3273	2	0	205204	188349	10159	C02
174302X2BZ12052009	1898	2008	17975	3771929	1906288	1719585	56194	10	6175	3789054	3791325	10280	C04
130374X2DZ05062014	237	900	6241	194830	106602	104409	3402	2	0	213234	194686	10370	C03
111181X2DZ29082011	1268	906	7713	1128597	571542	565203	18372	10	1563	1145804	1130743	10393	C23
082885X2BZ02062015	263	901	6576	218488	118347	116032	3740	3	124	235015	219006	10436	C02
074617X2BZ07102015	244	904	7445	200862	110476	107751	3493	2	0	220185	202076	10547	C02
073989X2BZ27092012	315	449	6885	123549	70682	69025	2215	1	0	137314	124760	10562	C03
073086X2BZ19032013	269	904	5950	224381	121686	119010	3833	2	0	243026	224178	10698	C02

Tabla B-2 Fragmento de la vista minable resultante de la fase de preparación de los datos.

D. *Class recall* y *class presicion* para los modelos obtenidos

En esta sección se detallan los valores de *class recall* y *class precision* para cada modelo obtenido durante la fase de modelado.

Clase predicha	*Class Precision*				
	Naive Bayes	***Decision tree***	***Random tree***	***Rule induction***	***K Nearest neighbors***
pred. C01	100.00%	96.90%	67.70%	99.37%	95.94%
pred. C02	84.00%	72.00%	76.92%	71.43%	90.91%
pred. C03	99.85%	99.70%	92.75%	99.85%	99.01%
pred. C04	72.73%	100.00%	90.91%	100.00%	100.00%
pred. C06	0.00%	71.72%	40.58%	67.31%	69.77%
pred. C07	33.33%	77.78%	32.14%	63.04%	70.37%
pred. C08	0.00%	59.00%	14.29%	55.06%	63.00%
pred. C10	0.00%	0.00%	0.00%	0.00%	0.00%
pred. C11	0.00%	36.84%	16.67%	46.15%	60.61%
pred. C12	47.06%	75.00%	0.00%	50.00%	55.56%
pred. C13	100.00%	90.00%	50.00%	100.00%	90.00%
pred. C15	0.00%	48.48%	0.00%	34.52%	57.58%
pred. C16	77.27%	64.71%	28.57%	56.25%	69.23%
pred. C17	27.63%	44.00%	11.11%	51.52%	90.00%
pred. C18	50.00%	82.14%	60.00%	100.00%	79.17%
pred. C19	1.33%	0.00%	0.00%	0.00%	60.00%
pred. C21	75.00%	77.78%	37.50%	69.80%	76.17%
pred. C22	2.61%	0.00%	0.00%	0.00%	0.00%
pred. C23	0.00%	86.36%	56.19%	84.68%	88.84%
pred. C24	11.90%	53.66%	26.83%	47.73%	78.05%
pred. C26	0.00%	55.30%	36.04%	48.63%	56.99%
pred. C27	0.00%	67.71%	32.76%	69.66%	75.22%
pred. C28	4.70%	66.67%	0.00%	37.50%	35.71%
pred. C29	14.40%	57.58%	18.18%	66.67%	66.67%
pred. C30	2.11%	50.00%	0.00%	0.00%	50.00%
pred. C31	0.00%	0.00%	0.00%	14.29%	50.00%
pred. C32	0.00%	0.00%	0.00%	0.00%	0.00%
pred. C33	19.67%	70.00%	0.00%	81.82%	55.56%
pred. C34	100.00%	94.59%	76.32%	100.00%	94.12%
pred. C35	0.00%	0.00%	0.00%	0.00%	0.00%
pred. C36	0.00%	30.00%	0.00%	20.51%	54.29%
pred. C37	92.34%	95.81%	76.42%	94.42%	91.86%
pred. C38	42.35%	80.82%	29.17%	76.19%	80.52%
pred. C39	75.84%	83.65%	52.44%	74.29%	83.65%
pred. C41	100.00%	85.91%	44.14%	93.00%	88.73%

pred. C42	0.00%	58.33%	25.00%	40.00%	75.00%
pred. C43	7.60%	0.00%	0.00%	0.00%	21.05%
pred. C44	60.61%	62.96%	0.00%	69.09%	65.77%
pred. C45	8.70%	0.00%	0.00%	50.00%	60.00%
pred. C47	0.00%	0.00%	0.00%	0.00%	0.00%
pred. C48	0.00%	34.15%	25.00%	41.59%	50.48%
pred. C49	95.76%	98.51%	61.35%	97.32%	95.93%
pred. C51	100.00%	92.31%	20.00%	100.00%	92.31%
pred. C52	0.00%	0.00%	0.00%	0.00%	0.00%
pred. C53	7.25%	0.00%	0.00%	18.18%	50.00%
pred. C54	36.42%	80.00%	57.14%	65.33%	85.00%
pred. C55	68.42%	84.21%	66.67%	65.52%	74.19%
pred. C56	82.86%	70.91%	53.23%	71.74%	79.55%
pred. C57	0.00%	81.73%	65.59%	85.02%	91.73%
pred. C58	0.00%	0.00%	0.00%	0.00%	15.87%
pred. C59	75.87%	84.70%	71.47%	82.98%	94.59%
pred. C60	0.00%	0.00%	0.00%	0.00%	0.00%
pred. C61	68.78%	72.78%	59.07%	74.92%	83.51%
pred. C62	13.99%	66.13%	11.76%	48.15%	47.30%
pred. C63	59.04%	61.51%	51.08%	65.07%	80.14%
pred. C64	0.00%	62.69%	61.90%	58.46%	67.71%
pred. C65	0.00%	43.01%	35.54%	38.83%	48.69%
pred. C66	13.17%	41.67%	10.53%	21.74%	65.57%
pred. C68	0.00%	83.75%	62.00%	83.33%	82.43%
pred. C69	57.69%	77.33%	52.17%	89.04%	82.93%
pred. C70	0.00%	85.96%	76.92%	77.42%	90.32%
pred. C71	72.81%	71.64%	49.08%	74.10%	79.57%
pred. C72	61.03%	80.13%	55.63%	79.87%	89.29%
pred. C74	75.82%	91.21%	52.63%	92.55%	86.27%
pred. C75	16.44%	83.33%	44.19%	78.85%	85.45%
pred. C76	62.30%	68.77%	54.73%	67.95%	71.60%
pred. C77	0.00%	45.24%	0.00%	38.30%	55.56%
pred. C80	81.97%	84.35%	61.16%	83.80%	89.62%
pred. C84	100.00%	96.82%	83.22%	98.73%	99.36%

Tabla D-3 Valor de *class precision* para cada clase de cada uno de los modelos obtenidos

Clase real	*Class Recall*				
	Naive Bayes	*Decision tree*	*Random tree*	*Rule induction*	*K Nearest neighbors*
true C01	98.09%	99.68%	62.74%	99.68%	97.77%
true C02	87.50%	75.00%	41.67%	83.33%	83.33%
true C03	97.62%	99.60%	94.45%	99.55%	99.21%
true C04	88.89%	94.44%	55.56%	38.89%	94.44%
true C06	0.00%	74.29%	40.00%	75.00%	85.71%
true C07	76.09%	76.09%	19.57%	63.04%	82.61%
true C08	0.00%	53.15%	4.50%	44.14%	56.76%
true C10	0.00%	0.00%	0.00%	0.00%	0.00%
true C11	0.00%	17.50%	5.00%	30.00%	50.00%
true C12	80.00%	60.00%	0.00%	40.00%	50.00%
true C13	90.00%	90.00%	20.00%	60.00%	90.00%
true C15	0.00%	29.91%	0.00%	27.10%	35.51%
true C16	77.27%	50.00%	18.18%	40.91%	81.82%
true C17	23.60%	12.36%	1.12%	19.10%	60.67%
true C18	100.00%	82.14%	21.43%	46.43%	67.86%
true C19	85.71%	0.00%	0.00%	0.00%	21.43%
true C21	48.80%	70.33%	1.44%	67.46%	70.33%
true C22	83.33%	0.00%	0.00%	0.00%	0.00%
true C23	0.00%	91.47%	80.06%	91.04%	90.90%
true C24	7.81%	34.38%	17.19%	32.81%	50.00%
true C26	0.00%	27.97%	15.33%	27.20%	42.15%
true C27	0.00%	65.66%	19.19%	62.63%	85.86%
true C28	93.75%	62.50%	0.00%	37.50%	31.25%
true C29	69.23%	73.08%	7.69%	61.54%	92.31%
true C30	25.00%	37.50%	0.00%	0.00%	12.50%
true C31	0.00%	0.00%	0.00%	6.67%	40.00%
true C32	0.00%	0.00%	0.00%	0.00%	0.00%
true C33	66.67%	38.89%	0.00%	50.00%	55.56%
true C34	86.05%	81.40%	67.44%	79.07%	74.42%
true C35	0.00%	0.00%	0.00%	0.00%	0.00%
true C36	0.00%	18.18%	0.00%	24.24%	57.58%
true C37	98.09%	98.56%	89.95%	97.13%	97.13%
true C38	98.63%	80.82%	19.18%	87.67%	84.93%
true C39	89.80%	86.97%	60.91%	89.24%	88.39%
true C41	15.28%	88.89%	34.03%	64.58%	87.50%
true C42	0.00%	70.00%	10.00%	20.00%	60.00%
true C43	74.00%	0.00%	0.00%	0.00%	12.00%
true C44	64.52%	68.55%	0.00%	61.29%	58.87%
true C45	66.67%	0.00%	0.00%	33.33%	50.00%

true C47	0.00%	0.00%	0.00%	0.00%	0.00%
true C48	0.00%	12.17%	7.39%	20.43%	46.09%
true C49	94.39%	94.91%	40.21%	94.91%	98.56%
true C51	100.00%	100.00%	25.00%	100.00%	100.00%
true C52	0.00%	0.00%	0.00%	0.00%	0.00%
true C53	71.43%	0.00%	0.00%	28.57%	28.57%
true C54	94.83%	96.55%	55.17%	84.48%	87.93%
true C55	76.47%	47.06%	5.88%	55.88%	67.65%
true C56	65.91%	88.64%	75.00%	75.00%	79.55%
true C57	0.00%	92.13%	68.54%	85.02%	95.51%
true C58	0.00%	0.00%	0.00%	0.00%	11.36%
true C59	70.05%	91.83%	63.86%	86.88%	95.30%
true C60	0.00%	0.00%	0.00%	0.00%	0.00%
true C61	30.47%	91.96%	73.64%	92.15%	88.97%
true C62	60.00%	41.00%	2.00%	13.00%	35.00%
true C63	81.60%	91.24%	89.97%	88.20%	85.53%
true C64	0.00%	40.00%	24.76%	36.19%	61.90%
true C65	0.00%	21.51%	15.86%	30.38%	50.00%
true C66	74.68%	6.33%	2.53%	6.33%	50.63%
true C68	0.00%	84.81%	39.24%	75.95%	77.22%
true C69	18.99%	73.42%	45.57%	82.28%	86.08%
true C70	0.00%	74.24%	45.45%	72.73%	84.85%
true C71	59.07%	68.33%	56.94%	73.31%	79.00%
true C72	44.91%	95.85%	65.28%	94.34%	94.34%
true C74	72.63%	87.37%	63.16%	91.58%	92.63%
true C75	21.05%	87.72%	33.33%	71.93%	82.46%
true C76	78.72%	84.80%	64.53%	77.36%	80.07%
true C77	0.00%	51.35%	0.00%	48.65%	54.05%
true C80	39.16%	97.23%	92.11%	97.54%	94.26%
true C84	98.11%	95.60%	77.99%	97.48%	98.11%

Tabla D-4 Valor de *class recall* para cada clase de cada uno de los modelos obtenidos

E. Árbol de decisión

A continuación de destallan las reglas obtenidas del modelo de árbol de decisión en RapidMiner:

```
Alto > 427.500
|   Alto > 1687.500
|   |   Alto > 2332
|   |   |   SOB255 > 291640
|   |   |   |   Ori0 > 791060.500
|   |   |   |   |   Alto > 4116.500: C84
|   |   |   |   |   Alto ≤ 4116.500: C69
|   |   |   |   Ori0 ≤ 791060.500
|   |   |   |   |   DTF255 > 4534634
|   |   |   |   |   |   Alto > 4117.500
|   |   |   |   |   |   |   Alto > 4305: C70
|   |   |   |   |   |   |   Alto ≤ 4305: C64
|   |   |   |   |   |   Alto ≤ 4117.500
|   |   |   |   |   |   |   Ori0 > 647939: C69
|   |   |   |   |   |   |   Ori0 ≤ 647939: C68
|   |   |   |   |   DTF255 ≤ 4534634
|   |   |   |   |   |   SOB255 > 338514.500
|   |   |   |   |   |   |   SMO0 > 49689.500
|   |   |   |   |   |   |   |   Ori255 > 7678146
|   |   |   |   |   |   |   |   |   SMO0 > 85060
|   |   |   |   |   |   |   |   |   |   SMO0 > 100160: C65
|   |   |   |   |   |   |   |   |   |   SMO0 ≤ 100160: C48
|   |   |   |   |   |   |   |   |   SMO0 ≤ 85060: C65
|   |   |   |   |   |   |   |   Ori255 ≤ 7678146
|   |   |   |   |   |   |   |   |   SMO0 > 53979: C63
|   |   |   |   |   |   |   |   |   SMO0 ≤ 53979
|   |   |   |   |   |   |   |   |   |   Ori255 > 7388510: C65
|   |   |   |   |   |   |   |   |   |   Ori255 ≤ 7388510: C63
|   |   |   |   |   |   |   SMO0 ≤ 49689.500
|   |   |   |   |   |   |   |   Ori0 > 505361
|   |   |   |   |   |   |   |   |   SOB255 > 380996.500: C63
|   |   |   |   |   |   |   |   |   SOB255 ≤ 380996.500: C62
|   |   |   |   |   |   |   |   Ori0 ≤ 505361: C65
|   |   |   |   |   |   SOB255 ≤ 338514.500
|   |   |   |   |   |   |   Ori0 > 394446
|   |   |   |   |   |   |   |   SMO0 > 82713.500
|   |   |   |   |   |   |   |   |   Ancho > 2502: C48
|   |   |   |   |   |   |   |   |   Ancho ≤ 2502: C63
|   |   |   |   |   |   |   |   SMO0 ≤ 82713.500
|   |   |   |   |   |   |   |   |   SMO0 > 27998.500: C65
|   |   |   |   |   |   |   |   |   SMO0 ≤ 27998.500
|   |   |   |   |   |   |   |   |   |   Alto > 3539: C48
|   |   |   |   |   |   |   |   |   |   Alto ≤ 3539: C65
|   |   |   |   |   |   |   Ori0 ≤ 394446: C42
|   |   |   |   SOB255 ≤ 291640
|   |   |   |   |   Alto > 2677
|   |   |   |   |   |   Ori0 > 304203.500
|   |   |   |   |   |   |   Alto > 3185.500
|   |   |   |   |   |   |   |   Alto > 3800
|   |   |   |   |   |   |   |   |   Alto > 4056
|   |   |   |   |   |   |   |   |   |   Alto > 4149.500: C70
|   |   |   |   |   |   |   |   |   |   Alto ≤ 4149.500: C69
|   |   |   |   |   |   |   |   |   Alto ≤ 4056: C51
|   |   |   |   |   |   |   |   Alto ≤ 3800
|   |   |   |   |   |   |   |   |   SMO0 > 25208.500
|   |   |   |   |   |   |   |   |   |   SMO0 > 104312
|   |   |   |   |   |   |   |   |   |   |   Ori0 > 385377
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 211317
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 415697.500: C77
```

```
|   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 415697.500: C12
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 211317
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO255 > 8400686.500: C11
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO255 ≤ 8400686.500: C28
|   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 385377
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO255 > 8383729: C77
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO255 ≤ 8383729: C15
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 104312
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 323814: C72
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 323814: C27
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 25208.500: C27
|   |   |   |   |   |   |   |   |   |   |   Alto ≤ 3185.500
|   |   |   |   |   |   |   |   |   |   |   |   Alto > 2880.500: C56
|   |   |   |   |   |   |   |   |   |   |   |   Alto ≤ 2880.500
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 431583.500: C56
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 431583.500: C55
|   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 304203.500
|   |   |   |   |   |   |   |   |   |   |   |   SOB255 > 168183
|   |   |   |   |   |   |   |   |   |   |   |   Alto > 3641: C13
|   |   |   |   |   |   |   |   |   |   |   |   Alto ≤ 3641
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 29611
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SOB0 > 8551783: C08
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SOB0 ≤ 8551783: C15
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 29611
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SOB255 > 276141.500: C42
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SOB255 ≤ 276141.500: C27
|   |   |   |   |   |   |   |   |   |   |   |   SOB255 ≤ 168183
|   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 57887
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 61345
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SOB255 > 119515: C08
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SOB255 ≤ 119515: C15
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 61345
|   |   |   |   |   |   |   |   |   |   |   |   |   |   INT255 > 21.500: C23
|   |   |   |   |   |   |   |   |   |   |   |   |   |   INT255 ≤ 21.500: C11
|   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 57887: C04
|   |   |   |   |   |   |   |   |   |   Alto ≤ 2677
|   |   |   |   |   |   |   |   |   |   |   SMO0 > 73028
|   |   |   |   |   |   |   |   |   |   |   |   Ancho > 2569
|   |   |   |   |   |   |   |   |   |   |   |   |   Ancho > 2609.500: C15
|   |   |   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 2609.500: C54
|   |   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 2569: C57
|   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 73028: C54
|   |   |   |   |   |   |   |   |   |   Alto ≤ 2332
|   |   |   |   |   |   |   |   |   |   |   SOB0 > 3308324.500
|   |   |   |   |   |   |   |   |   |   |   |   Ancho > 2029.500: C18
|   |   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 2029.500
|   |   |   |   |   |   |   |   |   |   |   |   |   INT255 > 15.500
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 213433: C71
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 213433: C17
|   |   |   |   |   |   |   |   |   |   |   |   |   INT255 ≤ 15.500: C61
|   |   |   |   |   |   |   |   |   |   |   SOB0 ≤ 3308324.500
|   |   |   |   |   |   |   |   |   |   |   |   Ancho > 1938
|   |   |   |   |   |   |   |   |   |   |   |   |   INT0 > 57455: C71
|   |   |   |   |   |   |   |   |   |   |   |   |   INT0 ≤ 57455: C61
|   |   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 1938
|   |   |   |   |   |   |   |   |   |   |   |   |   SOB255 > 149842.500
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 55819.500: C17
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 55819.500: C71
|   |   |   |   |   |   |   |   |   |   |   |   SOB255 ≤ 149842.500: C61
|   |   Alto ≤ 1687.500
|   |   |   SMO0 > 40413.500
|   |   |   |   SMO255 > 1281998
|   |   |   |   |   Ancho > 949.500
```

```
|   |   |   |   |   Ancho > 1982.500
|   |   |   |   |   |   Alto > 1372
|   |   |   |   |   |   |   SMO0 > 56218
|   |   |   |   |   |   |   |   Alto > 1537.500: C38
|   |   |   |   |   |   |   |   Alto ≤ 1537.500
|   |   |   |   |   |   |   |   |   Ancho > 1996: C41
|   |   |   |   |   |   |   |   |   Ancho ≤ 1996: C38
|   |   |   |   |   |   |   |   SMO0 ≤ 56218
|   |   |   |   |   |   |   |   |   Ori255 > 2360865.500: C38
|   |   |   |   |   |   |   |   |   Ori255 ≤ 2360865.500: C41
|   |   |   |   |   |   |   Alto ≤ 1372: C74
|   |   |   |   |   Ancho ≤ 1982.500
|   |   |   |   |   |   DTF0 > 1401634.500
|   |   |   |   |   |   |   SOB255 > 163500
|   |   |   |   |   |   |   |   DTF0 > 1458179.500: C39
|   |   |   |   |   |   |   |   DTF0 ≤ 1458179.500: C26
|   |   |   |   |   |   |   SOB255 ≤ 163500
|   |   |   |   |   |   |   |   Ori0 > 252854: C39
|   |   |   |   |   |   |   |   Ori0 ≤ 252854: C26
|   |   |   |   |   |   DTF0 ≤ 1401634.500
|   |   |   |   |   |   |   Alto > 1331
|   |   |   |   |   |   |   |   SOB255 > 109708: C76
|   |   |   |   |   |   |   |   SOB255 ≤ 109708: C26
|   |   |   |   |   |   |   Alto ≤ 1331: C74
|   |   |   |   Ancho ≤ 949.500
|   |   |   |   |   SOB255 > 90090.500
|   |   |   |   |   |   SMO0 > 68980: C59
|   |   |   |   |   |   SMO0 ≤ 68980: C49
|   |   |   |   |   SOB255 ≤ 90090.500: C24
|   |   |   SMO255 ≤ 1281998
|   |   |   |   SOB255 > 70010.500: C59
|   |   |   |   SOB255 ≤ 70010.500
|   |   |   |   |   SOB255 > 46550
|   |   |   |   |   |   INT0 > 16407.500
|   |   |   |   |   |   |   Ancho > 916: C15
|   |   |   |   |   |   |   Ancho ≤ 916: C07
|   |   |   |   |   |   INT0 ≤ 16407.500
|   |   |   |   |   |   |   Ori255 > 968932: C15
|   |   |   |   |   |   |   Ori255 ≤ 968932: C30
|   |   |   |   |   SOB255 ≤ 46550
|   |   |   |   |   |   SOB255 > 29953: C29
|   |   |   |   |   |   SOB255 ≤ 29953: C15
|   |   SMO0 ≤ 40413.500
|   |   |   SOB255 > 74913.500
|   |   |   |   Ancho > 1360
|   |   |   |   |   Alto > 1446
|   |   |   |   |   |   Ancho > 2004: C38
|   |   |   |   |   |   Ancho ≤ 2004
|   |   |   |   |   |   |   Ori0 > 185942
|   |   |   |   |   |   |   |   SOB255 > 166874
|   |   |   |   |   |   |   |   |   INT0 > 37458.500: C26
|   |   |   |   |   |   |   |   |   INT0 ≤ 37458.500: C39
|   |   |   |   |   |   |   |   SOB255 ≤ 166874
|   |   |   |   |   |   |   |   |   DTF0 > 1377364: C39
|   |   |   |   |   |   |   |   |   DTF0 ≤ 1377364: C26
|   |   |   |   |   |   |   Ori0 ≤ 185942
|   |   |   |   |   |   |   |   Ori255 > 2446775.500: C16
|   |   |   |   |   |   |   |   Ori255 ≤ 2446775.500: C26
|   |   |   |   |   Alto ≤ 1446
|   |   |   |   |   |   Alto > 1345
|   |   |   |   |   |   |   DTF255 > 1263160
|   |   |   |   |   |   |   |   Ori0 > 265100.500: C39
|   |   |   |   |   |   |   |   Ori0 ≤ 265100.500
```

```
|   |   |   |   |   |   |   |   |   SOB255 > 98624.500: C76
|   |   |   |   |   |   |   |   |   SOB255 ≤ 98624.500: C26
|   |   |   |   |   |   |   |   DTF255 ≤ 1263160: C75
|   |   |   |   |   |   |   Alto ≤ 1345
|   |   |   |   |   |   |   |   Ancho > 1934.500
|   |   |   |   |   |   |   |   |   Alto > 1129: C74
|   |   |   |   |   |   |   |   |   Alto ≤ 1129: C34
|   |   |   |   |   |   |   |   Ancho ≤ 1934.500
|   |   |   |   |   |   |   |   |   Ori255 > 2020390.500: C75
|   |   |   |   |   |   |   |   |   Ori255 ≤ 2020390.500
|   |   |   |   |   |   |   |   |   |   SOB255 > 100432: C26
|   |   |   |   |   |   |   |   |   |   SOB255 ≤ 100432: C33
|   |   |   |   |   Ancho ≤ 1360
|   |   |   |   |   |   SMO255 > 1300634.500
|   |   |   |   |   |   |   SMO0 > 15602
|   |   |   |   |   |   |   |   SOB255 > 109008
|   |   |   |   |   |   |   |   |   INT255 > 4.500: C80
|   |   |   |   |   |   |   |   |   INT255 ≤ 4.500
|   |   |   |   |   |   |   |   |   |   Ori255 > 1181381: C44
|   |   |   |   |   |   |   |   |   |   Ori255 ≤ 1181381: C21
|   |   |   |   |   |   |   |   SOB255 ≤ 109008
|   |   |   |   |   |   |   |   |   INT0 > 22182.500
|   |   |   |   |   |   |   |   |   |   Ancho > 942: C64
|   |   |   |   |   |   |   |   |   |   Ancho ≤ 942
|   |   |   |   |   |   |   |   |   |   |   INT255 > 5: C80
|   |   |   |   |   |   |   |   |   |   |   INT255 ≤ 5: C44
|   |   |   |   |   |   |   |   |   INT0 ≤ 22182.500
|   |   |   |   |   |   |   |   |   |   SMO0 > 28309: C24
|   |   |   |   |   |   |   |   |   |   SMO0 ≤ 28309
|   |   |   |   |   |   |   |   |   |   |   Ori255 > 1135391: C44
|   |   |   |   |   |   |   |   |   |   |   Ori255 ≤ 1135391: C21
|   |   |   |   |   |   |   SMO0 ≤ 15602
|   |   |   |   |   |   |   |   Ori0 > 131547.500: C21
|   |   |   |   |   |   |   |   Ori0 ≤ 131547.500
|   |   |   |   |   |   |   |   |   Ori255 > 1227796
|   |   |   |   |   |   |   |   |   |   Ori0 > 91918.500: C64
|   |   |   |   |   |   |   |   |   |   Ori0 ≤ 91918.500
|   |   |   |   |   |   |   |   |   |   |   DTF0 > 729832: C66
|   |   |   |   |   |   |   |   |   |   |   DTF0 ≤ 729832
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 1056.500: C24
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 1056.500: C64
|   |   |   |   |   |   |   |   |   Ori255 ≤ 1227796
|   |   |   |   |   |   |   |   |   |   SOB255 > 89056: C80
|   |   |   |   |   |   |   |   |   |   SOB255 ≤ 89056: C24
|   |   |   |   |   SMO255 ≤ 1300634.500
|   |   |   |   |   |   SMO255 > 1143163.500
|   |   |   |   |   |   |   Alto > 1345: C59
|   |   |   |   |   |   |   Alto ≤ 1345
|   |   |   |   |   |   |   |   SOB255 > 82362
|   |   |   |   |   |   |   |   |   Ori0 > 87657: C65
|   |   |   |   |   |   |   |   |   Ori0 ≤ 87657: C48
|   |   |   |   |   |   |   |   SOB255 ≤ 82362
|   |   |   |   |   |   |   |   |   Ori0 > 69363.500
|   |   |   |   |   |   |   |   |   |   Ancho > 911.500: C58
|   |   |   |   |   |   |   |   |   |   Ancho ≤ 911.500: C65
|   |   |   |   |   |   |   |   |   Ori0 ≤ 69363.500: C59
|   |   |   |   |   |   SMO255 ≤ 1143163.500
|   |   |   |   |   |   |   SMO0 > 16439
|   |   |   |   |   |   |   |   Ori0 > 125724.500: C59
|   |   |   |   |   |   |   |   Ori0 ≤ 125724.500: C15
|   |   |   |   |   |   |   SMO0 ≤ 16439
|   |   |   |   |   |   |   |   SOB0 > 1026135: C48
|   |   |   |   |   |   |   |   SOB0 ≤ 1026135: C59
```

```
|   |   |   SOB255 ≤ 74913.500
|   |   |   |   Alto > 493.500
|   |   |   |   |   Alto > 583.500
|   |   |   |   |   |   Ori255 > 1041633.500
|   |   |   |   |   |   |   SOB255 > 53920.500
|   |   |   |   |   |   |   |   Alto > 1142.500
|   |   |   |   |   |   |   |   |   Alto > 1345.500
|   |   |   |   |   |   |   |   |   |   Alto > 1490
|   |   |   |   |   |   |   |   |   |   |   Alto > 1534: C24
|   |   |   |   |   |   |   |   |   |   |   Alto ≤ 1534: C70
|   |   |   |   |   |   |   |   |   |   Alto ≤ 1490
|   |   |   |   |   |   |   |   |   |   |   Ancho > 1388: C26
|   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 1388
|   |   |   |   |   |   |   |   |   |   |   |   Ori255 > 1159447.500: C68
|   |   |   |   |   |   |   |   |   |   |   |   Ori255 ≤ 1159447.500: C69
|   |   |   |   |   |   |   |   |   Alto ≤ 1345.500
|   |   |   |   |   |   |   |   |   |   INT0 > 17260.500
|   |   |   |   |   |   |   |   |   |   |   Ori0 > 61366
|   |   |   |   |   |   |   |   |   |   |   |   Ancho > 913: C26
|   |   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 913: C36
|   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 61366: C06
|   |   |   |   |   |   |   |   |   |   INT0 ≤ 17260.500: C23
|   |   |   |   |   |   |   |   |   Alto ≤ 1142.500: C34
|   |   |   |   |   |   |   |   SOB255 ≤ 53920.500
|   |   |   |   |   |   |   |   |   SMO0 > 9212
|   |   |   |   |   |   |   |   |   |   Ori0 > 46392
|   |   |   |   |   |   |   |   |   |   |   Ori0 > 56787
|   |   |   |   |   |   |   |   |   |   |   |   INT0 > 20369
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 73070.500: C16
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 73070.500: C29
|   |   |   |   |   |   |   |   |   |   |   |   INT0 ≤ 20369
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 22868: C11
|   |   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 22868: C57
|   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 56787: C77
|   |   |   |   |   |   |   |   |   |   Ori0 ≤ 46392
|   |   |   |   |   |   |   |   |   |   |   INT255 > 15.500: C23
|   |   |   |   |   |   |   |   |   |   |   INT255 ≤ 15.500
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 13248: C23
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 13248: C08
|   |   |   |   |   |   |   |   |   SMO0 ≤ 9212
|   |   |   |   |   |   |   |   |   |   Ancho > 1418.500: C16
|   |   |   |   |   |   |   |   |   |   Ancho ≤ 1418.500
|   |   |   |   |   |   |   |   |   |   |   SMO255 > 1121608
|   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 53165.500: C08
|   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 53165.500
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 > 47538.500: C06
|   |   |   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 47538.500: C23
|   |   |   |   |   |   |   |   |   |   |   SMO255 ≤ 1121608
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 2321: C23
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 2321: C08
|   |   |   |   |   |   Ori255 ≤ 1041633.500
|   |   |   |   |   |   |   INT0 > 16551.500
|   |   |   |   |   |   |   |   SOB255 > 48909.500
|   |   |   |   |   |   |   |   |   SMO0 > 251.500
|   |   |   |   |   |   |   |   |   |   SMO255 > 1156693: C36
|   |   |   |   |   |   |   |   |   |   SMO255 ≤ 1156693: C06
|   |   |   |   |   |   |   |   |   SMO0 ≤ 251.500: C27
|   |   |   |   |   |   |   |   SOB255 ≤ 48909.500
|   |   |   |   |   |   |   |   |   Ori0 > 14417.500
|   |   |   |   |   |   |   |   |   |   Ancho > 900.500: C57
|   |   |   |   |   |   |   |   |   |   Ancho ≤ 900.500: C30
|   |   |   |   |   |   |   |   |   Ori0 ≤ 14417.500: C23
|   |   |   |   |   |   |   INT0 ≤ 16551.500
```

```
|   |   |   |   |   |   |   |   Alto > 748.500
|   |   |   |   |   |   |   |   |   SOB255 > 59242
|   |   |   |   |   |   |   |   |   |   SOB0 > 1010576
|   |   |   |   |   |   |   |   |   |   |   Ori255 > 989382
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 > 3468: C15
|   |   |   |   |   |   |   |   |   |   |   |   SMO0 ≤ 3468: C27
|   |   |   |   |   |   |   |   |   |   |   Ori255 ≤ 989382: C11
|   |   |   |   |   |   |   |   |   |   SOB0 ≤ 1010576: C59
|   |   |   |   |   |   |   |   |   SOB255 ≤ 59242
|   |   |   |   |   |   |   |   |   |   Ori0 > 25145.500
|   |   |   |   |   |   |   |   |   |   |   Ori0 > 31169.500: C57
|   |   |   |   |   |   |   |   |   |   |   Ori0 ≤ 31169.500: C27
|   |   |   |   |   |   |   |   |   |   Ori0 ≤ 25145.500: C23
|   |   |   |   |   |   |   |   |   Alto ≤ 748.500
|   |   |   |   |   |   |   |   |   |   DTF0 > 235305.500
|   |   |   |   |   |   |   |   |   |   |   Ancho > 822: C24
|   |   |   |   |   |   |   |   |   |   |   Ancho ≤ 822: C61
|   |   |   |   |   |   |   |   |   |   DTF0 ≤ 235305.500: C26
|   |   |   |   |   Alto ≤ 583.500
|   |   |   |   |   |   Ancho > 708: C37
|   |   |   |   |   |   Ancho ≤ 708: C26
|   |   |   |   Alto ≤ 493.500
|   |   |   |   |   Ancho > 686
|   |   |   |   |   |   Alto > 458: C74
|   |   |   |   |   |   Alto ≤ 458: C45
|   |   |   |   |   Ancho ≤ 686: C01
Alto ≤ 427.500
|   Alto > 266
|   |   Alto > 402: C01
|   |   Alto ≤ 402: C03
|   Alto ≤ 266: C02
```

F. Prototipo

El objetivo del desarrollo del prototipo es mostrar el funcionamiento del modelo obtenido como resultado del proceso de minería de datos. Por lo que no se contempla la integración con los sistemas existentes en la empresa.

El prototipo recibe como entrada un conjunto de datos correspondiente a una imagen, estos datos son obtenidos del proceso descripto en la sección 5.2.1, y devuelve como resultado la clase a la que pertenece la imagen. La implementación se realizó en un entorno web, empleando como lenguaje de programación JavaScript:

https://gabrm021.github.io/home.html

Imagen F-1 Interfaz web del prototipo de clasificación de imágenes

El sitio web cuenta con tres secciones:

Carga de datos: Al hacer clic en el botón **Elegir un archivo** se abre un cuadro de diálogo que nos permite seleccionar un archivo. El único formato de documento admitido es .txt. El formato del documento debe estar estructurado según las siguientes reglas:

- Cada línea representa una imagen.
- Cada valor de atributo se encuentra separado por el carácter ";".
- Los atributos que debe contener cada línea son los siguientes, en el orden que son mencionados:

Atributos	Tipo de dato
Identificador único de la imagen	Texto
Alto expresado en píxeles	Entero
Ancho expresado en píxeles	Entero
Cantidad de píxeles negros de la imagen original	Entero
Cantidad de píxeles blancos de la imagen original	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método DTF	Entero
Cantidad de píxeles blancos de la imagen resultante de la aplicación del método DTF	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método Integral	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método Integral	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método Smoothmedian	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método Smoothmedian	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método Sobel	Entero
Cantidad de píxeles negros de la imagen resultante de la aplicación del método Sobel	Entero

Tabla F-1 Atributos para cada línea del documento TXT

Un conjunto de documentos .txt para pruebas puede ser encontrado en el siguiente enlace: https://www.dropbox.com/sh/5iqoly1x8lpottf/AAA07NRkxHmi5MC_NdXLdOTqa?dl=0.

Cada archivo contiene información de imágenes seleccionadas aleatoriamente, con los atributos mencionados en la Tabla F-. A continuación, se muestra un ejemplo del archivo Test00.txt:

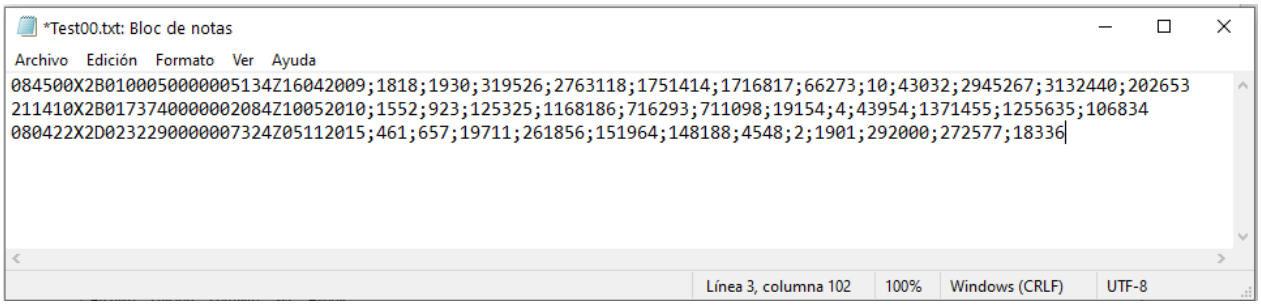


Imagen F-2 Resultados obtenidos del proceso de clasificación

La primera línea corresponde a la Imagen F-3 Resultados obtenidos del proceso de clasificación(084500X2B01000500000005134Z16042009) que se trata de un formulario de solicitud de seguro correspondiente a la clase C71 (Los datos sensibles fueron cubiertos de modo de mantener la confidencialidad de la información).

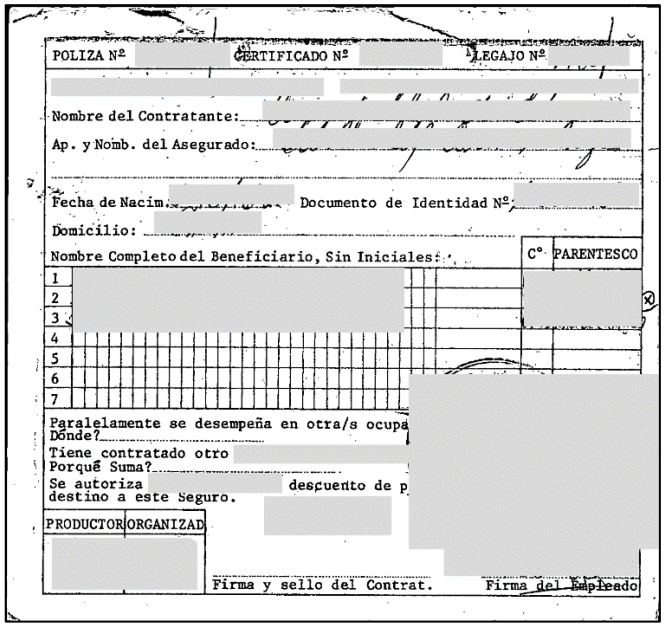


Imagen F-3 Resultados obtenidos del proceso de clasificación

La Imagen F-2 (211410X2B01737400000002084Z10052010) corresponde a un formulario de solicitud de aumento de la C49.

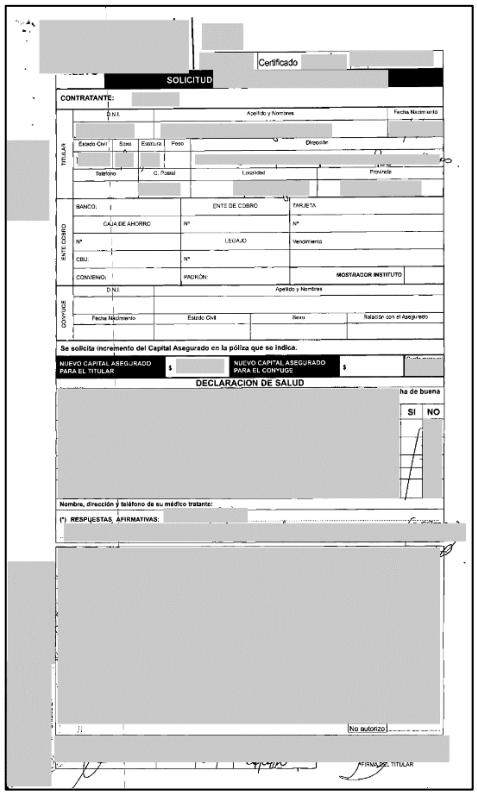


Imagen F-2 Resultados obtenidos del proceso de clasificación

Por último, la tercera línea corresponde al escaneo de un acuse de recibo perteneciente a la clase C01 (Imagen F-3)

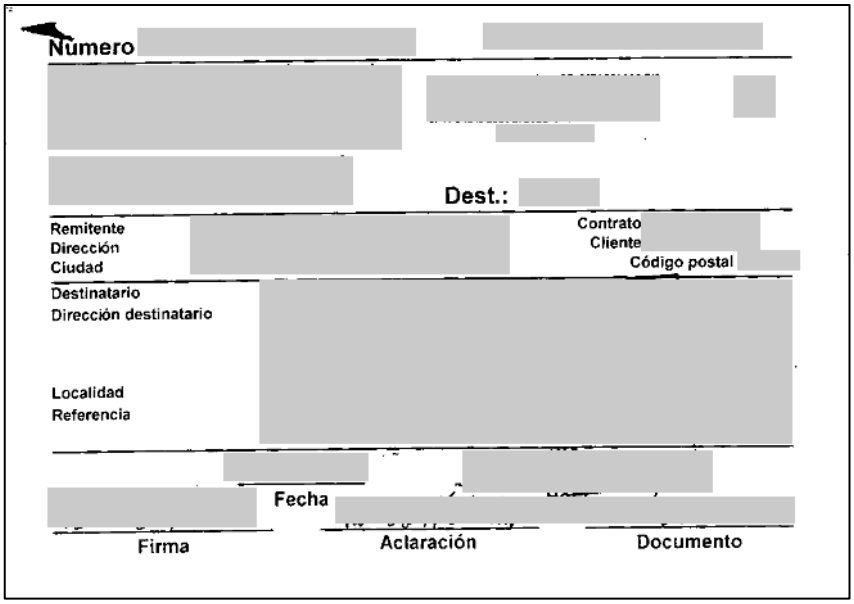


Imagen F-3 Resultados obtenidos del proceso de clasificación

En la Imagen F-4 se puede el procedimiento para la carga del archivo .txt, la validación de la extensión y el recorrido línea a línea del documento.

```
53      document
54          .getElementById('browseFile')
55          .addEventListener('change',
56              function () {
57
58                  document.getElementById("valval").innerHTML = "";
59
60                  if (browseFile.value.endsWith(".txt")){
61
62                      var fr = new FileReader();
63                      var alltext = ""
64                      var aline = ""
65                      var resulines =""
66
67                      fr.onload = function () {
68                          alltext = this.result;
69                          document.getElementById('procStatus').textContent = "En proceso";
70
71                          while (alltext.length > 0) {
72                              aline = getoneline(alltext)
73                              resulines = resulines +  "\n" + clasificarFun(aline);
74                              alltext = takeoneline(alltext,aline);
75                          }
76                          document.getElementById('contents').textContent = resulines;
77                          document.getElementById('procStatus').textContent = "Proceso de clasificación finalizado";
78                      };
79                      fr.readAsText(this.files[0]);
80                  }
81                  else {
82                      document.getElementById("valval").innerHTML = "Solo se adminten documentos .txt";
83                  }
84
85              }
86          );
87
```

Imagen F-4 Resultados obtenidos del proceso de clasificación

Estado del proceso de clasificación: Una vez seleccionado el archivo que contiene una o más filas con los datos de las imágenes a ser clasificadas, se inicia el proceso de clasificación. El algoritmo lee cada línea y pasa los datos a la función clasificarFun(img). Esta función implementa el árbol de decisión obtenido como resultado del proceso de minería de datos. Recibe la línea con los atributos, los cuales son identificados y evaluados por el árbol. Como resultado obtenemos la clase a la cual pertenece la imagen. Una vez finalizado el proceso, en esta sección se muestra el mensaje “Proceso de clasificación finalizado”. Un segmento de la función puede verse en la Imagen F-5.

```
105
106         function clasificarFun(img){
107             var imgarray = img.split(";");
108             var imgid = imgarray[0];
109             var alto = imgarray[1];
110             var ancho = imgarray[2];
111             var ori0 = imgarray[3];
112             var ori255 = imgarray[4];
113             var dtf0 = imgarray[5];
114             var dtf255 = imgarray[6];
115             var int0 = imgarray[7];
116             var int255 = imgarray[8];
117             var smo0 = imgarray[9];
118             var smo255 = imgarray[10];
119             var sob0 = imgarray[11];
120             var sob255 = imgarray[12];
121
122             if (alto > 427.500) {
123 >               if (alto > 1687.500) {···
298                 else {//alto ≤ 1687.500
299                     if (smo0 > 40413.500){
300 >                       if(smo255 > 1281998){···
352 >                       else {//smo255 ≤ 1281998···
371                     }
372                     else {//smo0 ≤ 40413.500
373 >                       if (sob255 > 74913.500){···
500 >                       else {//sob255 ≤ 74913.500···
653                     }
654                 }
655             }
656             else {
657                 if (alto > 266) {
658                     if (alto > 402) {clase = "C01"}
659                     else {clase = "C03"}
660                 }
661                 else {clase = "C02"}
662             }
663
664             return imgid + " " + clase;
665         }
```

Imagen F-5 Resultados obtenidos del proceso de clasificación

Resultados: En esta sección se muestran dos columnas, la primera con el identificador de la imagen y la segunda muestra la clase a la cual pertenece la imagen:

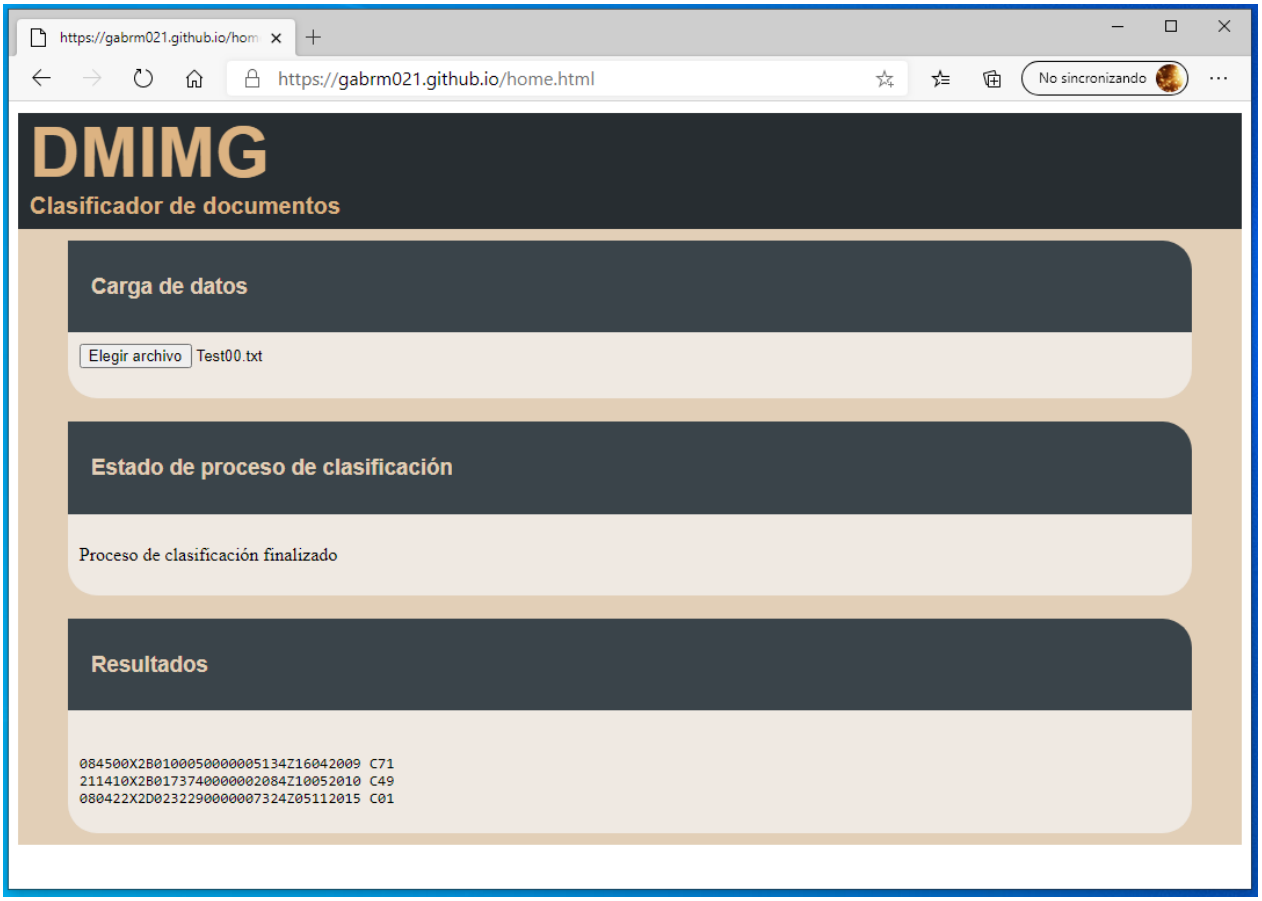


Imagen F-6 Resultados obtenidos del proceso de clasificación