Universidad de las Ciencias Informáticas

Facultad 1

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Título:

Solución para la clasificación de rasgos biométricos faciales.

Autores:

Anays Gómez García

Gregorio Ferrer Cordova

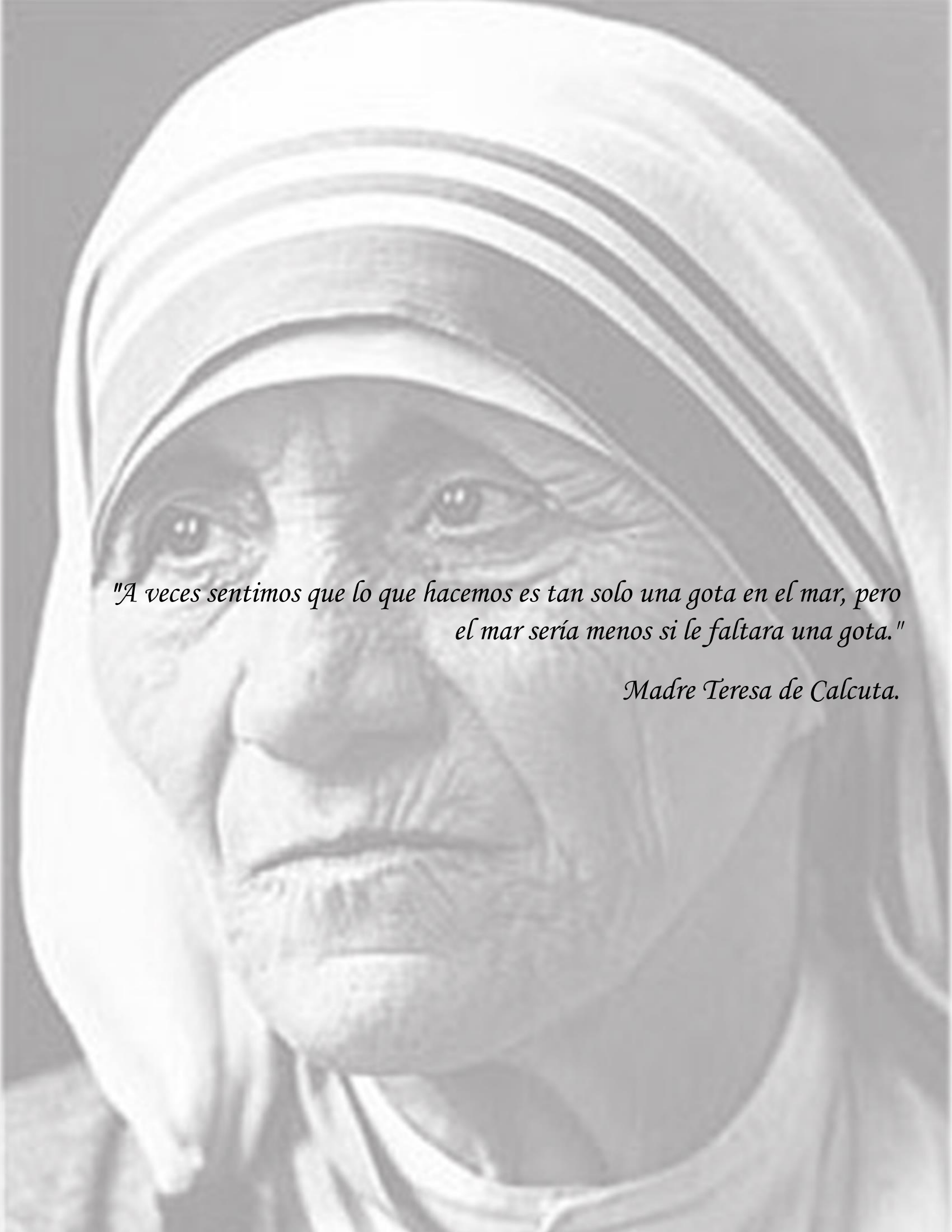
Tutores:

Lic. Yadier Perdomo Cuevas

Ing. Rafael Alberto Quiles Velázquez

La Habana, junio de 2014

Año 56 de la Revolución



"A veces sentimos que lo que hacemos es tan solo una gota en el mar, pero el mar sería menos si le faltara una gota."

Madre Teresa de Calcuta.

Declaración de Autoría

Declaramos que somos los únicos autores de este trabajo de diploma y conferimos a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales del mismo, con carácter exclusivo.

Para que así conste firmamos los presentes a los ___ días del mes de _____ del año _____.

_____

Anays Gómez García

Autor

_____

Gregorio Ferrer Cordova

Autor

_____

Ing. Rafael Alberto Quiles Velázquez

Tutor

_____

Lic. Yadier Perdomo Cuevas

Tutor

Datos del contacto

Tutor: Lic. Yadier Perdomo Cuevas

Licenciado en Ciencias de la computación en el año 2004. Profesor con categoría docente de Asistente, se desempeña como subdirector de tecnología del Centro de Identificación y Seguridad Digital (CISED) desde el año 2011. Posee diez años de experiencia laboral y nueve tutorando trabajos de diploma. Actualmente realiza su tesis de maestría relacionada con el uso de las técnicas de Minería de Datos en los procesos de reconocimiento facial.
Correo electrónico: ypc@uci.cu

Tutor: Ing. Rafael Alberto Quiles Velázquez

Ingeniero en Ciencias Informáticas. Profesor asociado al Centro de Identificación y Seguridad Digital (CISED), pertenece al Departamento de Desarrollo de Componentes, graduado en el curso 2012-2013. Tiene un año de experiencia en el tema Procesamiento de imágenes y señales digitales.
Correo electrónico: raquiles@uci.cu

Dedicatoria

A mi vieja por su amor incondicional y su apoyo en todo momento, eres especial mamá, dedico este trabajo de diploma a mi madre, porque fue la persona que en los momentos más difíciles de mi carrera siempre estuvo ahí para aconsejarme, que luchó conmigo en todas las cosas que necesité simplemente porque es mi madre Gloria Celis, para ti va dedicado este trabajo, y no solamente el trabajo sino más bien mi título porque no es sólo el trabajo de diploma sino que fueron cinco años de intenso estudio en el que nunca faltó un consejo. Simplemente por tu amor incondicional mamá a ti te dedico mi esfuerzo.

Gregorio Ferrer Cordova

A la memoria de mi madre, por su apoyo eterno e incondicional, siempre te estaré agradecida por tanto amor, dedicación y educación.

A mi abuela Silvia, por ser la reliquia de mi familia, como tú quedan pocas, te amo.

Y a mi familia, por estar a mi lado en los momentos más difíciles y los más hermosos de mi vida, los quiero a todos.

Anays Gómez García

Resumen

El reconocimiento facial como una técnica biométrica es uno de los modelos más efectivos para la identificación de personas. La incorporación de estas técnicas a sistemas y ambientes de seguridad es cada vez más común, por sus prestaciones y confiabilidad. Uno de los retos fundamentales que persisten a la hora de la implementación de estos sistemas es conseguir que las máquinas puedan llevar a cabo el análisis de rasgos morfológicos, la identificación en extensas bases de datos y demás tareas de manera rápida y libre de errores. El objetivo principal del trabajo es reducir el espacio de búsqueda en el proceso de reconocimiento facial, llevado a cabo en el Centro de Identificación y Seguridad Digital (CISED) de la Universidad de las Ciencias Informáticas (UCI), mediante el desarrollo de una solución que utilice técnicas de Minería de Datos, para agilizar el proceso de reconocimiento facial. Para lograr el objetivo se desarrolló una solución que emplea las técnicas de Minería de Datos: agrupamiento y clasificación, regidas por una metodología para la extracción del conocimiento (Crisp-DM). Como parte de la naturaleza experimental de la Minería de Datos la solución se basó en el ensayo, probándose con cinco vectores característicos. La solución propone clasificar las imágenes faciales agrupándolas en K grupos y luego aplicar un algoritmo de clasificación que utilice las asignaciones de las imágenes faciales a los grupos obtenidos. La investigación propone un nuevo enfoque basado en el PCA+K-means denominado PFA+K-means.

Palabras clave: agrupamiento, biometría, clasificación, Minería de Datos, reconocimiento facial.

Índice de contenido

Declaración de Autoría III

Datos del contacto IV

Dedicatoria V

Resumen VI

Índice de contenido VII

Índice de figuras IX

Índice de tablas XI

Introducción 1

1. Capítulo 1: Fundamentación Teórica 6

1.1 Introducción 6

1.2 Conceptos fundamentales 6

1.3 Minería de Datos (MD) 7

1.3.1 Tareas de la MD 7

1.4 Reducción de la dimensionalidad. 8

1.5 Técnicas de agrupamiento 10

1.6 Técnicas de clasificación 12

1.7 Metodologías, Herramientas y Tecnologías 16

1.7.1 Metodologías 16

1.7.2 Justificación de las metodologías escogidas 19

1.7.3 Herramientas y Tecnologías 20

1.8 Conclusiones parciales 26

2 Capítulo 2: Análisis y Diseño 27

2.1 Introducción 27

2.2 Propuesta de solución 27

2.3 Modelo de dominio 30

2.4 Aplicando Crisp-DM 31

2.4.1 Fase de Comprensión del negocio 31

2.4.2 Fase de Comprensión de los datos 31

2.4.3 Fase de Preparación de los datos 34

2.4.4 Fase de Modelado 34
2.4.5 Fase de Evaluación 38
2.4.6 Fase de Desarrollo 38
2.5 Aplicando XP 39
2.5.1 Fase de Planificación 39
2.5.2 Fase de Diseño 43
2.6 Conclusiones parciales 49
3 Capítulo 3: Implementación, Validación y Pruebas 50
3.1 Introducción 50
3.2 Implementación 50
3.2.1 Pautas de codificación 50
3.2.2 Tareas de ingeniería 51
3.2.3 Diagrama de componentes 51
3.2.4 Diagrama de despliegue 52
3.3 Validación 53
3.3.1 Agrupamiento 53
3.4 Pruebas 56
3.4.1 Pruebas de rendimiento 56
3.4.2 Pruebas unitarias 57
3.4.3 Pruebas de funcionalidad 59
3.5 Conclusiones parciales 60
Conclusiones 61
Recomendaciones 61
Bibliografía referenciada 62
Bibliografía consultada 65
Glosario de términos 67
Acrónimos 67
4 Anexos 68
4.1 Anexo 1: Variables identificadas en la hipótesis de la investigación. 68

4.2 Anexo 2: Entrevista realizada a Heydi Méndez Vázquez responsable de las investigaciones de reconocimiento facial desarrolladas en el CENATAV. 68

4.3 Anexo 3: Votación del Kdnuggest del 2007 respecto a las metodologías de MD más usadas. 69

4.4 Anexo 4: Descripción de las historias de usuario de la Solución para la clasificación de rasgos biométricos faciales. 69

4.5 Anexo 5: Descripción de las Tarjetas CRC pertenecientes a la capa del negocio. 72

4.6 Anexo 6: Diagrama de clases del diseño. 73

4.7 Anexo 7: Distribución de las Tareas de ingeniería para las iteraciones 2 y 3. 73

4.8 Anexo 8: Especificación de las Tareas ingenieriles. 74

4.9 Anexo 9: Tablas asociadas al procedimiento utilizado para obtener los valores del índice DF-A. 77

4.10 Anexo 10: Resultados de la Validación del agrupamiento. 86

4.11 Anexo 11: Código del algoritmo para entrenar el modelo. 88

4.12 Anexo 12: Casos de Prueba de funcionalidad. 88

Índice de figuras

FIGURA 1.1: RASGOS BIOMÉTRICOS (8). 6

FIGURA 1.2: ORGANIZACIÓN DE UN SISTEMA DE VERIFICACIÓN, COMPUESTO DE DOS FASES, DETECCIÓN DE ROSTROS Y VERIFICACIÓN (9). 7

FIGURA 1.3: PRINCIPALES TÉCNICAS Y MÉTODOS DE LA MINERÍA DE DATOS. 14

FIGURA 1.4: CICLO DE VIDA DE CRISP-DM (41). 18

FIGURA 2.1: FLUJO DE PROCESOS PARA EL AGRUPAMIENTO. 29

FIGURA 2.2: PROPUESTA DE SOLUCIÓN PARA LA CLASIFICACIÓN. 30

FIGURA 2.3: MODELO DE DOMINIO. 30

FIGURA 2.4: PROCEDIMIENTO IMAGEN REESCALADA. 31

FIGURA 2.5: PROCEDIMIENTO IMAGEN ICONO. 32

FIGURA 2.6: PROCEDIMIENTO HSV. 32

FIGURA 2.7: VECINDAD EXTRAÍDA PARA CADA UNO DE LOS PUNTOS ANTES LOCALIZADOS, A LA CUAL SE LE APLICA EL RECORRIDO EN ZIGZAG DANDO COMO SALIDA EL VECTOR CARACTERÍSTICO CORRESPONDIENTE. 33

FIGURA 2.8: FILTROS DE GABOR WAVELET. 33

FIGURA 2.9: DIAGRAMA DE CLASES DEL DISEÑO. 44

FIGURA 2.10: DIAGRAMA ENTIDAD-RELACIÓN. 44

FIGURA 2.11: ARQUITECTURA DE LA SOLUCIÓN PARA LA CLASIFICACIÓN DE RASGOS BIOMÉTRICOS FACIALES. 46

FIGURA 2.12: FILTRO DE LA FASE SELECCIÓN O EXTRACCIÓN DE CARACTERÍSTICAS SEGÚN EL ESTILO ARQUITECTÓNICO TUBERÍAS Y FILTROS. 47
FIGURA 3.1: DIAGRAMA DE COMPONENTES DE LA SOLUCIÓN. 52
FIGURA 3.2: DIAGRAMA DE DESPLIEGUE DE LA SOLUCIÓN PARA LA CLASIFICACIÓN DE RASGOS BIOMÉTRICOS FACIALES. 53
FIGURA 3.3: VALIDACIÓN DEL AGRUPAMIENTO PARA EL PROCEDIMIENTO DCT: TABLA (A) MUESTRA LOS GRUPOS Y EL CVI DF-A, TABLA (B) REPRESENTA LAS ESTADÍSTICAS CALCULADAS A PARTIR DEL MODELO DE REGRESIÓN, TABLA (C) MUESTRA LOS RESULTADOS DEL ESQUEMA BASADO EN VOTOS. 54
FIGURA 3.4: REGRESIÓN LINEAL ASOCIADA AL PROCEDIMIENTO DCT. 54
FIGURA 3.5: VALIDACIÓN DEL AGRUPAMIENTO PARA EL PROCEDIMIENTO GWT: TABLA (A) MUESTRA LOS GRUPOS Y EL CVI DF-A, TABLA (B) REPRESENTA LAS ESTADÍSTICAS CALCULADAS A PARTIR DEL MODELO DE REGRESIÓN, TABLA (C) MUESTRA LOS RESULTADOS DEL ESQUEMA BASADO EN VOTOS. 54
FIGURA 3.6: REGRESIÓN LINEAL ASOCIADA AL PROCEDIMIENTO GWT. 55
FIGURA 3.7: VALIDACIÓN DEL AGRUPAMIENTO PARA EL PROCEDIMIENTO IMAGEN ICONO: TABLA (A) MUESTRA LOS GRUPOS Y EL CVI DF-A, TABLA (B) REPRESENTA LAS ESTADÍSTICAS CALCULADAS A PARTIR DEL MODELO DE REGRESIÓN, TABLA (C) MUESTRA LOS RESULTADOS DEL ESQUEMA BASADO EN VOTOS. 55
FIGURA 3.8: REGRESIÓN LINEAL ASOCIADA AL PROCEDIMIENTO IMAGEN ICONO. 55
FIGURA 3.9: RESULTADOS DE LAS PRUEBAS DE RENDIMIENTO. 56
FIGURA 3.10: GRAFO DE FLUJO DEL MÉTODO PARA REALIZAR EL ENTRENAMIENTO DEL MODELO ASOCIADO AL CLASIFICADOR K-NN. 58
FIGURA 4.1: VOTACIÓN ENTRE LAS METODOLOGÍAS DE MINERÍA DE DATOS MÁS UTILIZADAS. 69
FIGURA 4.2: DIAGRAMA DE CLASES ASOCIADO AL EXTRACTOR DE CARACTERÍSTICAS. 73
FIGURA 4.3: PROCEDIMIENTO DCT. ÍNDICES DE VALIDACIÓN SIN NORMALIZAR PARA CADA CORRIDA DEL K-MEANS. 77
FIGURA 4.4: PROCEDIMIENTO DCT. ÍNDICES DE VALIDACIÓN NORMALIZADOS PARA CADA CORRIDA DEL K-MEANS. 78
FIGURA 4.5: PROCEDIMIENTO GWT. ÍNDICES DE VALIDACIÓN SIN NORMALIZAR PARA CADA CORRIDA DEL K-MEANS. 79
FIGURA 4.6: PROCEDIMIENTO GWT. ÍNDICES DE VALIDACIÓN NORMALIZADOS PARA CADA CORRIDA DEL K-MEANS. 80
FIGURA 4.7: PROCEDIMIENTO IC. ÍNDICES DE VALIDACIÓN SIN NORMALIZAR PARA CADA CORRIDA DEL K-MEANS. 81
FIGURA 4.8: PROCEDIMIENTO IC. ÍNDICES DE VALIDACIÓN NORMALIZADOS PARA CADA CORRIDA DEL K-MEANS. 82
FIGURA 4.9: PROCEDIMIENTO IR. ÍNDICES DE VALIDACIÓN SIN NORMALIZAR PARA CADA CORRIDA DEL K-MEANS. 83

FIGURA 4.10: PROCEDIMIENTO IR. ÍNDICES DE VALIDACIÓN NORMALIZADOS PARA CADA CORRIDA DEL K-MEANS. 84
FIGURA 4.11: PROCEDIMIENTO HSV. ÍNDICES DE VALIDACIÓN SIN NORMALIZAR PARA CADA CORRIDA DEL K-MEANS. 85
FIGURA 4.12: PROCEDIMIENTO HSV. ÍNDICES DE VALIDACIÓN NORMALIZADOS PARA CADA CORRIDA DEL K-MEANS. 86
FIGURA 4.13: VALIDACIÓN DEL AGRUPAMIENTO PARA EL PROCEDIMIENTO IMAGEN REESCALADA: TABLA (A) MUESTRA LOS GRUPOS Y EL CVI DF-A, TABLA (B) REPRESENTA LAS ESTADÍSTICAS CALCULADAS A PARTIR DEL MODELO DE REGRESIÓN, TABLA (C) MUESTRA LOS RESULTADOS DEL ESQUEMA BASADO EN VOTOS. 86
FIGURA 4.14: REGRESIÓN LINEAL ASOCIADA AL PROCEDIMIENTO IMAGEN REESCALADA. 87
FIGURA 4.15: VALIDACIÓN DEL AGRUPAMIENTO PARA EL PROCEDIMIENTO HSV: TABLA (A) MUESTRA LOS GRUPOS Y EL CVI DF-A, TABLA (B) REPRESENTA LAS ESTADÍSTICAS CALCULADAS A PARTIR DEL MODELO DE REGRESIÓN, TABLA (C) MUESTRA LOS RESULTADOS DEL ESQUEMA BASADO EN VOTOS. 87
FIGURA 4.16: REGRESIÓN LINEAL ASOCIADA AL PROCEDIMIENTO HSV. 87

Índice de tablas

TABLA 1.1: PRECISIÓN DE DIFERENTES MÉTODOS EN CUATRO BASES DE DATOS DE IMÁGENES FACIALES. 15
TABLA 2.1: ACTORES DEL SISTEMA. 39
TABLA 2.2: HU CARGAR CONJUNTO DE DATOS DE ENTRENAMIENTO. 41
TABLA 2.3: ESTIMACIÓN DE ESFUERZO POR HU. 42
TABLA 2.4: PLAN DE DURACIÓN DE ITERACIONES. 42
TABLA 2.5: PLAN DE ENTREGA. 43
TABLA 2.6: TARJETA CRC CORRESPONDIENTE A LA CLASE SOLUCION. 43
TABLA 3.1: DISTRIBUCIÓN DE TAREAS DE INGENIERÍA EN LA ITERACIÓN 1. 51
TABLA 3.2: MODELOS MÁS EFICIENTES PARA LOS PROCEDIMIENTOS IC, DCT, GWT, HSV E IR. 57
TABLA 3.3: RESULTADOS DEL CLASIFICADOR RESPECTO A LAS VARIACIONES DE LAS IMÁGENES EN LOS PROCEDIMIENTOS GWT, DCT E IC. 57
TABLA 3.4: CASO DE PRUEBA DE LA FUNCIONALIDAD: CLASIFICAR IMAGEN FACIAL. 60
TABLA 4.1: DESCRIPCIÓN DE LAS VARIABLES DEPENDIENTES E INDEPENDIENTES DE LA HIPÓTESIS. 68
TABLA 4.2: ENTREVISTA REALIZADA EN EL CENATAV A HEYDI MÉNDEZ VÁZQUEZ RESPONSABLE DE LAS INVESTIGACIONES DE RECONOCIMIENTO FACIAL 69
TABLA 4.3: HU_PROCESAR IMAGEN FACIAL 70
TABLA 4.4: HU_GENERAR FICHERO ARFF. 70
TABLA 4.5: HU_ETIQUETAR CONJUNTO DE DATOS DE ENTRENAMIENTO. 70
TABLA 4.6: HU_VISUALIZAR AGRUPAMIENTO. 71

TABLA 4.7: HU_CARGAR IMAGEN FACIAL. 71
TABLA 4.8: HU_CLASIFICAR IMAGEN FACIAL. 71
TABLA 4.9: TARJETA CRC CORRESPONDIENTE A LA CLASE ANALISISCARACTERISTICASPRINCIPALES. 72
TABLA 4.10: TARJETA CRC CORRESPONDIENTE A LA CLASE CLASIFICADOR. 72
TABLA 4.11: TARJETA CRC CORRESPONDIENTE A LA CLASE MODELO AGRUPAMIENTO ARFF. 73
TABLA 4.12: DISTRIBUCIÓN DE TAREAS DE INGENIERÍA EN LA ITERACIÓN 2. 73
TABLA 4.13: DISTRIBUCIÓN DE TAREAS DE INGENIERÍA EN LA ITERACIÓN 3. 74
TABLA 4.14: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_1. 74
TABLA 4.15: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_2. 74
TABLA 4.16: TAREA DE INGENIERÍA 2 CORRESPONDIENTE A LA HU_2. 75
TABLA 4.17: TAREA DE INGENIERÍA 3 CORRESPONDIENTE A LA HU_2. 75
TABLA 4.18: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_3. 75
TABLA 4.19: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_4. 76
TABLA 4.20: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_5. 76
TABLA 4.21: TAREA DE INGENIERÍA 2 CORRESPONDIENTE A LA HU_5. 76
TABLA 4.22: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_6. 76
TABLA 4.23: TAREA DE INGENIERÍA 1 CORRESPONDIENTE A LA HU_7. 77
TABLA 4.24: CASO DE PRUEBA DE LA FUNCIONALIDAD: CARGAR CONJUNTO DE DATOS DE ENTRENAMIENTO. 88
TABLA 4.25: CASO DE PRUEBA DE LA FUNCIONALIDAD: PROCESAR IMAGEN FACIAL. 89
TABLA 4.26: CASO DE PRUEBA DE LA FUNCIONALIDAD: GENERAR FICHERO ARFF. 89
TABLA 4.27: CASO DE PRUEBA DE LA FUNCIONALIDAD: ETIQUETAR CONJUNTO DE DATOS DE ENTRENAMIENTO. 90
TABLA 4.28: CASO DE PRUEBA DE LA FUNCIONALIDAD: VISUALIZAR AGRUPAMIENTO. 90
TABLA 4.29: CASO DE PRUEBA DE LA FUNCIONALIDAD: CARGAR IMAGEN FACIAL. 91
TABLA 4.30: CASO DE PRUEBA DE LA FUNCIONALIDAD: CLASIFICAR IMAGEN FACIAL. 91

Introducción

A lo largo de la historia, el ser humano ha desarrollado distintos métodos para reconocerse y distinguirse entre las demás personas, esto nace debido al interés natural del hombre, por querer protegerse y salvaguardar las cosas que le pertenecen, así como mantener la privacidad en sus acciones. El vertiginoso avance de una disciplina como la biometría, aparejado a los adelantos tecnológicos de las ciencias de la computación ha posibilitado el desarrollado de diversos métodos automáticos para la identificación de los individuos. Los sistemas de seguridad y acceso han evolucionado desde el uso de la huella dactilar, hasta el empleo de otras medidas físicas y de comportamiento (1). Una de estas innovadoras técnicas la constituye el reconocimiento facial, la cual cuenta entre sus muchas aplicaciones: el control de seguridad, de acceso a edificios, la identificación de criminales y las interfaces persona-ordenadores.

La ingeniería comenzó a mostrar interés en el reconocimiento facial a partir de 1960. Uno de los primeros investigadores en incursionar en el tema fue Woodrow W. Bledsoe el cual diseñó e implementó un sistema semiautomático, en el que algunas coordenadas del rostro fueron seleccionadas por un operador humano y luego las computadoras usaban esta información para el reconocimiento, además describió muchos de los problemas (variación de la iluminación, rotación del rostro, la expresión facial, envejecimiento, entre otros) que luego de 50 años el reconocimiento facial aún presenta (2).

A finales de la década, A. Jay Goldstein, Leon D. Harmon y Ann B. Lesk utilizaron 21 características subjetivas, como el color del cabello y grosor de los labios para automatizar el reconocimiento facial (3). En los años 70 se utilizó un enfoque que definía al rostro como un conjunto de parámetros geométricos para luego ejecutar algún reconocimiento de patrones, este enfoque fue utilizado por Kenade en 1973, para desarrollar un sistema completamente automatizado de reconocimiento facial basado en un algoritmo que extraía 16 parámetros faciales automáticamente (2), obteniendo resultados de alrededor del 45 al 75% de efectividad, concluyendo que se podía mejorar estos resultados eliminando características irrelevantes.

La década de los 80 trajo consigo el uso redes neuronales por parte de algunos investigadores en el área del reconocimiento facial. La primera vez que se introdujo el término de *Eigenfaces* en el procesamiento de imágenes, técnica que prevalecería en los siguientes años, fue por L. Sirovich y M. Kirby en 1986, sus métodos se basaban en el Análisis de Componentes Principales y sus trabajos se convertirían en la base de muchos de los nuevos algoritmos propuestos para el reconocimiento facial (2).

Desde los años 90 el avance en el campo del reconocimiento facial ha sido exponencial. Diversos han sido los enfoques propuestos que pueden alcanzar razonables rendimientos en términos de las tasas de reconocimiento, sin embargo el tiempo de cálculo para realizar la identificación de un rostro, a partir de una base de datos será proporcional al tamaño de la misma.

En un sistema de reconocimiento facial, cada rostro humano es preprocesado para realizar la extracción de características formándose un vector característico distintivo de cada persona, que será utilizado en el reconocimiento. El vector es almacenado para cuando una persona quiera ser identificada, pase por el mismo procedimiento de la extracción de sus características y sea comparado con cada uno de los vectores de la base de datos. La similitud entre el rostro de entrada y los rostros en la base de datos es medida por la distancia entre sus respectivos vectores. Luego el tiempo de corrida requerido para el proceso de reconocimiento, puede ser considerablemente reducido si el número de imágenes faciales a comparar es pequeño (4).

Con el objetivo de minimizar el problema de búsqueda de uno a muchos presente en los sistemas de reconocimiento facial donde el espacio muestral lo constituyen todas las imágenes faciales almacenadas, se pueden utilizar técnicas de Minería de Datos para tratar de sectorizar el espacio de búsqueda en pos de mejorar el proceso de reconocimiento. Las técnicas de Minería de Datos (5) intentan obtener patrones o modelos a partir de datos recopilados, las mismas se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o descriptivas.

En la actualidad son muchas las organizaciones, empresas internacionales y nacionales de la industria de las tecnologías biométricas, que dedican su trabajo diario al desarrollo de nuevas tecnologías en el campo del reconocimiento facial. En el ámbito nacional se destaca el Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), en el mismo se desarrollan investigaciones teóricas y aplicadas en el Reconocimiento de Patrones y Minería de Datos.

Otros centros en el país han incursionado en este tipo de estudio, entre los que se encuentra la Universidad de las Ciencias Informáticas (UCI) y en ella el Centro de Identificación y Seguridad Digital (CISED), donde se estudian los procesos para el reconocimiento de personas mediante sus rasgos biométricos. El CISED cuenta con un sistema de reconocimiento facial, que realiza la verificación e identificación de un individuo mediante la búsqueda en una base de datos del vector característico[1] más similar al de la persona analizada, siendo esta base de datos densa ya que los vectores utilizados presentan alta dimensionalidad (68x9 donde cada una de las 68 características extraídas para una imagen facial, tiene asociada un vector de 9 coeficientes) y la universidad cuenta con un total de personas enroladas del orden de las 10^4. Se ha constatado a través del criterio de los desarrolladores de la aplicación Ing. Rafael Alberto Quiles e Ing. Amelia Aguilera Reyes, por medio de pruebas de rendimiento en cuanto al tiempo realizadas a la misma, que para este conjunto de datos el proceso de reconocimiento demora un total de 4.27 segundos, realizando una comparación con todas las personas enroladas en la base de datos, trayendo como consecuencia

[1] En un sistema de reconocimiento facial un vector característico se refiere a un vector n-dimensional distintivo para cada persona, que contiene características del rostro, expresadas mediante valores cuantitativos.

demoras en la respuesta del subproceso de identificación y lentitud en el proceso, provocando aglomeraciones continuas en el punto de control de acceso al centro.

Ante esta dificultad se plantea como **problema científico** ¿Cómo mejorar los tiempos de respuesta en el proceso de reconocimiento facial llevado a cabo en el CISED?

Para dar solución al problema antes mencionado se define como **objeto de estudio** los procesos de clasificación de rasgos biométricos faciales.

Como una alternativa de solución al problema planteado, se define como **objetivo general** de la investigación, reducir el espacio de búsqueda en el proceso de reconocimiento facial llevado a cabo en el CISED, a través del desarrollo de una solución que utilice técnicas de Minería de Datos, para agilizar el proceso de reconocimiento facial.

La investigación se rige por la siguiente **hipótesis:** el desarrollo de una solución que agrupe y clasifique los rostros, mejorará los tiempos de respuesta del proceso de reconocimiento facial llevado a cabo en el CISED. La descripción de las variables identificadas en la hipótesis se muestra en los anexos. (Ver Anexo 1)

Los **objetivos específicos** derivados del objetivo general son los siguientes:

- Analizar los aspectos teórico-prácticos referentes a los sistemas de reconocimiento facial.
- Identificar las variantes de solución existentes en el mundo al problema planteado.
- Identificar las técnicas de Minerías de Datos utilizadas en los procesos de agrupamiento y clasificación.
- Diseñar la solución para la clasificación de rasgos biométricos faciales.
- Implementar la solución para la clasificación de rasgos biométricos faciales.
- Realizar pruebas a la solución para la clasificación de rasgos biométricos faciales.

Las **tareas de investigación** definidas para darle respuesta al objetivo de la investigación son:

- Definición de la fundamentación teórica del proceso de clasificación de rasgos biométricos faciales.
- Análisis de las tendencias actuales para la clasificación de rasgos biométricos faciales.
- Definición de los algoritmos de Minería de Datos a utilizar en la solución para clasificar rasgos biométricos faciales.
- Implementación de los algoritmos de Minería de Datos a utilizar en la solución.

- Definición de las metodologías, herramientas y tecnologías para implementar la solución de la investigación.
- Validación de los algoritmos utilizados en la solución mediante métodos de validación empleados en problemas de Minería de Datos.

El **método científico** de investigación es la forma de abordar la realidad, de estudiar los fenómenos de la naturaleza, la sociedad y el pensamiento, con el propósito de descubrir la esencia de los mismos y sus relaciones internas (6), en la presente investigación se utilizarán algunos de ellos según su clasificación: teóricos y empíricos, o métodos particulares.

Los **métodos empíricos** a utilizar durante la investigación son los siguientes:

- **Entrevistas** con el objetivo de conocer el criterio de expertos en el tema tratado en la presente investigación. (Ver Anexo 2)
- **Experimentación** para probar varios procedimientos para la extracción de los vectores característicos, como parte de la naturaleza experimental de la Minería de Datos basada en el ensayo.

Por otro lado los **métodos teóricos** que serán utilizados en el análisis e interpretación de la información obtenida son:

- **Analítico-Sintético** para consultar la bibliografía relacionada con el tema de la presente investigación.
- **Histórico-Lógico** con el objetivo de indagar sobre el avance alcanzado en el desarrollo de soluciones para la clasificación de rasgos biométricos faciales y los aportes que han brindado a la sociedad.
- **Hipotético-Deductivo** para arribar a conclusiones a partir de la hipótesis, que después pueden ser comprobadas experimentalmente.
- **Modelación** para modelar los diagramas de diseño de la solución para un mejor entendimiento.

La **justificación de la investigación** se fundamenta en que forma parte de los esfuerzos del centro para crear soluciones propias, representando un paso importante en la búsqueda de nuevas alternativas a las soluciones existentes. El espacio poblacional en un sistema de reconocimiento es creciente en el tiempo, por lo cual, los algoritmos de búsqueda y comparación poseen un alto costo computacional, por tanto se hace imprescindible brindarle una solución a este problema; en la investigación se propone la variante de agrupar y clasificar los rostros, con el objetivo de mejorar los tiempos de respuesta del sistema de reconocimiento facial existente en el centro, a través de la reducción del espacio de búsqueda en el proceso

de reconocimiento facial. Por otra parte la adquisición de *hardware* especializado con el objetivo de no depender del *software* requeriría un gasto económico para la institución.

El presente trabajo de diploma tiene estructurado el contenido en tres capítulos de la siguiente manera:

- **Capítulo 1: Fundamentación Teórica,** contiene los conceptos fundamentales relacionados con la investigación, un estudio del estado del arte sobre los algoritmos de Minería de Datos y las variantes existentes al problema, así como las metodologías, tecnologías y herramientas que serán utilizadas en el desarrollo de la solución.
- **Capítulo 2: Análisis y Diseño,** en este capítulo se presentan las características de la solución, el diseño de clases, así como los requerimientos funcionales y no funcionales identificados. Además quedan plasmadas las primeras fases del desarrollo de la solución de acuerdo a las metodologías seleccionadas.
- **Capítulo 3: Implementación, Validación y Pruebas,** este capítulo contiene los artefactos más importantes relacionados con la implementación de la solución, se validan los algoritmos utilizados, y se realizan pruebas de efectividad y funcionalidad.

1. Capítulo 1: Fundamentación Teórica

1.1 Introducción

En el presente capítulo se hace una breve descripción de los conceptos fundamentales relacionados con el tema de investigación, además un estudio del estado del arte sobre los algoritmos utilizados para la reducción de la dimensionalidad, de Minería de Datos (MD) más a fines con la problemática antes planteada y las principales variantes existentes en el mundo al problema planteado. Son seleccionadas las metodologías, las tecnologías y herramientas a utilizar en la investigación de acuerdo a las particularidades de la misma.

1.2 Conceptos fundamentales

- Biometría: es un término general utilizado alternativamente para describir una característica o un proceso. Como una característica: la biometría es una característica biológica (anatómica y psicológica) y de comportamiento que puede medirse y utilizarse en el reconocimiento autómata. Como un proceso: la biometría es un método automático de reconocimiento de individuos, basado en características biológicas y de comportamiento que se pueden medir (7).

Figura 1.1: Rasgos biométricos (8).

- Sistema biométrico: sistema automatizado utilizado por los dispositivos biométricos, teniendo en cuenta determinado ambiente o circunstancia, para realizar la verificación e identificación de un individuo a partir de sus rasgos físicos o de conducta, dependiendo de cuál sea usado por el dispositivo biométrico (1).

- Reconocimiento facial: proceso realizado, por una persona o un computador, para identificar un individuo a través de su rostro como rasgo físico de la biometría (9).

- Sistema de reconocimiento facial: es una aplicación dirigida por ordenador que identifica automáticamente a una persona en una imagen digital, al analizar las características faciales extraídas de la imagen y realizar una comparación con las características previamente almacenadas en la base de datos (9). El proceso realizado por este tipo de sistemas se muestra en la figura 1.2.

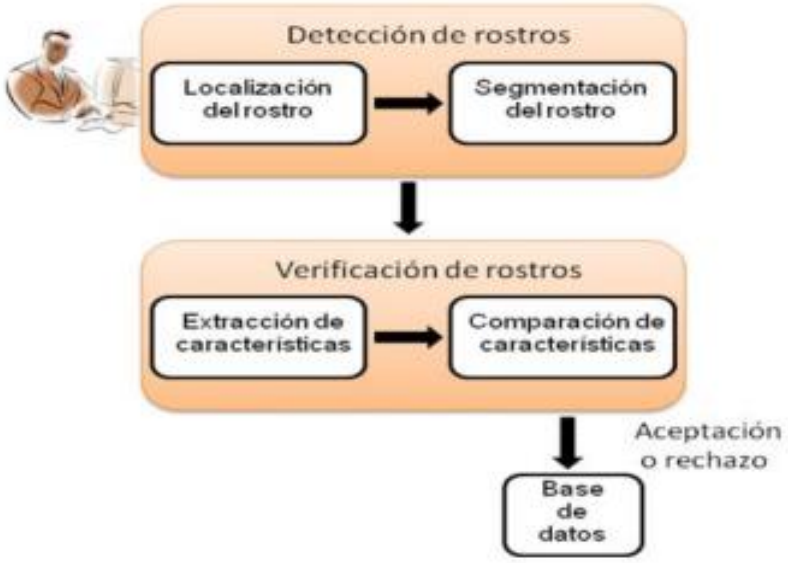


Figura 1.2: Organización de un sistema de verificación, compuesto de dos fases, detección de rostros y verificación (9).

- Minería de Datos: proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (5).

1.3 Minería de Datos (MD)

1.3.1 Tareas de la MD

Dentro de la MD se ha de distinguir tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de MD. Esto significa que cada una tiene sus propios requisitos y que el tipo de información obtenida con una tarea, puede diferir mucho de la obtenida con otra (5). Las mismas se clasifican en predictivas y descriptivas.

Predictivas

Se encarga de estimar valores futuros o desconocidos de una variable de interés (variable objetivo o dependiente) utilizando otras variables o campos de la base de datos (variables independientes o predictivas).

- **Clasificación:** predecir la clase de nuevas instancias de las que se desconoce su valor. Para ello, se tiene en cuenta que cada instancia (o registro de la base de datos) pertenece a una clase, indicándose a través del valor de un atributo llamado la clase de la instancia. Este

atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase (5). El resto de los atributos de la instancia se utilizan para predecir la clase.

- **Regresión:** consiste en aprender una función real que asigna a cada instancia un valor real. Se diferencia de la clasificación, en que el valor a predecir es numérico.

Descriptivas

Este tipo de tareas identifica patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos (5).

- **Agrupamiento:** es la tarea descriptiva por excelencia y consiste en obtener grupos naturales a partir de los datos. A diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta (5). El proceso de agrupamiento se realiza teniendo en cuenta el hecho de maximizar la distancia (o similitud) entre grupos y minimizar la distancia entre elementos de un mismo grupo.
- **Correlación:** se usa para examinar el grado de similitud de los valores de dos variables numéricas (5), teniendo en cuenta un valor r entre 1 y -1, lo que significa que si r toma valor 1 las variables están perfectamente correlacionadas, -1 están correlacionadas negativamente y en caso de tomar valor 0 no están correlacionadas.
- **Asociación:** muy similar a las correlaciones, tiene como objetivo identificar relaciones no explícitas entre atributos categóricos. Este tipo de tarea descriptiva tiene como característica el no ser necesario una relación causa-efecto para que los datos estén asociados (5).

El desarrollo de la presente investigación estará enfocado, en la clasificación y el agrupamiento, como técnicas de la MD, con el objetivo de alcanzar la meta propuesta para el trabajo.

1.4 Reducción de la dimensionalidad.

En muchos problemas prácticos de Minería de Datos, existen demasiadas características donde la inmensa mayoría representa información redundante o ruido (10), afectando significativamente en el rendimiento de los esquemas de aprendizaje, por lo que es necesario una selección o transformación de las características originales, para obtener un conjunto de características que sean lo más discriminante posibles entre sí. Este problema es conocido como “la maldición de la dimensionalidad” y es muy frecuente en el análisis de datos multivariantes[2] (11), donde la alta dimensionalidad afecta en la obtención de patrones útiles y robustos. Los procesos asociados a la resolución de este problema, constituyen un tipo de ingeniería de datos (10) y pueden aumentar considerablemente la

[2] El análisis multivariante clásico engloba, tradicionalmente, una serie de técnicas estadísticas para tratar con un conjunto de variables, que tienen sus orígenes en el álgebra lineal y la geometría, por esta razón, pese a su diversidad, se agrupan bajo el mismo término (5).

probabilidad de éxito de las técnicas de aprendizaje automático que son aplicadas a problemas reales de Minería de Datos. Experimentos arrojan que la intromisión de atributos insignificantes afecta el rendimiento de esquemas de aprendizaje, tales como los árboles de decisión, regresión lineal, aprendices basados en instancias y métodos de agrupamiento (10).

A continuación son explicados un conjunto de métodos, utilizados para la reducción de la dimensionalidad por la comunidad científica, en problemas similares al planteado en la investigación.

Análisis de Componentes Principales (PCA[3])

PCA es una técnica estadística que se basa en conceptos algebraicos, utilizada para la reducción de la dimensionalidad en el análisis de datos multivariantes. Con PCA se logra disminuir la dimensionalidad de los datos, proyectándolos en un espacio definido por $K < D$ componentes principales (12), siendo D la cantidad de características originales. La idea es obtener una base ortonormal compuesta por K vectores (componentes principales), que representen las direcciones de máxima varianza en el espacio de características original y proyectar los datos en ésta (13).

La propiedad esencial que se busca con PCA es la independencia estadística entre las nuevas características obtenidas (14). Sin embargo, esto es mucho más complejo de lograr que con una simple transformación lineal (14), en lugar de eso la matriz de covarianza que se obtiene es diagonal. Esto en cierta forma garantiza la eliminación de información redundante o ruido en los datos originales. Aunque bajo ciertas condiciones (14) se puede lograr que transformaciones que diagonalizen la matriz de covarianza también propicien esta propiedad. Esta técnica tiene una amplia variedad de aplicaciones, adaptaciones y denominaciones en otras áreas fuera de la estadística, lo que la convierte en una herramienta poderosa.

Para ampliar en el funcionamiento y fundamentos matemáticos del PCA se pueden consultar los siguientes artículos (13) (14) (15).

Análisis de Características Principales (PFA[4])

Una de las desventajas que presenta PCA, es la utilización de las medidas de todas las características originales (16) en la proyección a un espacio de menor dimensión. En (16) se propone un método para la reducción de la dimensionalidad, basado en la selección de un conjunto de características que contengan la información más relevante usando como base PCA. Sea A_q una matriz formada por los primeros q componentes principales y $V_1, V_2, \dots, V_n \in R^q$ las filas de esa matriz, donde cada vector V_i representará la proyección de la i-ésima característica a un espacio de menor dimensión (16), luego

[3] Por sus siglas en inglés: *Principal Components Analysis.*
[4] Por sus siglas en inglés: *Principal Feature Analysis.*

las características más relacionadas entre sí serán aquellas cuyos valores absolutos de V_i sean más similares. Para la elección del mejor conjunto se procede a encontrar los subconjuntos de características más relacionadas entre sí y elegir una característica de cada subconjunto, con lo que las características elegidas representarán a cada grupo óptimamente (16) en términos de alta varianza e insensibilidad al ruido.

En (16) se realizan un conjunto de experimentos donde se concluye que, el PFA tiene rendimiento comparable con el PCA en aplicaciones como la recuperación de imágenes en base de datos extensas, presentándose como principal ventaja del PFA que la selección del espacio de menor dimensión, se realiza a través de la selección de un subconjunto de características del espacio original y no utilizando todas las características originales como sucede con PCA.

Análisis Discriminante Lineal (LDA[5])

El LDA (17) es una técnica estadística utilizada para la reducción de la dimensionalidad, pero a diferencia de los métodos explicados anteriormente es una técnica supervisada, es decir que trabaja con conjunto de datos etiquetados. Su idea principal se basa en proyectar los datos en un subespacio donde mejor se discrimine las clases. El objetivo (12) es encontrar las direcciones que minimicen la varianza intraclase y maximicen la varianza interclase. Según (18) (19) este problema conduce al cálculo de los valores y vectores propios de $S_w^{-1}S_b$, donde S_w representa la matriz de covarianza intraclase y S_b la matriz de covarianza interclase.

A pesar de la tendencia de utilizar LDA sobre PCA en procesos de clasificación, se ha probado en estudios realizados (20), que bajo ciertas condiciones el PCA puede funcionar mejor que el LDA.

1.5 Técnicas de agrupamiento

El agrupamiento es la técnica descriptiva de Minería de Datos que tiene como objetivo, encontrar una partición de los datos en la cual los objetos pertenecientes a un mismo grupo sean homogéneos, mientras que aquellos que pertenecen a grupos diferentes sean lo más separados posibles entre sí (21).

K-medias[6]

El algoritmo K-means (21) es uno de los algoritmos de agrupamiento más conocido y popular, que intenta encontrar una partición óptima de los datos minimizando el criterio de la suma del error

[5] Por sus siglas en inglés: *Lineal Discriminant Analysis.*
[6] Del inglés: *K-means.*

cuadrático (21), a través de un procedimiento iterativo. El algoritmo pertenece a la categoría de los algoritmos escaladores de colinas[7]. El procedimiento estándar (21) del K-means es el siguiente:

1. Seleccionar K puntos iniciales como centroides[8].
2. Asignar cada instancia al centroide más cercano.
3. Recalcular el centroide de cada grupo.
4. Repetir pasos 2 y 3 hasta que los centroides no varíen.

La complejidad computacional del K-means es $O(NKdT)$, donde N representa la cantidad de datos, K el número de grupos, d la dimensión de los datos y T el número de iteraciones; como K, d y T son menores que N en la práctica, se puede afirmar que el K-means es aproximadamente lineal.

Una de las debilidades propias de los escaladores de colinas y por ende que presenta el K-means, es que no garantiza obtener el óptimo global por lo que la elección de los centroides iniciales, determinará la calidad de la solución de agrupamiento final. Aunque existen muchas propuestas enfocadas a mejorar esta selección inicial del K-means, no existe teóricamente un método eficiente para determinar esta partición inicial (21). Una de las estrategias generalmente utilizadas para minimizar la influencia de los puntos iniciales escogidos, es correr el algoritmo varias veces y elegir la mejor solución de agrupamiento. Otro problema que afecta a este algoritmo, es la necesidad de definir a priori el número de grupos presentes en los datos, cuestión que en la práctica generalmente no se conoce, sin embargo existen investigaciones y métodos (21) (22) (23) propuestos para lidiar con este problema.

A pesar de sus debilidades, K-means es el algoritmo de agrupamiento particional más utilizado en la práctica, por su simplicidad, fácil entendimiento y adaptabilidad a diferentes escenarios. Además los continuos mejoramientos y generalizaciones del algoritmo estándar, han garantizado su relevancia y considerable incremento de efectividad (22).

Algoritmo EM[9]

El agrupamiento basado en probabilidad puede mejorar algunas deficiencias (10) presentes en los enfoques heurísticos. La perspectiva probabilística del agrupamiento, se centra en determinar el conjunto de grupos más probables presentes en los datos (10). Su principal enfoque radica en que las instancias no pertenecen a un grupo en particular, sino más bien tienen una cierta probabilidad de pertenencia a cada uno de los grupos existentes. El fundamento matemático detrás del EM es un

[7] Son algoritmos que tratan de elegir en cada paso un estado cuyo valor heurístico sea mejor que el estado activo en ese momento.
[8] Se denomina centroide a la media (o media ponderada) de los puntos que pertenecen a un determinado grupo.
[9] Por sus siglas en inglés: *Expectation Maximization.*

modelo estadístico denominado mezcla finita[10], siendo una mezcla un conjunto de k distribuciones de probabilidad representando k grupos diferentes (10).

El algoritmo EM (5) (24) permite la estimación de parámetros en modelos probabilísticos. El procedimiento (5) es similar al del K-means, primero se realizan unas estimaciones iniciales de los parámetros del modelo (media, desviación típica) o también se pueden suponer asignaciones de las instancias a determinadas clases, entonces se utiliza esta información para estimar las probabilidades de pertenencia de cada instancia a cada uno de los grupos y se reestiman los parámetros, este procedimiento se repite hasta su convergencia la cual está teóricamente garantizada (5). Este algoritmo alterna entre dos fases, la fase *E* en la cual se calculan las probabilidades de pertenencia a los grupos para cada instancia y la fase *M* en donde se reestiman los parámetros del modelo con el objetivo de maximizar la verosimilitud de las distribuciones. Este procedimiento puede ser visto como la resolución de una secuencia de subproblemas de optimización, que son elegidos de forma tal que garanticen sus correspondientes soluciones y puedan converger a un óptimo local como solución final (24).

Aunque está garantizada la convergencia del algoritmo a un máximo, el mismo no es necesariamente global, con el objetivo de mejorarlo se recomienda repetir el procedimiento muchas veces con diferentes inicializaciones para los parámetros del modelo (10).

Para ampliar los conocimientos sobre los fundamentos del algoritmo EM se pueden consultar las bibliografías (5) (10) (21) (22) (24) (25).

Aunque el K-means y el EM fueron los métodos de agrupamiento analizados en la investigación existen otras propuestas que pueden ser consultadas en (26) (27) (28) (29) (30) (31).

1.6 Técnicas de clasificación

La clasificación es la técnica predictiva que se fundamenta en la predicción del valor de un atributo discreto denominado clase. El objetivo es inducir un modelo para poder predecir el valor de la clase dados los valores de los demás atributos (32).

K vecinos más cercanos[11]

El K-NN (5) (22) (32) es un método de aprendizaje automático basado en vecindad. Su principal característica (32) es la no existencia de un modelo asociado al aprendizaje, sino que las predicciones se realizan en función de la información extraída de un conjunto de ejemplos existentes previamente (5), por lo que generalmente se refieren a éste como un método retardado. La principal

[10] Del inglés: finite mixtures.
[11] Del inglés: *K-nearest neighbors (K-NN).*

diferencia entre los métodos retardados y los no retardados, es la relacionada a que el mayor costo computacional de los métodos retardados se centra en la etapa de la predicción, a diferencia de los no retardados que su fase más costosa es la del entrenamiento, donde se obtiene un modelo de generalización a partir del conjunto de ejemplos. El funcionamiento del algoritmo (22) es bastante simple, cuando un nuevo caso es presentado, se determinan los k vecinos más cercanos del conjunto de entrenamiento y se etiqueta al nuevo caso como perteneciente a la clase más numerosa dentro de la vecindad.

Existen algunos inconvenientes (22) (32) que afectan el rendimiento del K-NN y uno de ellos es la elección del k. Si la vecindad es muy pequeña los resultados pueden ser sensibles al ruido, mientras que la elección de una vecindad muy grande puede provocar la inclusión de muchas instancias de otras clases. En (22) se propone la resolución de este problema utilizando validación cruzada. Otro inconveniente es el relacionado con la elección de la clase más probable en la vecindad, esto pudiera traer consigo un problema si los vecinos más cercanos varían considerablemente en sus distancias, siendo los vecinos más próximos los más confiables para indicar la clase del objeto (22). Una solución al problema es penalizar aquellos vecinos más lejanos dentro de la vecindad. La elección de la medida de distancia es otra cuestión a tener en cuenta, aunque existen muchas medidas de distancia la más adecuada es aquella para la cual, una distancia pequeña entre dos objetos implique mayor probabilidad de pertenencia a una misma clase (22).

Aunque es una técnica de clasificación bastante fácil de entender e implementar, presenta buen rendimiento en muchas situaciones (22).

Existen diversas propuestas para realizar la tarea de clasificación, algunas más complejas que la explicada anteriormente y pueden ser consultadas en (5) (10) (22) (33) (34).

En la figura 1.3 se muestra mediante un mapa conceptual un resumen de las principales técnicas y métodos utilizados en MD.

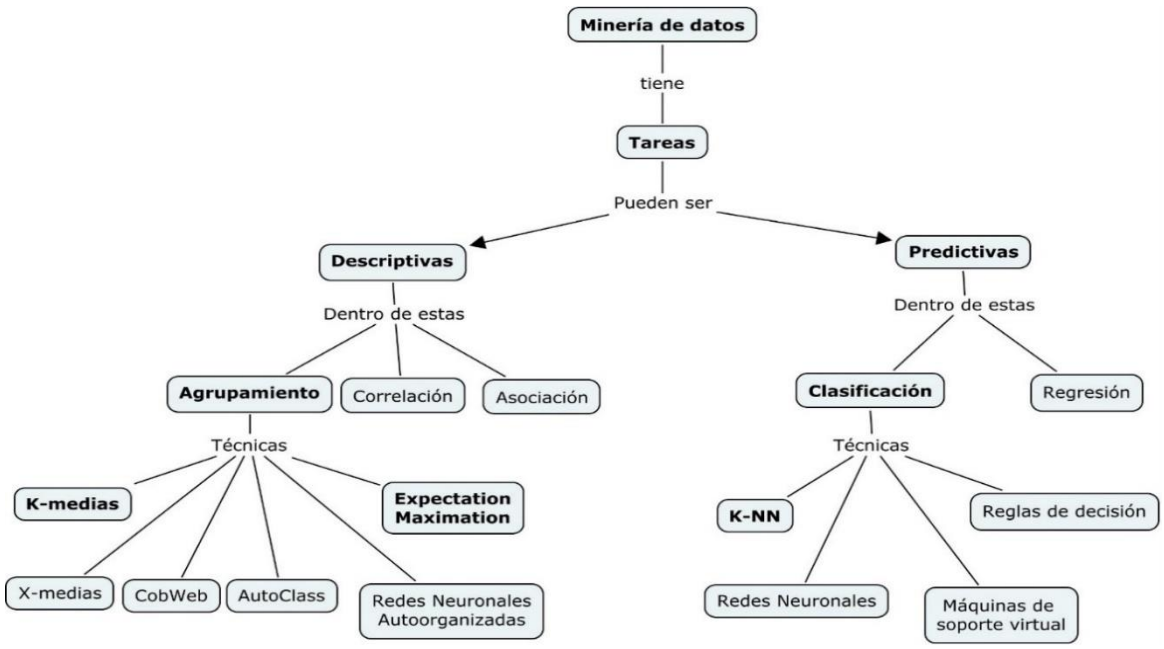


Figura 1.3: Principales técnicas y métodos de la Minería de Datos.

Variantes de solución existentes en el mundo

El agrupamiento de imágenes faciales es una línea de investigación utilizada con éxito en muchos campos, por ejemplo en la recuperación de imágenes faciales, si se pudiera agrupar diferentes tipos de imágenes, entonces se pudieran determinar similitudes presentes en varias de ellas, además de que el costo computacional de búsqueda de una imagen de interés se reduciría considerablemente (35).

Muchas han sido las variantes propuestas para realizar agrupamiento en rostros humanos, una manera directa de solucionar el problema es utilizar las técnicas tradicionales de agrupamiento como el K-means (22), sin embargo su aplicación en agrupamiento de imágenes faciales se ve limitada por la alta dimensionalidad presente en los vectores de representación (generalmente, el número de píxeles de la imagen). Para solucionar este problema se utiliza el PCA para reducir la dimensionalidad de los datos y luego aplicar el K-means, con lo que surge así una de las propuestas más famosas denominada PCA+K-means (35). Esta propuesta está sustentada en las ventajas que presenta el PCA en la reducción de ruido en los datos y en la estrecha relación que existe con el algoritmo K-means (36) (37) mostrándose notable rendimiento en el agrupamiento cuando los datos son reducidos con PCA, además de que la reducción de la dimensionalidad con esta técnica, descubre automáticamente un subespacio de agrupamiento (36) lo que demuestra porqué es beneficioso para el algoritmo K-means, ya que el agrupamiento en este subespacio es más efectivo que en el espacio original (36). En (38) se propone un enfoque que utiliza el método de reducción de dimensionalidad LDA en combinación con el K-means. Este enfoque se denomina LDA-Km, el

cual integra el proceso de agrupamiento con el de selección de un subespacio, en donde los datos simultáneamente son agrupados mientras se realiza la selección de un subespacio de características. Por último en (35) se propone un enfoque más novedoso denominado SCTR[12], que integra individualidades espaciales y estructuras de múltiples imágenes faciales para obtener un agrupamiento óptimo y realizar el aprendizaje de un subespacio de manera simultánea. En (35) se realiza una comparación de las variantes mencionadas anteriormente, el resultado de esta comparación se muestra en la tabla 1.1. La medida utilizada para realizar la comparación es la precisión (38), que relaciona las imágenes agrupadas en un determinado grupo con la clase a la que pertenecen, midiendo el grado en que cada grupo contiene a las imágenes de una determinada clase. En el caso de agrupamiento de imágenes faciales la clase se define como las imágenes que pertenecen a una misma persona.

Base de Datos	K-means	PCA+K-means	LDA-Km	SCTR
Yale	0.379	0.385	0.459	0.491
ORL	0.497	0.513	0.607	0.674
Umist	0.381	0.383	0.468	0.501
Feret	0.374	0.372	0.526	0.551

Tabla 1.1: Precisión de diferentes métodos en cuatro bases de datos de imágenes faciales.

Variante propuesta en la investigación

La variante propuesta está basada en el enfoque estándar **PCA+K-means** que se denominará **PFA+K-means**. El término PFA se refiere a la utilización de este método en la reducción de la dimensionalidad de los datos en vez de PCA como se propone inicialmente, esta elección se fundamenta en que la selección del espacio de menor dimensión se realiza a través de la selección de un subconjunto de características del espacio original, que contengan la información más relevante y no utilizando todas las características originales como sucede con PCA (16). En (16) se realizan experimentos que evidencian que el método presenta similar rendimiento al PCA en problemas de recuperación de imágenes faciales en extensas bases de datos.

Por su parte la elección del K-means es la más adecuada ya que, la mayoría de las propuestas estudiadas utilizan este algoritmo en su solución, tomándose en cuenta la fuerte relación existente entre el PCA y el K-means (36) (37). Aunque no es el PFA el que se aborda en (36) (37), el mismo tiene sus fundamentos en el PCA y presenta similares propiedades (16).

La propuesta está enfocada como un problema de Minería de Datos a resolver, por lo que se utilizará una metodología para la extracción del conocimiento (Crisp-DM) dirigiendo el proceso de solución,

[12] Por sus siglas en inglés: *Subspace Clustering Via Trace Ratio.*

fundamentándose en la naturaleza experimental presente en la resolución de estos problemas, con el objetivo de incorporar elementos de Minería de Datos a la propuesta inicial PCA+K-means. Otra de las cuestiones a tener en cuenta es el conjunto de datos que serán utilizados en la solución, que estará conformado por las imágenes faciales de la población perteneciente a la UCI. La universidad cuenta con una imagen facial por persona evidenciándose en (20) que esta propiedad puede conllevar a que el PCA tenga mejor rendimiento que el LDA.

Aunque el enfoque SCTR es el de mejores resultados, el equipo de investigación decidió no probar con este algoritmo ya que el objetivo fundamental es resolver el problema utilizando las principales técnicas, tareas y herramientas de MD. El enfoque no está incluido en las herramientas de MD Weka y RapidMiner las cuales fueron las herramientas analizadas (ver epígrafe 1.7.3), por lo que su uso supondría implementación por parte de los investigadores, siendo el mismo de muy alta complejidad. Además un análisis detallado abarcaría un tiempo considerable para llegar a un desarrollo del mismo, siendo mayor que el tiempo con que se cuenta para realizar la investigación.

1.7 Metodologías, Herramientas y Tecnologías

1.7.1 Metodologías

Se define como metodología al conjunto de métodos que se siguen en una investigación científica o en una exposición doctrinal (39). Llevada al campo de la informática, una metodología no sólo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevar a cabo las mismas.

Metodología para la extracción del conocimiento

Durante la presente investigación se utilizará la metodología Crisp-DM, como guía principal del proceso de extracción del conocimiento. Esta metodología, es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de MD (Ver Anexo 3). Propone cuatro niveles de abstracción, organizados de forma jerárquica entre tareas que van desde lo más general hasta lo más específico. Profundiza en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de MD, proponiendo un ciclo de vida divido en seis fases (ver figura 1.4):

- **Comprensión del negocio:** aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto (40). En esta fase, es de suma importancia la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de MD y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio.
- **Comprensión de los datos:** se recolectan los datos iniciales, estos son descritos y explorados, para luego verificar la calidad de los mismos. Esta fase junto a la preparación de

los datos y el modelado, son las fases del ciclo de vida del proceso de MD que requieren de mayor esfuerzo por parte del equipo de desarrollo.

- **Preparación de los datos:** obtiene la vista minable[13] de los datos luego de hacer una selección, limpieza, construcción, integración y formateo de los datos. Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente (40).
- **Modelado:** el equipo de desarrollo selecciona las técnicas de MD que más se ajustan en cuanto al problema que se necesita resolver, la disponibilidad de los datos adecuados, el cumplimiento de los requisitos, así como el tiempo en que puede obtenerse un modelo y el conocimiento de la técnica.
- **Evaluación:** la tarea principal de esta fase, es la evaluación de los datos teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error (40).
- **Desarrollo:** en esta fase se generan el plan de implantación para desplegar el resultado de la MD en la organización, tomando los resultados de la evaluación, y concluye con una estrategia para su implantación; el plan de monitoreo y mantenimiento para trazar estrategias de monitorización y mantenimiento que serán aplicadas a los modelos; el informe final donde se recogen las conclusiones finales del proyecto de MD desarrollado; y se realiza la revisión del proyecto definiendo que fue lo que se hizo correcta o incorrectamente y lo que requiere de mejoras.

[13] Se le denomina vista minable al subconjunto de los datos sobre el que se va a aplicar una técnica de MD.

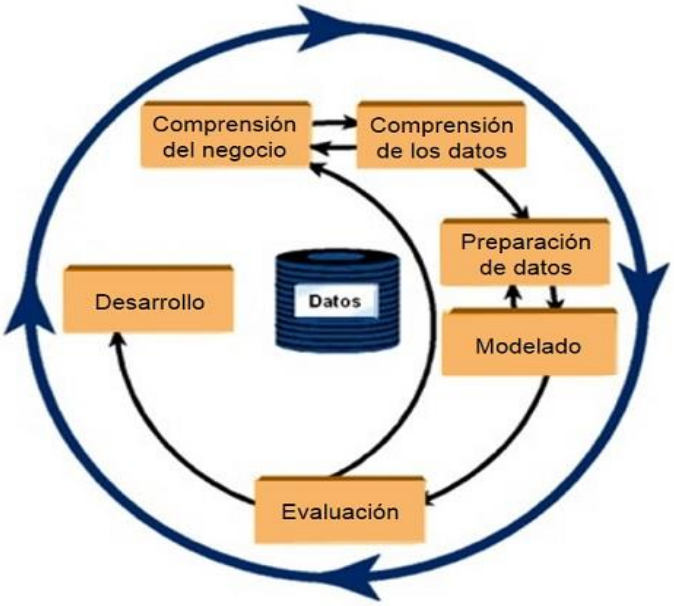


Figura 1.4: Ciclo de vida de Crisp-DM (41).

Metodología para el desarrollo del software

Las metodologías para el desarrollo del software son un conjunto de procedimientos, técnicas y ayudas a la documentación para el desarrollo de productos informáticos (39). Durante la presente investigación se escogió como metodología para el desarrollo del software: XP.

Esta metodología intenta reducir la complejidad del software por medio de un trabajo orientado al objetivo, basado en las relaciones interpersonales y la velocidad con la que trabaja el equipo de desarrollo. Se caracteriza por su adaptabilidad, pudiendo ser aplicada dinámicamente durante el ciclo de vida del software. Al estar sustentada en las relaciones interpersonales, establece una unión entre el cliente y el equipo de desarrollo, logrando satisfacer al cliente dándole el producto que necesita cuando lo necesita. Esta metodología propone un ciclo de vida divido en cuatro fases:

- **Exploración:** se plantean las historias de usuario (HU) importantes para la primera entrega del producto. Al unísono el equipo de desarrollo se familiariza con las herramientas, tecnologías y prácticas que se utilizarán en el proyecto. Esta fase concluye probando las tecnologías y construyendo un prototipo del sistema.
- **Planificación de la entrega:** se define el orden en que serán implementadas y entregadas las HU y se estiman los recursos necesarios y la cantidad de iteraciones. Como resultado de esta fase se obtiene el Plan de entregas. Además es desarrollado el Plan de iteraciones considerando las HU, la velocidad con que marcha el proyecto, los recursos disponibles y las prioridades del cliente.

- **Iteraciones:** esta fase incluye varias iteraciones sobre el sistema antes de ser entregado. El trabajo realizado en la misma, se expresa en las tareas de programación previamente asignadas a un miembro del equipo de desarrollo como responsable.
- **Producción:** son aplicadas pruebas unitarias y funcionales para determinar si el producto está en condiciones de ser trasladado al entorno cliente. El equipo de desarrollo prevé tareas de ajuste pero no se realizarán desarrollos funcionales.

Durante el ciclo de vida del proyecto, la metodología genera los siguientes artefactos:

- Historias de usuario.
- Plan de iteraciones.
- Metáfora.
- Tareas de ingeniería.
- Pruebas de aceptación.
- Pruebas unitarias y de integración.
- Código.

1.7.2 Justificación de las metodologías escogidas

Luego de realizar un estudio de las metodologías que fueron descritas anteriormente y teniendo en cuenta las particularidades de la solución que se desea desarrollar, se elige XP ya que el sistema al cual se va integrar la solución utiliza esta metodología como política del proyecto, siendo este aspecto importante a la hora de realizar una elección, ya que toda la documentación del sistema de reconocimiento facial está generada a través de las pautas que rigen el desarrollo utilizando XP. El empleo de otra metodología implicaría cambios en la forma de representar artefactos, o demás aspectos propios del desarrollo del software, siendo este proceso de cambio un contratiempo que puede resultar engorroso. Por otra parte XP es necesaria para seguir el ciclo de vida de la solución que se pretende desarrollar, ya que el usuario final es el propio CISED y los requerimientos de la solución pueden sufrir cambios, debido a que el centro puede necesitar nuevas funcionalidades durante el proceso de desarrollo de la solución. Además se cuenta con poco personal, por lo que no se pueden definir roles por etapas en el equipo de desarrollo así como generar múltiples artefactos.

Por otra parte se eligió como metodología para la extracción del conocimiento a Crisp-DM, a raíz de una encuesta realizada en el sitio KDnuggets en los años 2004 y 2007 con el objetivo de conocer la metodología de extracción del conocimiento más utilizada por la comunidad científica, evidenciándose que Crisp-DM en ambos años resultó la de mayor cantidad de votos con un 42% del total. Además el uso de la metodología Crisp-DM para la presente investigación tendrá un papel imprescindible, ya que orientará en cada una de las tareas y actividades a realizar como parte del

enfoque de Minería de Datos utilizado en la investigación. Guiará en los procesos de preparación de los datos, donde serán seleccionados, integrados y transformados; así como el modelado y evaluación de los resultados obtenidos para su uso.

1.7.3 Herramientas y Tecnologías

Lenguaje de modelado:

Un lenguaje para el modelado de objetos, es un conjunto estandarizado de símbolos y de modos de disponerlos para modelar un diseño de software orientado a objetos (42). En el proceso de desarrollo de la solución para la clasificación de rasgos biométricos faciales, se utilizará el Lenguaje Unificado de Modelado (UML[14]). Este lenguaje tiene una notación gráfica muy expresiva, que permite representar en mayor o menor medida todas las fases de un proyecto informático. UML provee de un mayor rigor en la especificación, permite realizar una verificación y validación del modelo realizado, además de automatizar determinados procesos, generar código a partir de los modelos y a la inversa (a partir del código fuente generar los modelos) (42). Permite documentar todos los artefactos de un proceso de desarrollo (requerimientos, arquitectura, pruebas, versiones). Una de sus principales características está relacionada a la posibilidad de poder conectarse con lenguajes de programación (ingeniería directa e inversa).

Herramientas de modelado:

- **Visual Paradigm para UML:** es una herramienta para el desarrollo de aplicaciones utilizando modelado UML. Es ideal para Ingenieros de Software, así como para Analistas y Arquitectos de sistemas que estén interesados en construir software a gran escala y necesiten confiabilidad y estabilidad en el desarrollo orientado a objetos. Brinda navegación intuitiva entre la escritura del código, su visualización y un ambiente visualmente superior de modelado.
- **Rational Rose:** proporciona un conjunto de prestaciones controladas por modelo para el desarrollo de aplicaciones de software. Permite a los clientes y a los diseñadores generar modelos UML de arquitecturas de software, necesidades empresariales, activos reutilizables y comunicación de nivel de gestión, además contiene un entorno de modelado visual que permite agilizar el desarrollo de aplicaciones y unifica el equipo del proyecto proporcionando una ejecución y una notación de modelos UML comunes.

[14] Por sus siglas en inglés: *Unified Modeling Language.*

Luego de estudiar estas dos herramientas se llega a la conclusión de que la más apropiada es **Visual Paradigm** en su versión **8.0**, ya que es un potente generador de informes en formato PDF[15]/HTML[16], permite realizar ingeniería inversa y directa, soporta varios usuarios trabajando sobre un mismo proyecto y permite agilidad en el trabajo del analista. Además se cuenta con experiencia trabajando en la misma.

Lenguajes de programación:

Lenguaje artificial que puede ser usado para controlar el comportamiento de una máquina, especialmente una computadora. Estos se componen de un conjunto de reglas sintácticas y semánticas que permiten expresar instrucciones que luego serán interpretadas (43). En el desarrollo de este trabajo fueron analizados los lenguajes de programación C#, C++ y Java, abordando sus principales características.

- **C#:** es un lenguaje de programación que se ha diseñado para generar diversas aplicaciones que se ejecutan en .NET Framework. Este lenguaje es simple, eficaz, con seguridad de tipos y orientado a objetos. Las numerosas innovaciones de C# permiten desarrollar aplicaciones rápidamente, mantener la expresividad y elegancia de los lenguajes de estilo de C. Aunque C# forma parte de la plataforma .NET es una interfaz de programación de aplicaciones (API), mientras que C# es un lenguaje de programación independiente diseñado para generar programas sobre dicha plataforma.
- **C++:** La intención de su creación fue extender al exitoso lenguaje de programación C con mecanismos que permitan la manipulación de objetos. Es un lenguaje sencillo que no está especializado en ningún tipo de aplicación, versátil, flexible, conciso y muy eficiente. Una particularidad del C++ es la posibilidad de redefinir los operadores y de poder crear nuevos tipos que se comporten como tipos fundamentales.
- **Java:** Está inspirado en C++ y se proyectó con la finalidad de obtener un producto de pequeñas dimensiones, simple y portátil sobre diferentes plataformas y sistemas operativos ya sea a nivel de código fuente como a nivel de código binario. Es un lenguaje multiplataforma y sencillo, además es independiente de la arquitectura del procesador. Su sintaxis ha sido trabajada mejorando la de C++ logrando mayor sencillez y legibilidad.

Los lenguajes de programación seleccionados para la implementación de la solución son C# y C++. El primero de ellos por su sencillez, además de que no es necesario emplear mucho tiempo en la programación ya que presenta bibliotecas de clases muy completas y bien diseñadas. Estas

[15] Por sus siglas en inglés: *Portable Document Format.*
[16] Lenguaje de programación web, sus siglas en inglés significan: *HyperText Markup Language.*

responden a las necesidades del cliente de una aplicación en cuanto a la velocidad en la manipulación de los datos así como una interfaz que sea lo más rápida posible y que permita una fácil navegación. El lenguaje C++ en la utilización de una biblioteca de clases[17] que contiene el algoritmo Active Shape Model (ASM), necesaria para la ubicación de los puntos característicos en el rostro. Por otra parte el Sistema de reconocimiento facial está desarrollado en C#, con el uso de las bibliotecas de clases mencionadas anteriormente, lo que su uso facilitaría la integración de la solución.

Entornos de Desarrollo Integrado (IDE[18]):

Un entorno de desarrollo integrado, es un programa informático compuesto por varias herramientas de programación, que puede dedicarse a uno o varios lenguajes de programación en conjunto. Consiste en un editor de código, un compilador, un depurador y un constructor de interfaz gráfica (GUI[19]).

Para el desarrollo de la solución que se desea proponer, se analizaron los entornos siguientes, teniendo en cuenta los lenguajes de programación antes seleccionados:

- **MonoDevelop:** es un entorno de desarrollo integrado libre y gratuito, diseñado primordialmente para C# y otros lenguajes .NET como Nemerle, Boo, Java (vía IKVM.NET) y en su versión 2.2 Python. El IDE incluye manejo de clases, ayuda incorporada, completamiento de código, Stetic (diseñador de GUI) integrado, soporte para proyectos y un depurador integrado desde la versión 2.2. Cuenta con una función de autocompletado, útil a la hora de escribir código; por otro lado no cuenta con funciones de depuración que hacen difícil en algunos casos encontrar un error específico.
- **Visual Studio.Net:** es un entorno de desarrollo integrado para sistemas operativos Windows. Soporta múltiples lenguajes de programación tales como C++, C#, Visual Basic .NET, entre otros. Visual Studio permite a los desarrolladores crear aplicaciones, sitios y aplicaciones web, así como servicios web en cualquier entorno que soporte la plataforma .NET (a partir de la versión .NET 2002). Cuenta con varias versiones, una de las analizadas fue Visual Studio 2012, esta última versión de la suite de desarrollo, integra todas las capacidades de la versión 2010 (rapidez en el desarrollo, segura, confiable y administrable), pero ha sufrido un completo cambio en el gráfico de la interfaz y se han sumado prestaciones tanto a los módulos existentes, como a los de nueva creación.

[17] Por sus siglas en inglés: *dynamic-link library (dll).*
[18] Por sus siglas en inglés: *Integrated Development Environment.*
[19] Por sus siglas en inglés: *Graphical User Interface.*

Luego del análisis de ambos entornos de desarrollo integrado, se decidió escoger el Visual Studio.NET en su versión 2012, debido a que brinda calidad, rapidez, seguridad y confiabilidad. Es un entorno para sistemas operativos Windows, que permite modelar la arquitectura y el diseño de la solución. Además brinda aprovechamiento de las últimas características de los lenguajes de programación, mejoras en rendimiento, disponibilidad y compatibilidad, productividad para el entorno, herramientas de prueba y aprovechamiento de herramientas y conceptos de agilidad.

Herramientas de MD:

- **Weka[20]:** contiene una colección de herramientas de visualización y algoritmos de aprendizaje automático para análisis de datos, permitiendo la experimentación con estos. Weka brinda la posibilidad de realizar manipulaciones sobre los datos aplicando filtros en dos niveles: atributos e instancias y realizar operaciones de filtrado en cascada, siendo la entrada de cada filtro la salida del filtro anterior. Entre otras funcionalidades permite realizar la asociación y el agrupamiento de forma simbólica o numérica.
- **RapidMiner:** es un programa informático para el análisis y la MD, que incluye cientos de métodos para integración, transformación, modelación, preprocesamiento y visualización de datos. Puede ser utilizada en varias plataformas y sistemas operativos. RapidMiner es fácilmente la interfaz gráfica de usuario más potente e intuitivo para el diseño de los procesos de análisis, por lo que es una de las herramientas analíticas y de Minería de Datos más conocidas y utilizadas para proyectos reales. Una de sus características principales es que permite representar de forma interna los procesos de análisis de datos en ficheros XML.

Luego de analizar las herramientas anteriores, se decidió utilizar durante el desarrollo de la solución, la herramienta **Weka** en su versión **3.7.10**, ya que es un entorno visual fácil de usar, que permite reconocer errores, aplicar soluciones rápidas y utilizar los algoritmos EM, K-means y K-NN. Además se cuenta con experiencia en el trabajo con la herramienta.

Herramienta para hojas de cálculo:

Una hoja de cálculo es un tipo de documento, que permite manipular datos numéricos y alfanuméricos dispuestos en forma de tablas compuestas por celdas. Existen diversas herramientas para hojas de cálculo difundidas en el mercado, entre estas se encuentra: Calc de OpenOffice.org, Calc integrada a LibreOffice, Microsoft Excel integrada a Microsoft Office, entre otras. De este conjunto se decidió utilizar **Microsoft Excel** en su versión del **2013** ya que:

- Provee un conjunto de plantillas para presupuestos, calendarios, formularios, informes. etc.

[20] Por sus siglas en inglés: *Waikato Environment for Knowledge Analysis.*

- ❖ Contiene una herramienta para el análisis rápido, que permite convertir los datos en un gráfico o tabla.
- ❖ Permite lograr resultados profesionales con rapidez.
- ❖ Con la herramienta Recomendaciones de gráficos, puede seleccionarse el gráfico que muestra mejor las ideas que se desean presentar.
- ❖ Se cuenta con experiencia trabajando en la misma.

Gestores de base de datos:

Los gestores de base de datos son un conjunto de programas que permiten el almacenamiento, modificación y extracción de la información, además de proporcionar herramientas para añadir, borrar, modificar y analizar los datos. Entre los gestores de base de datos más utilizados a nivel mundial se encuentran Oracle, Microsoft SQL Server 2008, **MySQL** y **PostgreSQL**.

- ❖ MySQL: es un sistema de gestión de base de datos relacional capaz de almacenar una enorme cantidad de datos de gran variedad. Este gestor utiliza el lenguaje de consulta estructurado (SQL), dispone de almacenamiento de procedimientos, disparadores[21] y vistas, además de ser GPL[22] por lo que no tiene costo.
- ❖ PostgreSQL: es un gestor de base de datos objeto-relacional que utiliza un modelo cliente-servidor. Brinda integridad referencial, además se pueden realizar varias acciones al unísono sobre una misma tabla. Algunas de las características que lo distinguen son:
    - Licencia BSD[23] que permite su uso en software privativo.
    - Presenta un alto nivel de seguridad de los datos al gestionar usuarios y contraseñas.
    - Está disponible para software libre y privativo por lo que lo convierte en multiplataforma.

Luego de estudiar los gestores anteriores, se decidió escoger **PostgreSQL** en su versión **9.1**, ya que se cuenta con experiencia trabajando en este gestor, siendo fácil la creación de la base de datos de la solución que se desea desarrollar, así como las consultas a la misma.

Mapeo objeto-relacional (ORM[24]):

Un ORM es un modelo de programación que consiste en la transformación de las tablas de una base de datos, en una serie de entidades que simplifiquen las tareas básicas de acceso a los datos para el programador (44).

[21] Del inglés: triggers.
[22] Por sus siglas en inglés: *General Public Licence.*
[23] Por sus siglas en inglés: *Berkeley Software Distribution*.
[24] Por sus siglas en inglés: *Object Relational Mapping.*

Telerik OpenAccess, es una herramienta de mapeo diseñada para generar códigos de acceso a datos. Esta herramienta funciona para todas las plataformas .NET, ofrece soporte integrado para más de 12 bases de datos, incluyendo Oracle, MySQL y PostgreSQL, y proporciona una perfecta integración con Visual Studio permitiendo crear bases de datos independientes del código. Además utiliza poco espacio de memoria en el cliente, y genera modelos con el potente diseñador visual para OpenAccess ORM (45). Por otra parte el equipo de desarrollo posee experiencia en el trabajo con la herramienta.

Plataforma de desarrollo:

Una plataforma de desarrollo es el entorno de software común en el cual se desenvuelve la programación de un grupo definido de aplicaciones. Durante el desarrollo de la presente investigación se hará uso del **Framework.net** en su versión **4.0** como plataforma de desarrollo. Brinda mejoras en seguridad, programación paralela, rendimiento y diagnóstico, e incluye simplificación y transparencia en las implementaciones. Otras de las características que lo distingue de versiones anteriores son:

- Incluye un nuevo Runtime para ejecución de código dinámico, que simplifica y facilita el desarrollo de código dinámico en .NET.
- Proporciona recolección de elementos no utilizados en segundo plano.
- Multilenguaje.

Bibliotecas de clases utilizadas:

En décadas pasadas una biblioteca era un conjunto de programas que contenían cientos de rutinas (una rutina es un procedimiento o función bien verificados, en determinado lenguaje de programación), hoy se conocen como un conjunto de clases de programación orientada a objetos, las cuales contienen métodos que son útiles para los programadores. También, existen diversas bibliotecas de clases desarrolladas por terceros que están disponibles a través de la web. En la presente investigación se trabajará con aquellas que permitan la detección del rostro en una imagen digital. Entre las bibliotecas que permiten esta funcionalidad se encuentran **OpenCV** y **Emgu CV**. La primera de ellas es una biblioteca libre, de visión artificial originalmente desarrollada por Intel, es multiplataforma, existiendo versiones para GNU/Linux, Mac OS X y Windows; contiene más de 500 funciones que abarcan una gran gama de áreas en el proceso de visión, como reconocimiento de objetos (reconocimiento facial), calibración de cámaras, entre otras. Presenta la limitación de sólo ser compatible con los lenguajes de programación C++, C y Python, debido a esta limitación es necesario utilizar un contenedor para poder disponer de OpenCV en C#, por lo que se decidió utilizar **Emgu CV**. Esta es una plataforma cruzada, envoltorio de la biblioteca OpenCV para procesamiento

de imágenes, siendo posible poder utilizar OpenCV en plataforma .Net. Fue seleccionada para su uso, ya que es multiplataforma, libre y dispone el algoritmo AdaBoost para detección de rostro, necesario para desarrollar esta fase en el proceso de reconocimiento facial de la solución a implementar.

Active Shape Mode (ASM) es un modelo definido por una serie de puntos, así como la conexión entre ellos. El modelo ASM descubre las principales variaciones en los datos de entrenamiento utilizando PCA, que permite al modelo reconocer automáticamente si el contorno es un posible buen contorno del objeto. Es utilizada para corregir factores que pueden influir negativamente sobre las potencialidades biométricas de los rostros humanos, como son: la pose, la iluminación, la expresión, los accesorios utilizados, entre otros.

1.8 Conclusiones parciales

El análisis de las características y funcionamiento de algunos de los algoritmos utilizados para la reducción de la dimensionalidad, así como para el agrupamiento y la clasificación en la MD, evidenciaron que con la aplicación de estos al sistema de reconocimiento facial actualmente utilizado en el CISED, podrá alcanzarse el objetivo propuesto en la presente investigación. La selección de las metodologías XP y Crisp-DM, conjuntamente con los lenguajes de programación C# y C++, el IDE Visual Studio.Net en su versión 2012, la herramienta CASE[25] Visual Paradigm 8.0 para modelar UML, el gestor de base de datos PostgreSQL 9.1 y Weka 3.7.10 como herramienta de MD, demostró que con las potencialidades de estas, se podrá obtener una solución que supla las necesidades existentes en el centro.

[25] Por sus siglas en inglés: *Computer Aided Software Engineering.*

2 Capítulo 2: Análisis y Diseño

2.1 Introducción

En el capítulo se describe la propuesta de solución mediante la ejecución de las tareas de MD: agrupamiento y clasificación. Se desglosan las fases de la metodología Crisp-DM, mediante las tareas realizadas en cada una de estas. Además se desarrollan las fases iniciales de la metodología de desarrollo XP: planificación y diseño, presentándose los diferentes artefactos que genera la misma.

2.2 Propuesta de solución

La propuesta de solución como se explicaba en el capítulo 1 será una variante del PCA+K-means denominada PFA+K-means, viéndose como un problema de Minería de Datos a resolver a través de la utilización de la metodología Crisp-DM para dirigir el proceso de solución. Básicamente lo que se pretende utilizar es la propuesta inicial PCA+K-means y desarrollarla a través de un conjunto de procesos utilizados en Minería de Datos.

La solución está formada por dos fases: una primera fase en donde se agruparán las imágenes faciales en K grupos y una segunda que se encargará de clasificar una nueva imagen facial en alguno de los grupos obtenidos en la fase anterior, integrándose el clasificador como una biblioteca de clases al sistema de reconocimiento facial.

Se puede afirmar entonces que la propuesta incluye principalmente el desarrollo de dos tareas de Minería de Datos: agrupamiento y clasificación, las cuales serán explicadas en detalles a continuación.

Agrupamiento

El primer paso a realizar en esta fase es la extracción de los vectores característicos de las imágenes faciales pertenecientes a la población UCI (10788 imágenes), por medio de cinco procedimientos: Imagen Reescalada (IR), Imagen Icono (IC), Transformada Discreta del Coseno (DCT[26]), Transformada de Wavelet Gabor (GWT[27]) y el espacio de color HSV[28]. Sea $v \in R^{Cxc}$ la forma del vector característico asociado a los procedimientos DCT, GWT y HSV, donde C representa la cantidad de características y c los valores asociados a cada característica, se necesita transformar este vector a la forma $v' \in R^{C}$. Con el objetivo de transformar v a v' se procede de la siguiente forma: en DCT se elige el primer coeficiente de su recorrido en zigzag que es el valor más significativo de entre todos sus valores, en cambio en GWT y HSV se calcula la media de sus valores. El paso anterior fue necesario para poder aplicar la siguiente etapa de la propuesta, que consta de la selección o extracción de las principales características de estos vectores a través del método de reducción de la

[26] Por sus siglas en inglés: *Discrete Cosine Transform.*
[27] Por sus siglas en inglés: *Gabor Wavelet Transform.*
[28] Del inglés: *Hue, Saturation, Value.*

dimensionalidad PFA. En el siguiente paso se generaron dos ficheros de tipo ARFF[29] de las imágenes faciales, uno conteniendo todas las características y el otro sólo las principales, con el objetivo de utilizar la herramienta de Minería de Datos Weka en la solución. Una vez generados estos ficheros se procede a estimar el número de grupos presentes en los datos mediante una combinación del algoritmo de agrupamiento EM y el método de validación cruzada, presente en la colección de algoritmos de Weka, esto se realizó utilizando el fichero ARFF de todas las características. Los restantes procesos utilizaron el fichero ARFF que contiene únicamente las principales características de las imágenes faciales. Después de estimado el posible número K de grupos presentes en los datos, se define un intervalo $[K-5, K+5]$ en el cual se ejecutó el K-means 10 veces variando sus puntos iniciales, utilizando la propuesta K-means++ contenida en Weka, con lo que se obtuvo un conjunto de modelos de agrupamiento que fueron validados mediante dos métodos, con el objetivo de seleccionar el número de grupos que mejor se ajusta a los datos y propicie la solución de agrupamiento más adecuada[30]. El primer método consiste en medir la calidad del agrupamiento mediante un conjunto de índices de validación (CVIs[31]) que brindan información sobre la alta cohesión y separabilidad presente en la solución y un segundo método basado en un esquema de votos. Una vez determinado el K que mejor se ajusta a los datos, se procede a elegir el modelo dentro de las corridas del K-means, para este K, que mejor índice de agrupamiento presente. Después de obtenido el mejor modelo para cada uno de los distintos procedimientos (IR, IC, DCT, GWT, HSV) se analiza si estos modelos cumplen con los requisitos del negocio, que son totalidad, exclusividad, homogeneidad y adecuada distribución de las imágenes faciales en los grupos detectados.

La figura 2.1 muestra el flujo de procesos para realizar el agrupamiento:

[29] Por sus siglas en inglés: *Attribute-Relation File Format.*

[30] Nótese que se le confiere gran importancia a la selección del K y a las variaciones de los puntos iniciales, ya que es bien conocido que el K-means depende en gran medida de la elección de los centroides iniciales, como parte de su naturaleza de algoritmo escalador de colina, y necesita la información del número de grupos a priori siendo vital para determinar una adecuada solución de agrupamiento.

[31] Por sus siglas en inglés: *Clustering Validity Indices.*

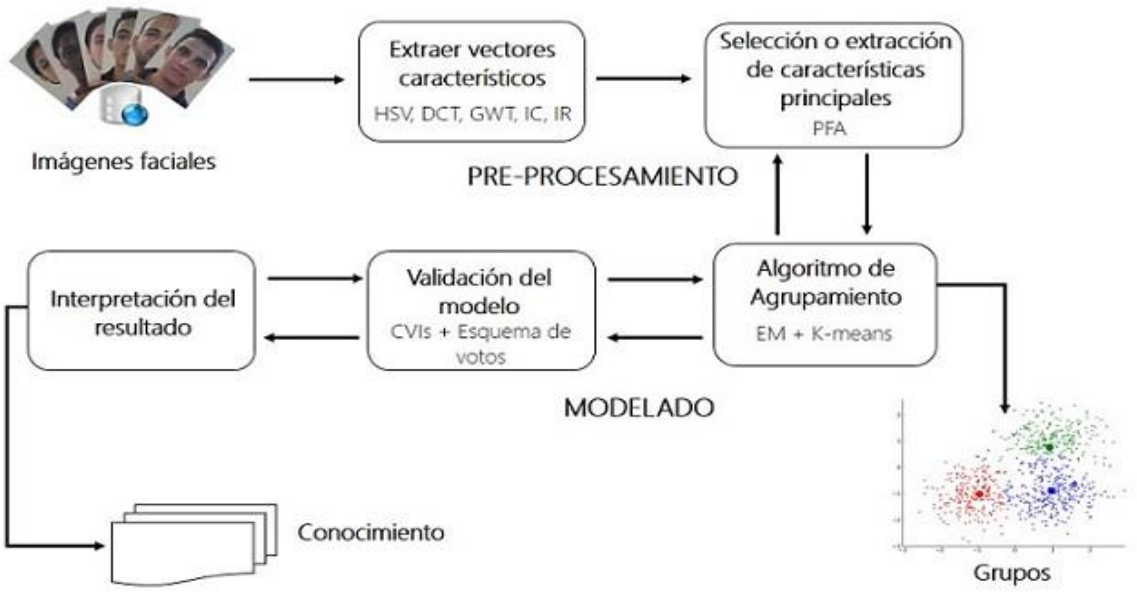


Figura 2.1: Flujo de procesos para el agrupamiento.

Clasificación

Una vez obtenido el modelo de agrupamiento para cada de uno de los procedimientos aplicados en la investigación, se procede a la fase de clasificación utilizando está información, o sea las asignaciones de las distintas imágenes faciales a los grupos o clases. Para esta fase se propone etiquetar la base de datos de imágenes faciales con el objetivo de conocer en cada momento el grupo al que pertenece una imagen facial determinada, además de poder realizar visualizaciones de los grupos para estudios más detallados de las características similares en rostros pertenecientes a la población UCI. En la siguiente etapa se implementa un clasificador que en la solución propuesta es el K-NN, específicamente 1-NN, donde los datos de interés para entrenar el mismo son los centroides obtenidos en la fase anterior. Con el objetivo de proponer un único procedimiento para extraer los vectores característicos, se le harán validaciones al clasificador para cada uno de los procedimientos utilizando el índice de precisión, que brinda una medida de qué tan exacto es el mismo. Después de tener el clasificador y el procedimiento a utilizar, se seguirán los siguientes pasos: dada una nueva imagen facial, se extraerá el vector característico asociado y las principales características de este vector, por último se aplicará el 1-NN para determinar a qué grupo pertenece.

En resumen, el flujo que sigue esta fase se ilustra a continuación:

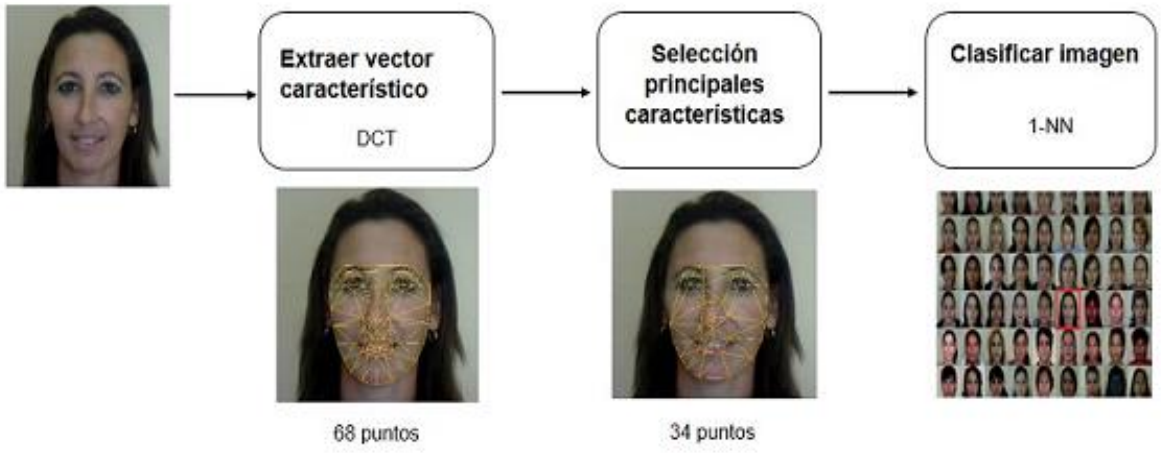


Figura 2.2: Propuesta de solución para la clasificación.

2.3 Modelo de dominio

El modelo de dominio es la representación gráfica de los conceptos claves del dominio del problema. Ayuda a un mejor entendimiento de los objetos relacionados con la solución, así como las relaciones entre estos y a delimitar el alcance del problema. Además permite describir las clases más importantes dentro del contexto de la solución. La representación de los conceptos fundamentales y las relaciones entre ellos se muestran en la figura 2.3.

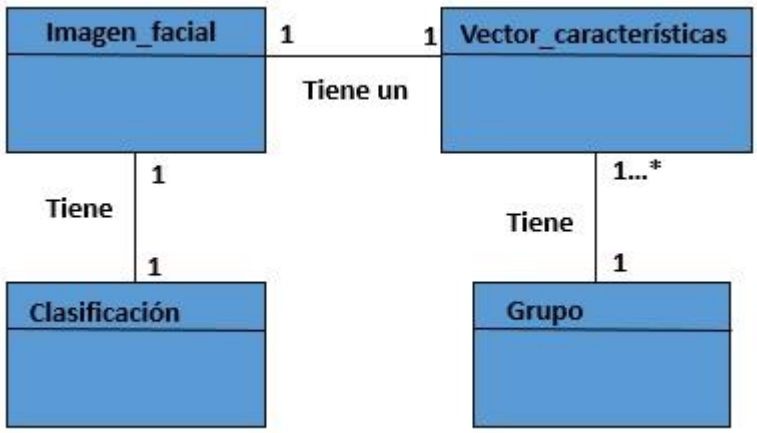


Figura 2.3: Modelo de dominio.

2.4 Aplicando Crisp-DM

2.4.1 Fase de Comprensión del negocio

Los objetivos del negocio, como se explicó en el diseño teórico de la investigación, están relacionados con la reducción del espacio de búsqueda para mejorar el proceso de reconocimiento facial. Después de analizados los objetivos, las tareas de Minería de Datos que se determinaron para resolver el problema fueron: agrupamiento y clasificación.

2.4.2 Fase de Comprensión de los datos

En esta fase de la metodología se recolectan los datos iniciales y se describen, con el objetivo de examinar propiedades presentes en los mismos. En la investigación se utilizó una base de datos de imágenes faciales pertenecientes a la UCI que presenta las siguientes propiedades: la población UCI está conformada por 10788 individuos de los cuales el 41% representan al sexo femenino y el 59% al masculino, la población la componen estudiantes con edades entre los 18 y 24 años (representando un 38% de la población total), trabajadores con un 61% entre los 24 y 60 años de edad y familiares de diversas edades representando sólo el 1%. Una vez conocidas las propiedades de los datos se procede a una normalización (9) de las imágenes con el objetivo de prepararlas para la extracción de los vectores característicos mediante cinco procedimientos:

- **Imagen Reescalada:** propone llevar la imagen a una escala de 100x100, conformándose un vector característico n dimensional con los valores de los píxeles de la región de interés (rostro).

Figura 2.4: Procedimiento Imagen Reescalada.

- **Imagen Icono:** como primer paso se reescala la imagen a 96x96, después se divide la imagen en secciones de 8x8, eligiéndose de cada sección el pixel más representativo (media), generando así una nueva imagen de 12x12 (Imagen Icono). Al igual que en el procedimiento anterior el vector característico n dimensional lo componen los valores de los píxeles de la región del rostro.

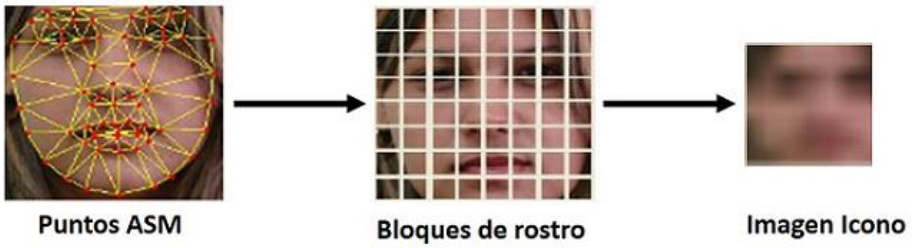


Figura 2.5: Procedimiento Imagen Icono.

- **HSV:** Generalmente una imagen viene dada en el formato RGB, pero para esta variante es necesario su conversión a HSV (matiz, saturación y valor). Luego de que la imagen es convertida se realiza el mismo proceso que en la Imagen Reescalada. A diferencia de la Imagen Icono y la Imagen Reescalada, el HSV genera un vector de la forma $v \in R^{Cxc}$.

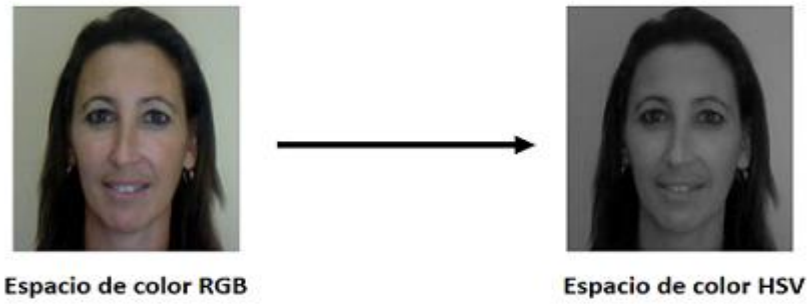


Figura 2.6: Procedimiento HSV.

- **DCT** (Transformada Discreta del Coseno): Tal como se describe en (9), la DCT tiene una buena capacidad de compactación de la energía al dominio transformado, es decir, que la DCT concentra la mayor parte de la información en pocos coeficientes transformados. Una vez que se hayan seleccionado los 68 puntos extraídos del rostro generados por ASM, se obtiene para cada uno de ellos una vecindad de 7x7 píxeles de la cual se extraen las características correspondientes. Estas características son extraídas a partir del filtro DCT que utiliza la fórmula que se muestra en la ecuación 2.1, es decir, para cada punto se va a obtener una matriz de transformación a partir de la vecindad antes definida (7x7 píxeles dando como resultado 49 coeficientes) donde la esquina superior izquierda recoge los valores más significativos (color rojo). Luego de ser distribuidos estos valores se realiza un recorrido en zigzag a la matriz dando como respuesta el vector característico mostrado en la figura 2.7.

$$B_{pq} = \propto_p \propto_q \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_{mn} \cos \frac{\pi(2m+1)p}{2M} \cos \frac{\pi(2n+1)q}{2N} \qquad \begin{matrix} 0 \le p \le M-1 \\ 0 \le q \le N-1 \end{matrix}$$

$$\propto_p = \begin{cases} \frac{1}{\sqrt{M}}, & p = 0 \\ \sqrt{\frac{2}{M}}, & 1 \leq p \leq M-1 \end{cases} \qquad \propto_p = \begin{cases} \frac{1}{\sqrt{N}}, & q = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq q \leq N-1 \end{cases}$$

Ecuación 2.1: Fórmula para calcular la DCT en una matriz 2D.

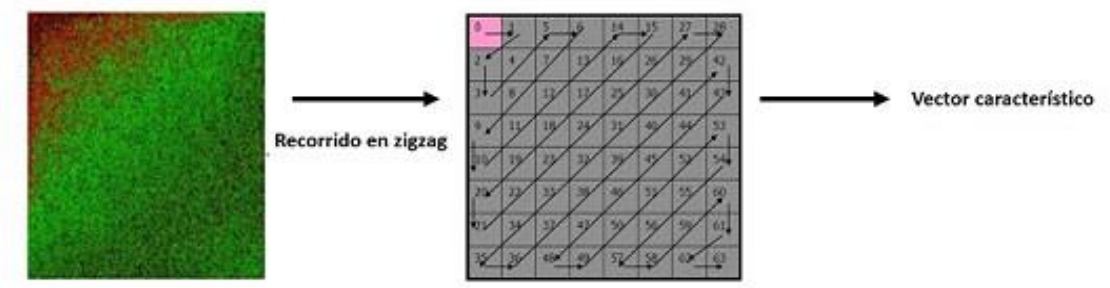


Figura 2.7: Vecindad extraída para cada uno de los puntos antes localizados, a la cual se le aplica el recorrido en zigzag dando como salida el vector característico correspondiente.

- **GWT** (Transformada de Gabor Wavelet): es una técnica de filtrado en dos dimensiones utilizada en el procesamiento de imágenes bidimensionales, principalmente porque permite resaltar los bordes de la imagen aunque persistan cambios de iluminación. Se aplican 40 filtros de wavelet, un filtro por cada escala en todas las direcciones posibles (0…4 frecuencia y 0…7 orientaciones). Se generó el mapa de características donde por cada uno de los 68 puntos definidos por ASM, se analiza el valor de intensidad según los filtros de Gabor.

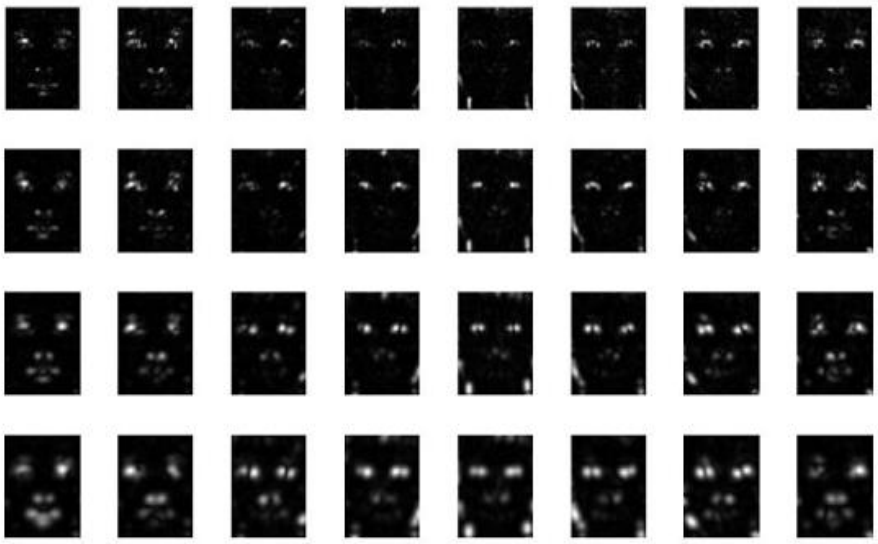


Figura 2.8: Filtros de Gabor Wavelet.

Después de obtenidos los vectores característicos asociados a cada imagen facial, se procede a realizar una exploración de los datos, para dirigir las tareas de MD y refinar la descripción inicial de los mismos. En esta fase se realizó una estimación inicial del número de grupos con el objetivo de descubrir estructuras de agrupamiento y proporcionar un punto de partida en el desarrollo de esta tarea. En la estimación se utilizó el método de validación cruzada combinado con la técnica de agrupamiento EM (23) presente en la herramienta Weka. Como última tarea la metodología propone evaluar la calidad de los datos, durante la tarea se comprobó que no existieran imágenes faciales que no presentaran rostros de individuos de la UCI y que cada imagen facial contara con un vector característico de acorde a los procedimientos utilizados.

2.4.3 Fase de Preparación de los datos

En esta fase de la metodología se aplicó el análisis de características principales, con el objetivo de realizar una selección de los atributos más relevantes como parte de la tarea de selección de los datos. La siguiente tarea consiste en la limpieza de los datos, constatándose en la investigación que cinco imágenes faciales no formaron parte del análisis ya que la biblioteca de clase ASM no detectó sus puntos característicos. Por último en la tarea de formateo de los datos se generaron los ficheros ARFF para ser utilizados en la herramienta Weka en los procesos de modelado. Esta fase es de mucha importancia ya que el éxito de los procesos que se aplicarán a continuación, depende considerablemente de la calidad con que hayan sido preparados los datos.

2.4.4 Fase de Modelado

Esta fase es la encargada de seleccionar las técnicas de modelado que mejor se ajustan a los datos obtenidos en la fase anterior. Luego se construye y evalúa el modelo.

Selección de las técnicas de modelado

- **Agrupamiento:** se procede a la obtención de varios modelos, mediante la corrida del algoritmo K-means, variando el número de grupos obtenidos en la estimación inicial.
- **Clasificación:** es aplicado el algoritmo 1-NN durante el proceso de clasificación de las imágenes faciales.

Generación de la prueba de diseño

Como parte de la generación de la prueba de diseño se utilizó el siguiente procedimiento:

- **Agrupamiento:** se realizaron diez iteraciones o corridas del algoritmo K-means, para un conjunto de K definidos en una vecindad del número de grupos estimados en la fase de la comprensión de los datos, específicamente se utilizaron once K distintos. Como parte de la

naturaleza del K-means su solución depende de la elección de los centroides iniciales, por lo que en cada corrida se variaron los puntos iniciales escogidos a través de la propuesta K-means++ (46), presente en la colección de algoritmos de Weka.

- **Clasificación:** se utilizó el algoritmo 1-NN, con los centroides como datos de entrenamiento y un conjunto de imágenes faciales con variaciones de brillo, rotación y escala como datos de prueba utilizando la herramienta Weka.

Evaluación del modelo

- **Agrupamiento:** teniendo en cuenta que la validación de los resultados obtenidos por los algoritmos de agrupamiento es una parte fundamental del proceso, además de que una de las cuestiones que más debilita a estas técnicas es la elección del número correcto de grupos, son utilizados los índices de validación de agrupamiento[32] (CVIs[33]) y un esquema basado en votos. La idea principal es combinar un conjunto de métodos para obtener una conclusión sólida y acertada. Para esto el primer paso es aplicar un método heurístico para estimar un número de grupos iniciales (EM + validación cruzada), una vez determinado un valor, se utilizarán los CVIs (47) y un esquema basado en votos (48) para obtener el número de grupos más adecuado presente en los datos. A continuación se define la notación (47) utilizada por los índices aplicados en la investigación y después se describen cada uno de los mismos.

  Notación

  Sea X el conjunto de N objetos representados como vectores en un espacio $F - dimensional$ $X = \{x_1, x_2, \dots, x_N\} \subseteq R^F$. Un agrupamiento en X es un conjunto de grupos disjuntos que dividen a X en K grupos $C = \{c_1, c_2, \dots, c_K\}$ donde $\cup_{c_k \in C} c_k = X, c_k \cap c_l = \emptyset, \forall k \neq l$. El centroide de un grupo c_k es su vector medio, $\overline{c_k} = \frac{1}{|c_k|}\sum_{x_i \in c_k} x_i$ y similarmente se define el centroide del conjunto de datos como la media de todos los vectores $\bar{X} = \frac{1}{N}\sum_{x_i \in X} x_i$. La distancia euclideana entre dos objetos se define como $d_e(x_i, x_j)$. Cada índice irá acompañado de una flecha arriba (↑) o abajo (↓) donde la flecha de abajo indica que un valor pequeño del índice significa una mejor partición, en cambio una flecha hacia arriba indica todo lo contrario.

[32] Un índice de validación del agrupamiento es una medida derivada de una solución de agrupamiento obtenida, que cuantifica propiedades de la solución tales como compactación intragrupos y separación intergrupos. Cada índice captura un aspecto específico del modelo de agrupamiento que sugiere que tan adecuada es la solución obtenida (49).

[33] Por sus siglas en inglés: *Cluster Validity Indices.*

Índices utilizados:

1. Davies–Bouldin (DB↓): es uno de los índices de validación más utilizado. Estima la cohesión basándose en la distancia entre los puntos de un grupo y su centroide, y la separación por medio de la distancia entre centroides. Se define como:

$$DB(C) = \frac{1}{K}\sum_{c_k \in C} max_{c_l \in C \setminus c_k}\left\{\frac{S(c_k) + S(c_l)}{d_e(\bar{c_k}, \bar{c_l})}\right\}$$

$$S(c_k) = \frac{1}{|c_k|}\sum_{x_i \in c_k} d_e(x_i, \bar{c_k})$$

Ecuación 2.2: Fórmula para calcular el índice Davies–Bouldin.

2. Calinski–Harabasz (CH↑): se trata de un índice de proporción donde se estima la cohesión por medio de las distancias de los puntos en un grupo a su centroide y la separación por la distancia de los centroides al centroide de todo el conjunto de datos. Se define como:

$$CH(C) = \frac{N-K}{K-1}\frac{\sum_{c_k \in C}|c_k| d_e(\bar{c_k}, \bar{X})}{\sum_{c_k \in C}\sum_{x_i \in c_k} d_e(x_i, \bar{c_k})}$$

Ecuación 2.3: Fórmula para calcular el índice Calinski–Harabasz.

3. Dunn (D↑): al igual que el índice CH es un índice de proporción. Propone la variante de estimar la cohesión mediante la distancia al vecino más cercano y la separación por el grupo de mayor diámetro. Este índice está definido por la siguiente ecuación:

$$D(C) = \frac{min_{c_k \in C}\left\{min_{c_l \in C \setminus c_k}\{\partial(c_k, c_l)\}\right\}}{max_{c_k \in C}\{\Delta(c_k)\}}$$

$$\partial(c_k, c_l) = min_{x_i \in c_k} min_{x_j \in c_l}\{d_e(x_i, x_j)\}$$

$$\Delta(c_k) = max_{x_i, x_j \in c_k}\{d_e(x_i, x_j)\}$$

Ecuación 2.4: Fórmula para calcular el índice Dunn.

Este índice presenta muchas variantes para calcular ∂ y Δ, en la investigación se utilizaron las siguientes:

$$\partial(c_k, c_l) = d_e(\bar{c_k}, \bar{c_l})$$

$$\Delta(c_k) = \frac{2}{|c_k|}\sum_{x_i \in c_k} d_e(x_i, \bar{c_k})$$

Ecuación 2.5: Variante para calcular el índice Dunn utilizada en la investigación.

4. DF-A: en (49) se propone la variante de combinar varios índices con el objetivo de hacer más confiable la validación de las soluciones de agrupamiento. Sean un

conjunto de modelos definidos para un número específico de grupos (k) para los cuales se le han calculado un conjunto de índices de validación $\{s_1, s_2, \dots, s_N\}$, entonces se define $d_i = \max(s_i)$, luego queda definido el índice DF-A como:

$$DF - A = \frac{1}{N}\sum_{i=1}^{N} d_i$$

Ecuación 2.6: Fórmula para calcular el índice DF-A.

Además de los CVIs, se utilizó el esquema basado en votos para corroborar los resultados emitidos por los índices empleados. Se fundamenta en la generación de una cierta cantidad de modelos de agrupamiento para cada K y en la recogida de votos a partir de los mejores índices.

- **Procedimiento para utilizar los CVIs**

Por cada modelo generado a partir del K-means se procede a calcular los índices DB, CH y D, después se normalizan estos valores entre 0 y 1, para que los índices con mayores valores no tengan mayor peso al ser combinados. Luego de normalizados los valores se procede a calcular el índice DF-A, el cual determinará el número de grupos más adecuado según el que presente mayor valor. Posteriormente se procede a realizar una regresión lineal (5) con los valores de K y sus respectivos DF-A para comprobar un conjunto de estadísticas enfocadas a que los resultados obtenidos no fueron por azar, concluyéndose que la elección realizada por este método es confiable.

- **Procedimiento para utilizar el esquema de votos**

Para cada uno de los modelos generados con K-means, fueron calculados los índices CH, DB y D. Luego se determinan los mejores índices por cada corrida, asignándose un voto al K al que pertenecen. Finalizado el paso anterior quedarán asignados a cada número de grupos un número de votos, eligiéndose el K que mayor cantidad de votos presente.

❖ **Clasificación:** para la evaluación del modelo de clasificación fue aplicado un índice de precisión, este índice mide la cantidad de veces que una imagen facial, con variaciones en la iluminación, escala y pose perteneciente a una persona determinada, se clasifica dentro del grupo al que pertenece la imagen de la misma persona que formó parte del análisis de agrupamiento. Esta relación se puede cuantificar de la siguiente manera:

$$índice = \frac{CC}{TC}$$

donde CC representa la cantidad de imágenes, con distintas variaciones, que se clasificaron en el grupo que contiene la imagen facial de la persona a la que representan y TC el total de imágenes a clasificar.

2.4.5 Fase de Evaluación

En esta fase se realiza la evaluación de los resultados respecto a los objetivos definidos en el negocio y las deficiencias del modelo. Además se revisa el proceso de acuerdo a elementos que puedan ser mejorados.

Evaluar los resultados

En la investigación se observó que los modelos de agrupamiento generados por los procedimientos Imagen Reescalada y HSV, arrojaron una distribución de imágenes faciales en los grupos inadecuada ya que presentaban grupos con bajos porcientos con respecto al total de imágenes y los índices de validación eran muy bajos en correspondencia con los demás procedimientos.

Los parámetros utilizados para evaluar los modelos de agrupamiento con respecto al negocio fueron:

1. Totalidad: todos los individuos deberán pertenecer a un grupo.
2. Exclusividad: relacionada con la pertenencia de un individuo a un único grupo.
3. Homogeneidad: se refiere a que los vectores de características pertenecientes a un grupo específico, deben ser lo más compactos y semejantes entre sí.

Evidenciándose estas propiedades en los modelos obtenidos con los procedimientos IC, DCT y GWT.

2.4.6 Fase de Desarrollo

Esta fase de la metodología se encarga de la implementación de las conclusiones obtenidas durante el proceso de MD en el negocio.

Desarrollo del plan

Una vez determinado el procedimiento que mejor se ajusta a los datos, a partir del que mejor índice de precisión presentó, se propone etiquetar la base de datos del sistema de reconocimiento facial para saber en cualquier momento a que grupo o clase pertenece una imagen determinada. Después se implementa el clasificador a través de una dll para integrarse al sistema y dada una nueva imagen facial clasificarla y reducir el espacio de búsqueda en el proceso de reconocimiento. Teniendo en cuenta que el clasificador tiene un índice de precisión, se propone adaptar el sistema a que realice una petición a la dll si no es encontrada la imagen facial dentro del grupo que se estimó, con el

objetivo de que el clasificador responda con el próximo grupo o clase más probable. Esto permite mayor exactitud y mantiene la sectorización del espacio muestral.

2.5 Aplicando XP

2.5.1 Fase de Planificación

En esta fase se pone en práctica la comunicación con el cliente, como uno de los valores definidos por esta metodología, al elaborar la planificación por etapas, definir las historias de usuario y el orden en que serán implementadas y entregadas de acuerdo con la velocidad del proyecto.

Actores del sistema

Generalmente los actores del sistema, son todas aquellas personas que de una forma u otra estarán interactuando con el mismo y obtendrán algún resultado de los procesos que se ejecutan. En el caso de la solución, los actores del sistema identificados no son personas sino otros sistemas, estos se muestran en la tabla 2.1.

Actor del sistema	Descripción
Sistema Multibiométrico	Es el encargado de enrolar, identificar y verificar a una persona a través del rostro o la huella dactilar. Realiza la identificación de forma distribuida, por lo que puede particionar la base de datos y asignar una partición a cada nodo de clúster de procesamiento.
Sistema de reconocimiento facial	Encargado de realizar los procesos de enrolamiento e identificación de una persona por medio de sus características faciales.

Tabla 2.1: Actores del sistema.

Captura de requerimientos

Capturar los requerimientos de la solución es una tarea fundamental en esta fase, ya que un error al identificarlos puede traer como consecuencia, que la solución final no cumpla con las expectativas del cliente y no sea funcional para el ambiente donde será desplegada. Los requerimientos de la solución pueden ser funcionales y no funcionales. Los requerimientos funcionales son especificaciones de los servicios y funciones que debe proporcionar la solución, de la manera en que debe reaccionar a entradas particulares y la forma de comportarse ante situaciones específicas (50). Mientras que los requerimientos no funcionales conforman restricciones de los servicios o funciones que proporciona la solución, incluye restricciones de tiempo, sobre el proceso de

desarrollo, estándares, entre otras (50). Los requerimientos identificados para el desarrollo de la solución son los siguientes:

Requerimientos funcionales (RF)

RF 1. Cargar conjunto de datos de entrenamiento.

RF 2. Procesar imagen facial.

RF 2.1. Detectar rostros.

RF 2.2. Normalizar imagen facial.

RF 2.3. Extraer vector de características.

RF 3. Generar fichero ARFF.

RF 4. Etiquetar conjunto de datos de entrenamiento.

RF 5. Visualizar agrupamiento.

RF 6. Cargar imagen facial.

RF 7. Clasificar imagen facial.

Requerimientos no funcionales (RNF)

Requerimientos del producto

1. Requerimientos de portabilidad

RNF 1.1. La solución será usada bajo sistemas operativos Windows XP o superior.

RNF 1.2. Framework de desarrollo: .NET Framework 4.0.

RNF 1.3 Microsoft Visual C++ 2010 x86 Redistribuible.

RNF 1.4 Microsoft Visual C++ 2010 x86 Runtime.

RNF 1.5 Gestor de base de datos: PostgreSQL 9.1.

RNF 1.6 PC Intel Pentium 4 o superior.

RNF 1.7 CPU 2.13 GHZ o superior.

RNF 1.8 2 GB de memoria RAM o superior.

2. Requerimientos de usabilidad

RNF 2.1 La solución debe ser de fácil manejo para los usuarios que tengan niveles básicos sobre computación y Minería de Datos.

Requerimientos organizacionales

3. Requerimientos de implementación

RNF 3.1 Lenguaje de programación: Visual C++ y C#.

RNF 3.2 IDE: Visual Studio 2012.

RNF 3.3 Para el Modelado de UML se utiliza: Visual Paradigm 8.0.

RNF 3.4 Bibliotecas de clases: Active Shape Model (ASM), Emgu CV.

Historias de usuario

Las historias de usuario (en lo adelante HU) son descripciones cortas de una necesidad del cliente del software, su utilización es común cuando se aplican marcos de trabajo ágiles, tales como Scrum o Extreme Programming (XP). A continuación se muestra la HU correspondiente al RF 1, el resto de las HU se muestra en los anexos. (Ver Anexo 4)

Historia de Usuario	
Número: HU_1	**Nombre de Historia:** Cargar conjunto de datos de entrenamiento.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial.	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 1	**Iteración asignada:** 1
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe cargar un conjunto de imágenes faciales de entrenamiento desde un directorio físico y almacenarlas en la base de datos.	
Observaciones: En caso de no poder cargar el conjunto de datos no podrán ejecutarse los demás procesos.	

Tabla 2.2: HU Cargar conjunto de datos de entrenamiento.

Plan de iteraciones

Debido a que XP asume que con un poco de planificación, un poco de codificación y unas pocas pruebas se puede decidir si se está siguiendo un camino acertado o equivocado, evitando así tener que echar marcha atrás demasiado tarde, se define el plan de iteraciones que se ejecutará antes de iniciar el desarrollo de la solución.

En el desarrollo de la solución se llevarán a cabo tres iteraciones, durante estas se implementarán las funcionalidades antes identificadas. La tabla 2.3 muestra la estimación de esfuerzos por cada HU, mientras que la tabla 2.4 registra la asignación de las HU a cada iteración y la duración de estas.

No.	Historias de Usuario	Estimación (Semana)
HU_1	Cargar conjunto de datos de entrenamiento	1
HU_2	Procesar imagen facial	3
HU_3	Generar fichero ARFF	1
HU_4	Etiquetar conjunto de datos de entrenamiento	1
HU_5	Visualizar agrupamiento	2
HU_6	Cargar imagen facial	1
HU_7	Clasificar imagen facial	3

Tabla 2.3: Estimación de esfuerzo por HU.

Iteración	Código	Historias de Usuario	Duración (semanas)
1	HU_1	Cargar conjunto de datos de entrenamiento	5
	HU_2	Procesar imagen facial	
	HU_3	Generar fichero ARFF	
2	HU_4	Etiquetar conjunto de datos de entrenamiento	3
	HU_5	Visualizar agrupamiento	
3	HU_2	Cargar imagen facial	4
	HU_7	Clasificar imagen facial	

Tabla 2.4: Plan de duración de iteraciones.

Plan de entrega

Gran parte del éxito del proyecto XP se debe, a que el cliente es quien conduce constantemente el trabajo hacia lo que aportará mayor valor de negocio, por lo que éste es quien define la prioridad que tendrá cada HU y de acuerdo con esta prioridad determina conjuntamente con el equipo de desarrollo cuál funcionalidad implementar primero, conformando así un plan de entrega.

No. Entrega	Fin de Iteración
1era Entrega (1era Iteración)	Febrero del 2014

2da Entrega (2da Iteración)	Abril del 2014
3era Entrega (3era Iteración)	Mayo del 2014

Tabla 2.5: Plan de entrega.

2.5.2 Fase de Diseño

Tarjetas CRC

Las tarjetas CRC (Clase - Responsabilidad - Colaboración), constituyen uno de los artefactos de la metodología XP que guía el proceso de desarrollo de la solución propuesta. La tabla 2.6 muestra la Tarjeta CRC definida para la clase *Solucion*, el resto de las Tarjetas CRC se muestran en los anexos. (Ver Anexo 5)

Nombre de la clase: Solucion	
Responsabilidades	1. PrepararDatos 2. GenerarARFF 3. EntendimientoDatos 4. Modelado 5. GuardarImagen 6. InsertarImagenEnBaseDatos 7. InsertarVectoresEnBaseDatos 8. LimpiarBaseDatos
Colaboradores	1. AccesoDatosGestion 2. AnalisisCaracteristicasPrincipales 3. ModeloAgrupamientoARFF 4. Clasificador

Tabla 2.6: Tarjeta CRC correspondiente a la clase Solucion.

Diagrama de clases del diseño

El diagrama de clases del diseño, permite describir la estructura de la solución mostrando sus clases, atributos y las relaciones entre ellos. La estructura que guiará el desarrollo de la solución, está dada por el diagrama de clases del diseño que se muestra en la figura 2.9. En los anexos se muestra el diagrama de clases del extractor de características. (Ver Anexo 6)

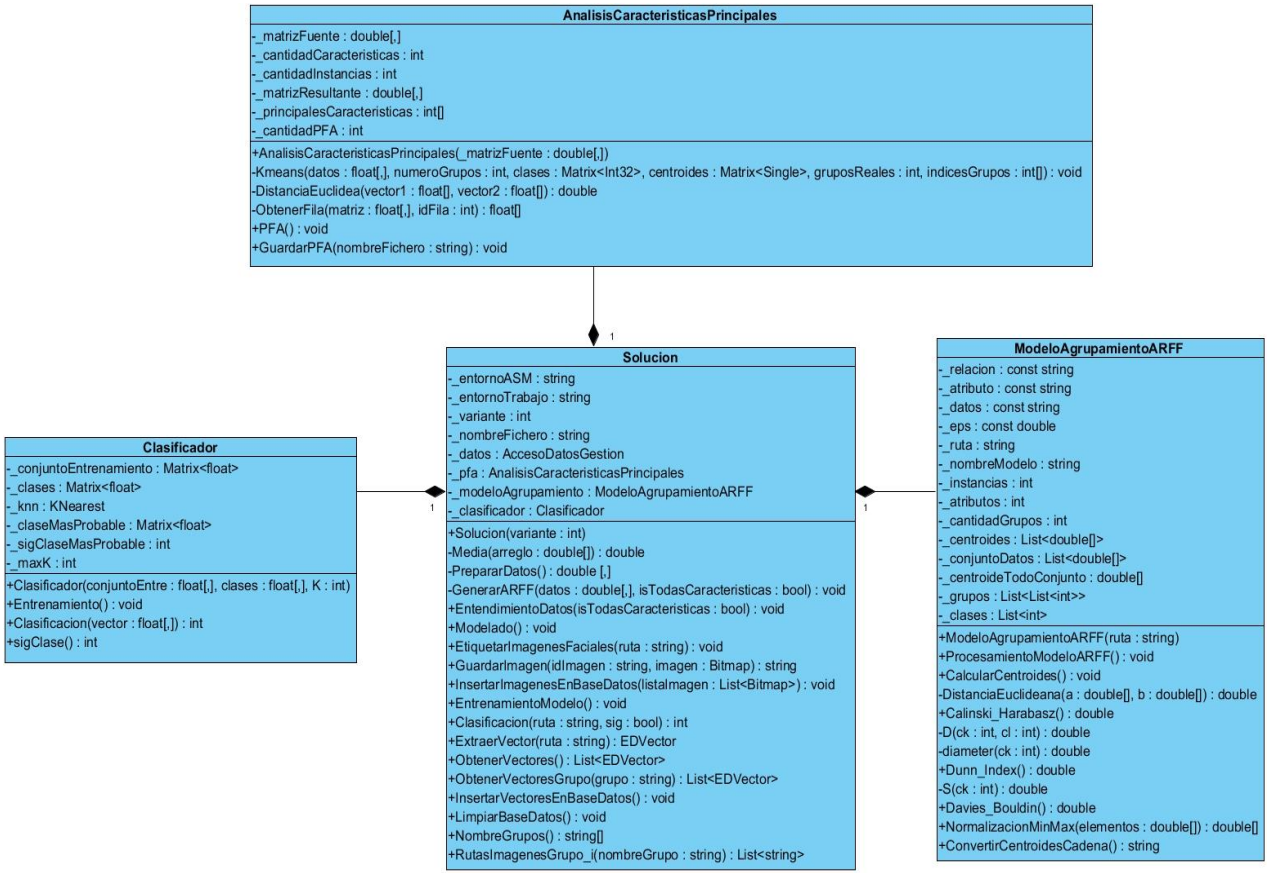


Figura 2.9: Diagrama de clases del diseño.

Modelo de datos de la solución

Para el modelado de los datos, en la figura 2.10 se propone el diagrama entidad-relación, que refleja la interrelación entre las tablas de la base de datos así como sus propiedades.

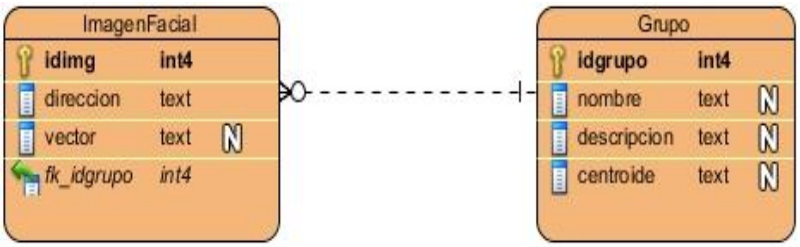


Figura 2.10: Diagrama entidad-relación.

Arquitectura de la solución

La arquitectura de software puede considerarse como el “puente” entre los requerimientos de la solución y la implementación. Constituye un artefacto de la actividad de diseño, que servirá de medio de comunicación entre los miembros del equipo de desarrollo, los clientes y usuarios finales, dado que contempla los aspectos que interesan a cada uno. Además pasa a ser la base del diseño de la solución a desarrollar, razón por la cual la arquitectura es considerada como plan de diseño de la solución, debido a que es usada como guía para el resto de las tareas del desarrollo (51).

Durante el desarrollo de la solución, se propone el uso de una **Arquitectura N capas**, como se muestra en la figura 2.11. Esta arquitectura tiene como objetivo principal reducir dependencias entre artefactos, para esto propone una estructura jerárquica en la que cada capa se sirve de los servicios que le brinda la capa inmediata inferior y a su vez proporciona servicios a su capa inmediata superior. Esta arquitectura brinda ventajas a los desarrolladores, al permitir la reutilización de componentes y un mejor mantenimiento. Se basa en una distribución jerárquica de roles y responsabilidades para proporcionar una división de los problemas a resolver. Los roles indican el tipo y la forma de interacción con otras capas, mientras que las responsabilidades representan la funcionalidad que implementan (52).

La arquitectura propuesta en la presente investigación, consta de 3 capas:

- **Capa Presentación:** contiene la interfaz gráfica de prueba de la solución y en ella las funcionalidades que ayudarán a obtener la clasificación final del rostro.
- **Capa del Negocio:** encargada de modelar la lógica de la solución al representar la organización de las clases y las relaciones que existen entre estas. Se comunica con la Capa de Acceso a Datos para gestionar la información al modificarla o consultarla. En esta capa se definen las clases encargadas de realizar todas las funcionalidades que deberá presentar la solución propuesta, tales como generar un fichero ARFF y extraer vectores de características.
- **Capa de Acceso a Datos:** tiene como objetivo principal acceder y modificar los datos que residen en ella. Es la capa más crítica y sensible de la arquitectura. Recibe solicitudes de almacenamiento y recuperación desde la capa inmediata superior.

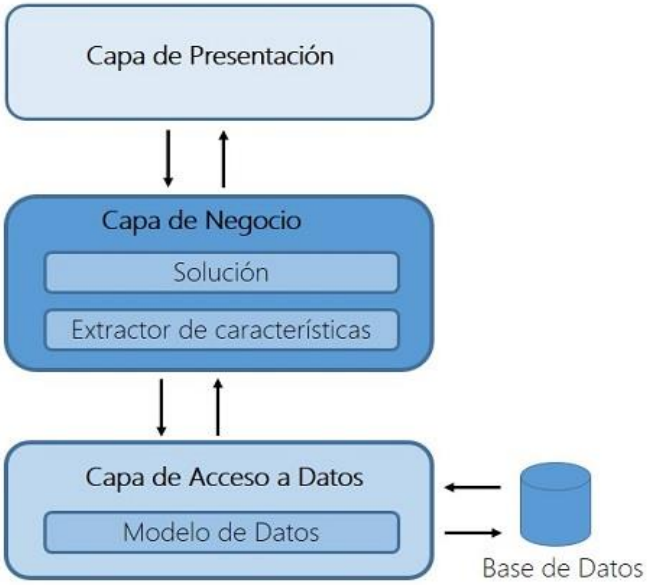


Figura 2.11: Arquitectura de la solución para la clasificación de rasgos biométricos faciales.

Según Buschman et al. (1996) un patrón arquitectónico es una descripción de un problema particular y recurrente de diseño, que aparece en contextos específicos y presenta un esquema genérico demostrado con éxito para su solución. El esquema de solución se especifica mediante la descripción de los componentes que la constituyen, sus responsabilidades y desarrollos, así como también la forma como estos colaboran entre sí (51).

El patrón arquitectónico propuesto es **Tuberías y Filtros**, ya que es apropiado para sistemas que implementan transformaciones de datos en pasos sucesivos, como es el caso de la solución a desarrollar. El filtro está diseñado para recibir la entrada de datos de una forma y producir la salida de datos de otra forma específica (9). La figura 2.12 muestra los filtros asociados a la fase de selección o extracción de características, reflejada anteriormente en la propuesta de solución. Señalar que todas las fases en la propuesta de solución, harán uso del estilo arquitectónico debido a las ventajas que brinda. Los filtros definidos como pasos secuenciales del procesamiento son:

- Entrada: Imagen facial capturada.
- Filtro 1: Detección del rostro.
- Filtro 2: Normalizar rostro.
- Filtro 3: Puntos Característicos.
- Filtro 4: Análisis de Características Principales.
- Salida: Vector característico.

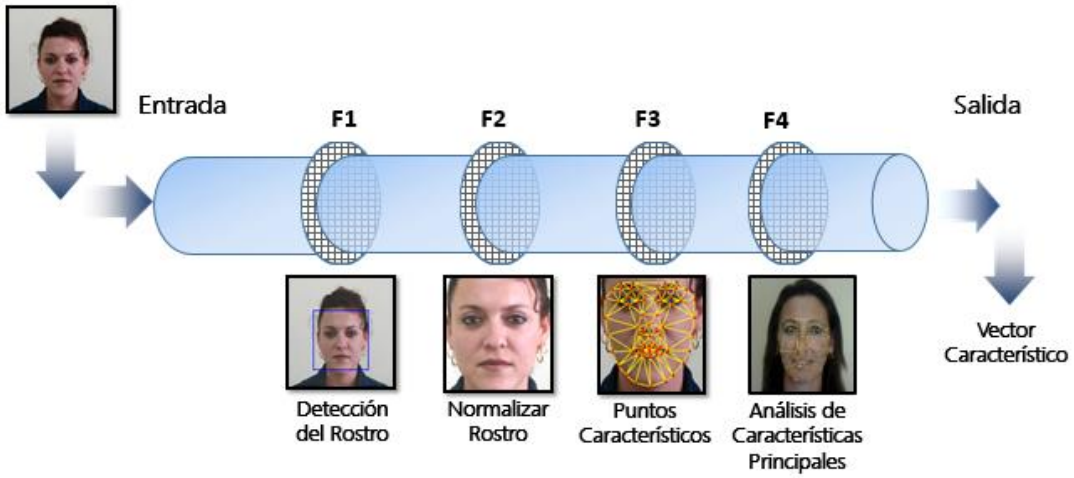


Figura 2.12: Filtro de la fase selección o extracción de características según el estilo arquitectónico Tuberías y Filtros.

Patrones de diseño

Un patrón de diseño es una solución a un problema en un contexto particular que describe las clases y objetos que se relacionan entre sí. En la ingeniería del software, un patrón constituye el apoyo para la solución a los problemas más comunes que se presentan durante las diferentes etapas del ciclo de vida del software. Además los patrones facilitan la reutilización de diseños, código y arquitecturas de software que han tenido éxito.

En resumen los patrones de diseño se caracterizan por ser soluciones concretas y técnicas, se utilizan en situaciones frecuentes y favorecen la reutilización de código (52).

Los patrones GRASP[34] son guías o principios que sirven para asignar responsabilidades a las clases y a objetos. Entre los patrones de diseño GRASP utilizados se encuentran:

- **Experto:** es el principio básico de asignación de responsabilidades. Indica, por ejemplo, que la responsabilidad de la creación de un objeto o la implementación de un método, debe recaer sobre la clase que conoce toda la información necesaria para crearlo. De este modo se obtendrá un diseño con mayor cohesión y así la información se mantiene encapsulada (disminución del acoplamiento). Este patrón se evidencia en las clases *Solucion* y *AnalisisCaracteristicasPrincipales*.
- **Creador:** el patrón creador ayuda a identificar quién debe ser el responsable de la creación (o instanciación) de nuevos objetos o clases. La nueva instancia deberá ser creada por la clase que presente algunas de las siguientes propiedades: tiene la información necesaria para realizar la creación del objeto, usa directamente las instancias creadas del objeto,

[34] Por sus siglas en inglés: *General Responsibility Assignment Software Patterns.*

almacena o maneja varias instancias de la clase, contiene o agrega la clase (53). Este patrón se pone de manifiesto en la clase *VectorCaracteristico* en la creación de instancias de clases relacionadas con la misma.

- **Bajo Acoplamiento:** el acoplamiento es una medida de la fuerza con que una clase está conectada a otras clases, con que las conoce y con que recurre a ellas. Una clase con bajo o débil acoplamiento no depende de muchas otras. Ejemplo de este patrón se presencia en la clase *ModeloAgrupamientoARFF*.
- **Alta Cohesión:** es la medida de cuán relacionadas y enfocadas están las responsabilidades de una clase. Cuando una clase presenta baja cohesión puede realizar acciones no afines o un trabajo extenso, mientras que cuando la cohesión es alta su utilidad también se eleva ya que resulta fácil darle mantenimiento, entenderla y reutilizarla. En la solución este patrón se evidencia en las clases *Solucion*, *Clasificacion* y *AnalisisCaracteristicasPrincipales*.
- **Controlador:** sirve como intermediario entre una determinada interfaz y el algoritmo que la implementa. Este patrón sugiere que la lógica de negocios debe estar separada de la capa de presentación, con el objetivo de aumentar la reutilización de código y a la vez tener un mayor control. El patrón se pone de manifiesto en la clase *Solucion*.

Por su parte los patrones GOF[35] describen soluciones simples y elegantes a problemas específicos en el diseño de software orientado a objetos. Entre estos fueron utilizados:

- **Patrones Creacionales:** se ocupan del proceso de creación de clases y objetos, son los encargados de abstraer el proceso de instanciación o creación de objetos. Estos ayudan a que la solución sea independiente de cómo sus objetos son creados, integrados y representados.
  - **Singleton (Instancia única):** asegura que una clase tiene una sola instancia y proporciona un punto de acceso global a ella.
    Por ejemplo:

```
solucion = Solucion.ObtenerInstancia();
public static Solucion ObtenerInstancia(int variante)
{
    if (_solucion == null){
      _solucion = new Solucion(variante);
      return _solucion;
    }
}
```

[35] Por sus siglas en inglés: Gang of Four.

2.6 Conclusiones parciales

La descripción detallada de la propuesta de solución permitió el desarrollo de dos tareas de la MD: agrupamiento y clasificación, así como la elección de los algoritmos a utilizar en cada fase, empleando la metodología de extracción del conocimiento Crisp-DM conjuntamente con XP. La definición del modelo de dominio permitió delimitar el alcance del problema a resolver. Fueron definidas las historias de usuario, que describen las principales funcionalidades que debe cumplir la solución y los planes de duración y entrega, con el objetivo de implementar la solución en el tiempo requerido. Se seleccionó la arquitectura N capas y los patrones de diseño a utilizar durante la implementación para lograr una solución acorde a las necesidades del cliente.

3 Capítulo 3: Implementación, Validación y Pruebas

3.1 Introducción

En el presente capítulo se definen las pautas de codificación que debe cumplir el equipo de desarrollo, así como las tareas de ingeniería necesarias para saber qué hacer en el desarrollo de cada HU. Son creados el diagrama de componente y el de despliegue, describiéndose en cada uno los elementos que los conforman. Además se realizan las pruebas de efectividad y funcionalidad y se validan los modelos utilizados en la fase de MD, para determinar si la solución desarrollada cumple con los requerimientos definidos por el cliente en las primeras fases del desarrollo del producto.

3.2 Implementación

3.2.1 Pautas de codificación

XP enfatiza en la comunicación de los programadores a través del código, con lo cual es indispensable que se sigan ciertos estándares de programación. Estos mantienen el código legible para los miembros del equipo facilitando los cambios.

Para la implementación de la solución, las pautas de codificación están compuestas por reglas de codificación, en las que están presentes las estructuras *camel case* y *pascal case* definidas en (54). Las reglas de codificación son las siguientes:

1. El nombre del proyecto debe ser escrito en español.
2. El código fuente debe ser escrito en español.
3. Cada funcionalidad debe tener comentario de su funcionamiento en español.
4. Los nombres de las clases, interfaces, métodos, propiedades y enumeradores deben tener la estructura *pascal case*. Ejemplo: GenerarARFF.
5. Los atributos y las constantes deben tener la estructura "_"*camel case*. Ejemplo: _conjuntoEntrenamiento.
6. Los parámetros deben tener la estructura *camel case*. Ejemplo: ejemploClasificar.
7. Los nombres de los métodos deben reflejar la acción a realizar. Ejemplo: GuardarImagen (nombre del método que guarda las imágenes en una ruta especificada).

3.2.2 Tareas de ingeniería

Durante el transcurso de las iteraciones, se realiza la implementación de las HU definidas por el cliente y descritas por el equipo de desarrollo en la etapa de planificación (9). Dichas HU se descomponen en partes más pequeñas, llamadas tareas de ingeniería. Estas son asignadas a un programador y a una iteración determinada, siendo el programador el responsable de su implementación durante la iteración asignada. La tabla 3.1 muestra las tareas de ingeniería correspondientes a la iteración 1 del proceso de desarrollo. Las tareas de las siguientes iteraciones, así como la descripción de estas se muestran en los anexos. (Ver Anexo 7 y Anexo 8 respectivamente)

Iteración 1	
Historia de Usuario	**Tareas de ingeniería**
Cargar conjunto de datos de entrenamiento.	1. Cargar desde una dirección específica las imágenes faciales.
Procesar imagen facial.	1. Detectar el rostro en la imagen facial. 2. Normalizar la imagen facial. 3. Extraer vector de características.
Generar fichero ARFF.	1. Convertir los vectores de la base de datos a formato ARFF para ser interpretado por la herramienta de MD.

Tabla 3.1: Distribución de tareas de ingeniería en la iteración 1.

3.2.3 Diagrama de componentes

Los diagramas de componentes son un modelado de la vista estática y dinámica del software. Para representar la división de la solución en componentes y las dependencias entre estos fue definido un diagrama de componentes, el cual se muestra en la figura 3.1.

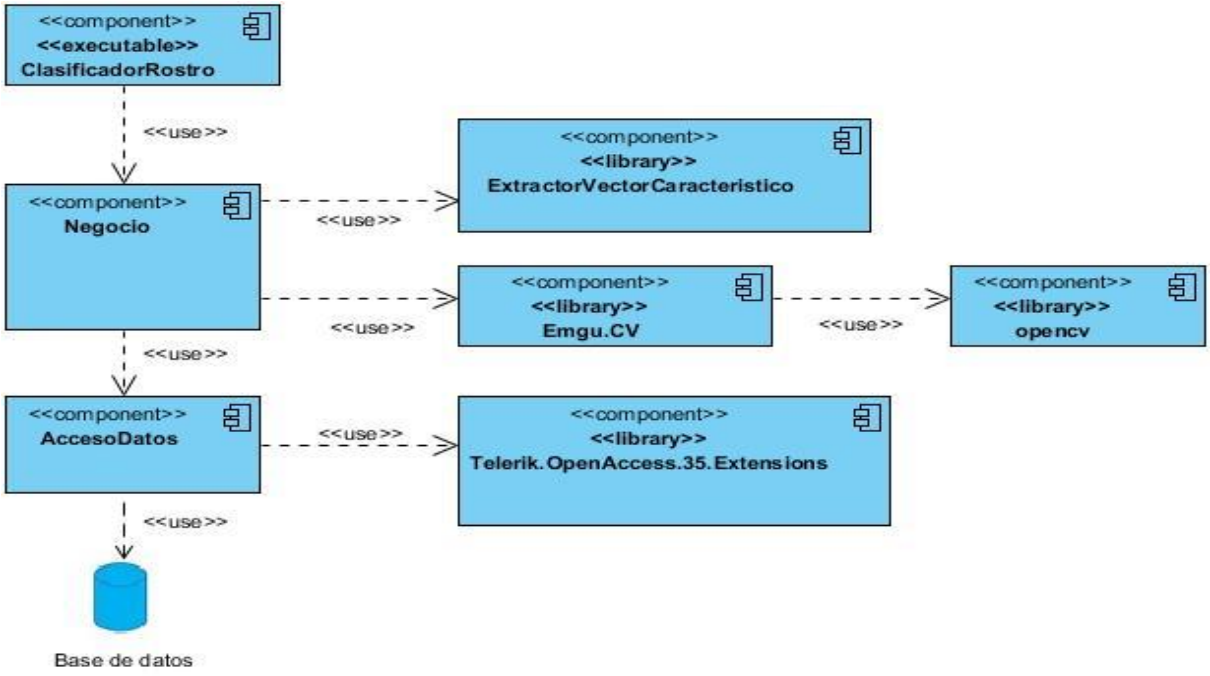


Figura 3.1: Diagrama de componentes de la solución.

Descripción de componentes

- **ClasificadorRostro:** componente que permitirá visualizar las funcionalidades de la solución.
- **Negocio:** componente que se comunica con el componente inmediato superior, recibiendo las peticiones y presentando los resultados de estas y con el componente inmediato inferior, consultando y verificando sus informaciones.
- **AccesoDatos:** componente que tiene como tarea principal acceder y modificar los datos de la solución.
- **ExtractorVectorCaracteristico:** componente encargado de realizar todo el proceso de extracción de características.
- **Emgu.CV:** componente contenedor de la opencv.
- **Telerik.OpenAccess.35.Extensions:** componente encargado del trabajo con la base de datos.
- **opencv:** componente encargado del procesamiento de imágenes.

3.2.4 Diagrama de despliegue

El diagrama de despliegue captura los elementos de configuración del procesamiento, las relaciones entre estos y visualiza su distribución en los nodos físicos. El modelo de despliegue de la solución para la clasificación de rasgos biométricos faciales, está compuesto por la biblioteca de enlace dinámico, la cual será integrada al sistema de reconocimiento facial que estará instalado en una

computadora. El cual intercambiará información con el servidor de base de datos mediante el protocolo TCP/IP.

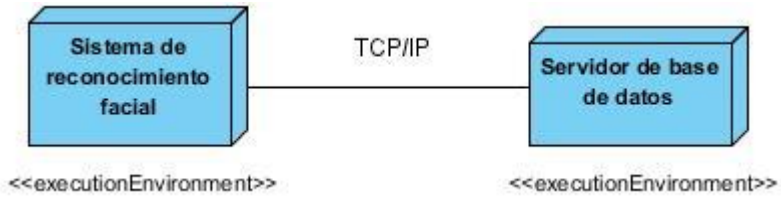


Figura 3.2: Diagrama de despliegue de la Solución para la clasificación de rasgos biométricos faciales.

3.3 Validación

3.3.1 Agrupamiento

Como parte de la validación del modelo de agrupamiento, para cada procedimiento utilizado en la investigación, se muestran los distintos valores de K y su valor $DF - A$ correspondiente, la regresión lineal representada por la función lineal que mejor se ajusta a los datos y un coeficiente de determinación (R^2) que brinda la información de qué tan adecuada es la recta de ajuste, siendo un valor muy cercano a 1 un ajuste correcto y un valor cercano a 0 un ajuste inadecuado. También se muestran dos estadísticas derivadas del modelo de regresión (55), una para determinar la probabilidad de que los resultados se produjeron por azar y otra para determinar si el valor de K es útil para estimar el índice propuesto ($DF - A$), en este caso es comparado su valor absoluto con 2.262157, que es el valor crítico para la distribución t con un nivel de confianza del 95% y un determinado grado de libertad y si es mayor entonces se puede afirmar la incidencia de la variable en los resultados obtenidos. El cálculo de las estadísticas anteriores se realizó con la herramienta Excel. Además pueden consultarse las tablas asociadas al procedimiento utilizado para obtener los valores del índice DF-A y la tabla correspondiente a los resultados obtenidos por el esquema basado en votos. En las figuras 3.3 a la 3.8 se muestran los resultados de estas validaciones para los procedimientos Transformada Discreta del Coseno (DCT), Transformada de Gabor Wavelet (GWT) e Imagen Icono (IC). Debido a lo extensión de estos resultados los restantes se muestran en los anexos. (Ver Anexo 9 y Anexo 10)

K	DF-A
31	**0.953433**
32	0.851667
33	0.6928
34	0.708233
35	0.674433
36	0.6247
37	0.5555
38	0.4527
39	0.665733
40	0.462867
41	0.401433

(a)

Estadísticas	Valor
Probabilidad	0.000155
Utilidad	-6.22171

(b)

	CH	Dunn	DB	Total
31	**10**	**3**	**4**	**17**
32	0	3	1	4
33	0	0	0	0
34	0	0	2	2
35	0	0	1	1
36	0	0	1	1
37	0	0	0	0
38	0	0	0	0
39	0	3	1	4
40	0	1	0	1
41	0	0	0	0

(c)

Figura 3.3: Validación del agrupamiento para el procedimiento DCT: tabla (a) muestra los grupos y el CVI DF-A, tabla (b) representa las estadísticas calculadas a partir del modelo de regresión, tabla (c) muestra los resultados del esquema basado en votos.

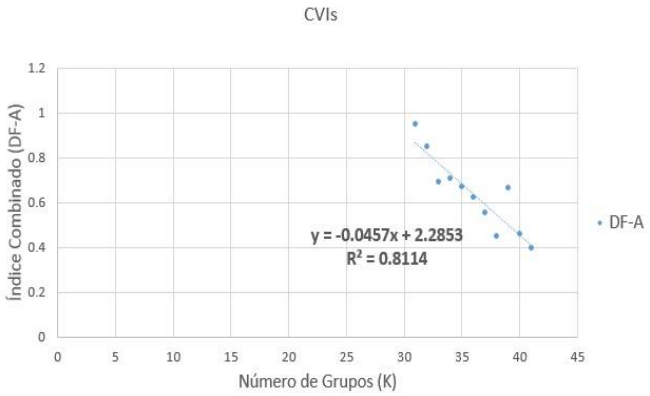


Figura 3.4: Regresión lineal asociada al procedimiento DCT.

K	DF-A
29	**0.966**
30	0.906
31	0.823067
32	0.799567
33	0.788433
34	0.780067
35	0.714467
36	0.627967
37	0.618567
38	0.6378
39	0.560967

(a)

Estadísticas	Valor
Probabilidad	3.21E-07
Utilidad	-13.292

(b)

	CH	Dunn	DB	Total
29	**10**	**2**	**2**	**14**
30	0	0	1	1
31	0	0	0	0
32	0	0	1	1
33	0	1	2	3
34	0	1	0	1
35	0	2	0	2
36	0	1	0	1
37	0	2	2	4
38	0	1	1	2
39	0	0	1	1

(c)

Figura 3.5: Validación del agrupamiento para el procedimiento GWT: tabla (a) muestra los grupos y el CVI DF-A, tabla (b) representa las estadísticas calculadas a partir del modelo de regresión, tabla (c) muestra los resultados del esquema basado en votos.

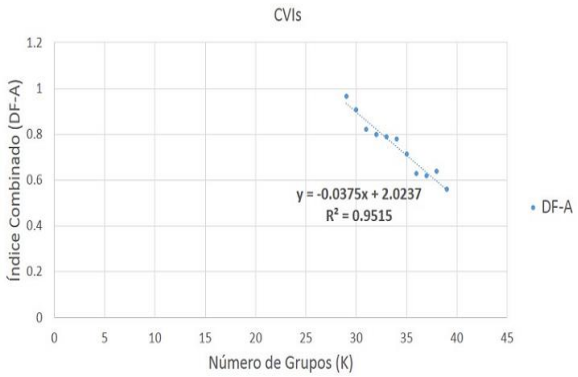


Figura 3.6: Regresión lineal asociada al procedimiento GWT.

K	DF-A
30	**0.941467**
31	0.843767
32	0.874133
33	0.6977
34	0.661233
35	0.647433
36	0.5814
37	0.570933
38	0.549267
39	0.491533
40	0.433933

(a)

Estadísticas	Valor
Probabilidad	6.7E-07
Utilidad	-12.1969

(b)

	CH	Dunn	DB	Total
30	**10**	**5**	**3**	**18**
31	0	1	3	4
32	0	1	1	2
33	0	0	0	0
34	0	1	0	1
35	0	1	1	2
36	0	0	0	0
37	0	0	1	1
38	0	1	1	2
39	0	0	0	0
40	0	0	0	0

(c)

Figura 3.7: Validación del agrupamiento para el procedimiento Imagen Icono: tabla (a) muestra los grupos y el CVI DF-A, tabla (b) representa las estadísticas calculadas a partir del modelo de regresión, tabla (c) muestra los resultados del esquema basado en votos.

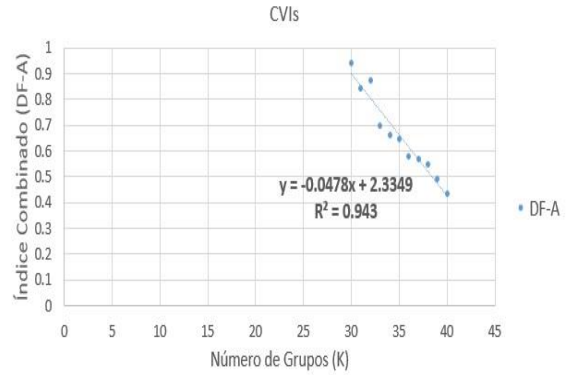


Figura 3.8: Regresión lineal asociada al procedimiento Imagen Icono.

3.4 Pruebas

Para determinar si el producto desarrollado cumple con la calidad requerida en aspectos técnicos y funcionales, se utilizan como instrumentos de verificación las pruebas al software. En este proceso se ejecutan pruebas dirigidas a componentes del software o al sistema de software en su totalidad, en este caso se ejecutarán pruebas dirigidas a la solución desarrollada, con el objetivo de medir el grado en que cumple con los requerimientos definidos en las primeras fases de su desarrollo.

3.4.1 Pruebas de rendimiento

Tiempo

Para realizar las pruebas de tiempo se eligieron 10 imágenes faciales, representadas en la figura por el número de prueba, perteneciendo cada imagen a los grupos más densos. Una vez seleccionadas las imágenes se realizó la identificación mediante dos vías, la primera comparando con todas las imágenes faciales y la segunda sectorizando el espacio de búsqueda mediante la solución propuesta. Concluyéndose que el tiempo promedio que demora la primera vía utilizada en identificar una persona es de 4.27 segundos en una base de datos del orden de las 10^4 personas enroladas. Una vez aplicada la solución pudo corroborarse que el tiempo promedio del proceso de reconocimiento mejora en 2.63 segundos. En la figura 3.9 se muestran dichos resultados, siendo TD el tiempo que demora la comparación de 1: N, TDC el tiempo luego de aplicada la solución y TM el tiempo mejorado.

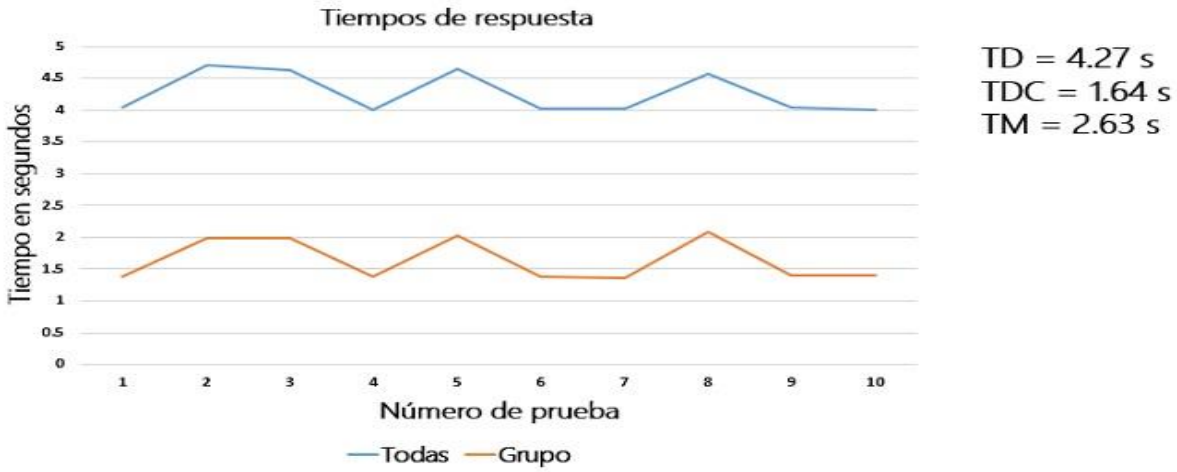


Figura 3.9: Resultados de las pruebas de rendimiento.

Efectividad

A continuación la tabla 3.2 muestra los modelos más eficientes para cada procedimiento representado a través del índice CH. En esta tabla los valores señalados en color rojo, representan la exclusión del procedimiento de la fase de clasificación, debido a que los modelos de agrupamiento generados por estos procedimientos, arrojaron una distribución de imágenes faciales en los grupos inadecuada, ya que presentaban grupos con bajos porcientos con respecto al total de imágenes y los índices de validación eran muy bajos en correspondencia con los demás procedimientos.

Procedimiento	CH
Imagen Icono	640.522
DCT	416.3606
GWT	388.006
HSV	276.1486
Imagen Reescalada	184.9653

Tabla 3.2: Modelos más eficientes para los procedimientos IC, DCT, GWT, HSV e IR.

La tabla 3.3 muestra los resultados del índice de precisión del clasificador K-NN para los procedimientos GWT, DCT e IC, las imágenes fueron separadas en dependencia de su categoría, es decir imágenes con variaciones en brillo, escala y pose. Concluyéndose que el procedimiento con mejor rendimiento en dos de las tres variaciones presentes es el GWT.

Procedimiento	K-NN			
	Brillo	Escala	Pose	Total
GWT	60.94817196	**37.43486974**	**68.07228916**	**55.4723039**
DCT	**79.71014493**	16.32736156	49.83948636	48.74865447
Imagen Icono	36.52849741	24.93573265	29.01023891	30.1401201

Tabla 3.3: Resultados del clasificador respecto a las variaciones de las imágenes en los procedimientos GWT, DCT e IC.

3.4.2 Pruebas unitarias

Pruebas de caja blanca

El enfoque estructural o de caja blanca se centra en la estructura interna del programa analizando los caminos de ejecución. Las pruebas de caja blanca intentan garantizar que:

- Se ejecutan al menos una vez todos los caminos independientes de cada método.
- Se utilizan las decisiones en su parte verdadera y en su parte falsa.
- Se ejecuten todos los bucles en sus límites.
- Se utilizan todas las estructuras de datos internas.

El método del camino básico propuesto por McCabe[36], permite obtener una medida de la complejidad de un diseño procedimental y utilizar esta medida como guía para la definición de una serie de caminos básicos de ejecución, diseñando casos de prueba que garanticen que cada camino se ejecuta al menos una vez.

La idea es derivar casos de prueba a partir de un conjunto dado de caminos independientes, por los cuales puede circular el flujo de control. Para obtener dicho conjunto de caminos independientes, se construye el grafo de flujo asociado al código y se calcula su complejidad ciclomática (52). La complejidad ciclomática, es una medida que proporciona una idea de la complejidad lógica de un programa. Para calcularla existen tres vías diferentes:

1. $V\,(G) = a - n + 2$, siendo a el número de arcos o aristas del grafo y n el número de nodos.
2. $V\,(G) = r + 1$, siendo r el número de regiones cerradas del grafo.
3. $V\,(G) = c + 1$, siendo c el número de nodos de condición.

Al aplicar el método del camino básico, es necesario realizar los pasos que se describen a continuación, ejemplificados en el algoritmo implementado para entrenar el clasificador, que se muestra en los anexos: (Ver Anexo 11)

1. Dibujar el grafo de flujo asociado al diseño o al código fuente.

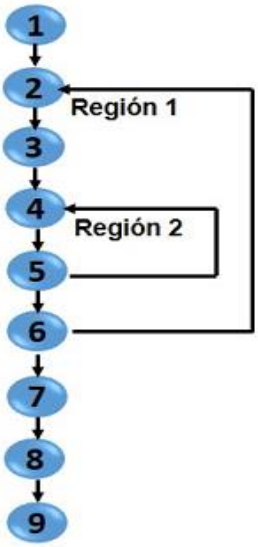


Figura 3.10: Grafo de Flujo del método para realizar el entrenamiento del modelo asociado al clasificador K-NN.

[36] Thomas J. McCabe, impulsor de esta métrica en 1976.

2. Calcular la complejidad ciclomática del grafo.

Fórmula 1	Fórmula 2	Fórmula 3
$V(G) = a - n + 2$ $V(G) = 10 - 9 + 2 = 3$	$V(G) = r + 1$ $V(G) = 3$	$V(G) = c + 1$ $V(G) = 2 + 1 = 3$

3. Determinar el conjunto básico de caminos independientes.
    a. 1-2-3-4-5-6-7-8-9
    b. 1-2-3-4-5-4-5-6-7-8-9
    c. 1-2-3-4-5-6-2-3-4-5-6-7-8-9
4. Preparar los casos de prueba para ejecutar cada camino del conjunto básico.

    Camino a (1-2-3-4-5-6-7-8-9)

    - Entrada:----------
    - Salida: modelo entrenado.
    - Precondición: El proceso de agrupamiento de las imágenes faciales debe estar realizado, con el objetivo de utilizar los centroides de cada grupo como datos de entrenamiento del modelo.

3.4.3 Pruebas de funcionalidad

Las pruebas de funcionalidad se enfocan en las acciones por parte del usuario y las respuestas por parte de la solución. Estas pruebas son realizadas con el objetivo de comprobar los requerimientos, considerando que una funcionalidad tiene éxito cuando se comporta de la manera esperada por el cliente. Para comprobar que las funcionalidades de la solución estén cumplidas, se diseñó para cada HU un caso de prueba de funcionalidad.

Pruebas de caja negra

Las pruebas de caja negra se llevan a cabo sobre la interfaz del software, obviando el comportamiento interno y la estructura del programa. Estas se centran en las funciones, entradas y salidas, es decir, que a través de los casos de prueba de funcionalidad se demuestra la operatividad de las funciones de la solución, que la entrada se acepta de forma correcta y que el resultado producido es el correcto. Se realizaron pruebas a diferentes funcionalidades de la solución, a continuación se muestra una de ellas, las restantes se muestran en los anexos. (Ver Anexo 12)

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 1	**HU:** Clasificar imagen facial.
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para verificar que la solución clasifica una imagen facial.	
Condiciones de ejecución: La imagen facial debe ser cargada previamente.	
Entrada/Pasos de ejecución: 1. Cargar imagen facial desde una dirección física. 2. Seleccionar la funcionalidad Clasificar imagen facial.	
Resultado esperado: Se muestra el número del grupo al que pertenece la imagen facial.	
Evaluación de la prueba: Satisfactoria	

Tabla 3.4: Caso de Prueba de la funcionalidad: Clasificar imagen facial.

3.5 Conclusiones parciales

Durante el desarrollo del presente capítulo fueron definidas las pautas de codificación con el objetivo de estandarizar el código para un mejor entendimiento por parte de otros desarrolladores, y las tareas de ingeniería a realizar por cada historia de usuario para guiar en la implementación de los requerimientos de la solución. Además fueron diseñados el diagrama de componentes para identificar los principales componentes y la relación entre ellos, y el diagrama de despliegue que permite conocer cómo se implantará la solución en el entorno del cliente.

Las validaciones realizadas a los modelos de agrupamiento, permitieron concluir que la población UCI cuenta con $K = 31$ clasificaciones por rasgos faciales determinadas con el procedimiento de extracción DCT, $K = 29$ con GWT y $K = 30$ con la Imagen Icono, además los modelos más eficientes asociados a estos procedimientos contienen un índice de agrupamiento CH de 416.3606, 388.006 y 640.522 respectivamente, con una probabilidad de que estos resultados fueron elegidos al azar del orden entre 10^{-4} y 10^{-7}. Además el clasificador propuesto presenta mejor rendimiento con el procedimiento GWT con un 55% de precisión, siendo el mejor en dos de las tres variaciones utilizadas en la prueba.

Por último las pruebas de rendimiento de tiempo permitieron comprobar la utilidad de la solución, mediante la disminución en 2.63 segundos del tiempo que presentaba el Sistema de reconocimiento facial en el proceso de identificación de una persona.

Conclusiones

Con el resultado de la presente investigación se puede concluir que:

- El estudio de los elementos teóricos permitió definir una propuesta de solución de acuerdo a las necesidades existentes en el centro.
- El empleo de la metodología Crisp-DM complementada para su correcta ejecución con XP permitió una solución para clasificar imágenes faciales en la población UCI y una estrategia de integración al sistema de reconocimiento facial.
- La solución propuso una estimación de las posibles clasificaciones de rasgos faciales presentes en la población UCI, presentándose la cantidad de clases o grupos como 31, 29 y 30 para los procedimientos DCT, GWT e IC respectivamente.
- Las validaciones realizadas a los modelos de agrupamiento permitieron confiabilidad y certeza en los resultados obtenidos sustentado en la probabilidad de que estos hayan sido por azar del orden de entre 10^{-4} y 10^{-7}.
- Se propuso un clasificador con un índice de precisión de 55% presente en el procedimiento GWT, concluyendo que era el procedimiento más efectivo en cuanto a variaciones en escala y pose.
- Las pruebas de rendimiento con respecto al tiempo, evidenciaron el cumplimiento de la hipótesis al reducir el tiempo en 2.63 segundos del proceso de reconocimiento utilizado en el Sistema de reconocimiento facial del CISED.

Recomendaciones

Para seguir profundizando sobre el uso de las técnicas de Minería de Datos en los procesos de reconocimiento facial se recomienda:

- Probar con otros algoritmos de reducción de dimensionalidad como LDA.
- Estimar el número de grupos con otros procedimientos existentes.
- Probar con otros algoritmos de agrupamiento de MD.
- Clasificar utilizando otros modelos como las redes neuronales o las máquinas de soporte vectorial.

Bibliografía referenciada

1. Reyes Pineda, Yusnieira y Santana Fernández, Ramón. *Trabajo de Diploma: Sintetizador de Huellas Dactilares.* La Habana : Universidad de las Ciencias Informáticas, 2011.

2. Marqués, Ion. *Face Recognition Algorithms.* s.l. : Universidad del país Vasco, 2010.

3. Biometria. [En línea] 15 de noviembre de 2013. http://www.biometria.gov.ar/metodos-biometricos/facial.aspx.

4. Kwan-Ho, LI, y otros. *AN EFFICIENT HUMAN FACE INDEXING SCHEME USING EIGENFACES.* s.l. : The Hong Kong Polytechnic University, 2003. pág. 4.

5. Hernández Orallo, José, Ramírez Quintana, María José y Ferri Ramírez, César. *Introducción a la Minería de Datos.* s.l. : Pearson Prentice Hall, 2004. 84-205-4091-9.

6. Hernández Meléndrez, Edelsys. Cómo escribir tesis. s.l. : Escuela Nacional de Salud Pública, 2006.

7. Biometrics. [En línea] [Citado el: 15 de noviembre de 2013.] http://www.biometrics.gov/Documents/FAQ.pdf.

8. Martínez Díaz, Yoanna, Méndez Vázquez, Heidy y García Reyes, Edel. *Reconocimiento de rostros en video.* s.l. : Centro de Aplicaciones de Tecnologías de Avanzada, 2012. 2072-6287.

9. Aguilera Reyes, Amelia y Quiles Velázquez, Rafael Alberto. *Sistema de identificación de personas basado en rasgos faciales.* La Habana : Universidad de las Ciencias Infromáticas, 2013.

10. H. Witten, Ian y Frank, Ebie. *Data Mining: Practical machine learning tools and techniques.* s.l. : Morgan Kaufmann, 2005.

11. Peña, Daniel. *Análisis de datos multivariantes.* Madrid : McGraw-Hill, 2002.

12. Nikolaou, Nikolaos. *Reducing Dimensionality for Face Recognition Miniproject for .* s.l. : CDT Computer Science, University of Manchester.

13. Shlens, Jonathon. *A Tutorial on Principal Component Analysis.* s.l. : New York University Systems Neurobiology Laboratory, 2009.

14. Shashua, Amnon. *Introduction to Machine Learning: Class Notes 67577.* s.l. : arXiv preprint arXiv, 2009. 0904.3664.

15. I. Smith, Lindsay. *A tutorial on principal components analysis.* USA : Cornell University, 2002.

16. Sean Cohen, Ira, y otros. *Feature Selection Using Principal Feature Analysis.* s.l. : Beckman Institute for Advanced Science and Technology, 2007.

17. Fukunaga, Keinosuke. *Introduction to statistical pattern recognition.* s.l. : Academic press, 1990.

18. R. Webb, Andrew. *Statistical paterrn recognition.* s.l. : John Wiley & Sons, 2003.

19. A. Fisher, Ronald. *The statistical utilization of multiple measurements.* s.l. : Annals of eugenics, 1938.

20. M. Martínez, Alexi y C. Kak, Avinash. *Pca versus Ida.* s.l. : Pattern Analysis and Machine Intelligence. IEEE , 2001.

21. Xu, Rui y Wunsch, Don. *Clustering.* s.l. : John Wiley & Sons, 2008.

22. Wu, Xindong y Kumar, Vipin. *The top ten algorithms in data mining.* s.l. : CRC Press, 2009.

23. Murty, Gorti Satyanarayana, Kumar, V. Vijaya y Naresh, Tangudu. *Learning Number of Clusters in Unlabeled Dataset using Rotation Estimation.* s.l. : Learning, 2012.

24. B. Do, Chuong y Batzoglou, Serafim. *¿What is the expectation maximization algorithm?* s.l. : Nature biotechnology, 2008.

25. Roche, Alexis. *EM algorithm and variants: An informal tutorial.* s.l. : arXiv preprint arXiv, 2011. 1105.1476.

26. Kohonen, Teuvo. *The self-organizing map.* s.l. : Proceedings of the IEEE, 1990.

27. —. *Self-organizing maps.* s.l. : Springer, 2001.

28. Pelleg, Dan. *X-mean: Extending K-means with Efficient Estimation of the Number of Clusters.* s.l. : ICML, 2000.

29. Cheeseman, Peter y Stutz, John. *Bayesian Classification (AutoClass): Theory and Results.*

30. H. Fisher, Douglas. *Knowledge acquisition via incremental conceptual clustering.* Irvine : Irvine Computational Intelligence Project, Department of Information and Computer Science, University of California, 1987.

31. H. Gennari, John, Langley, Pat y Fisher, Doug. *Models of incremental concept formation.* s.l. : Artificial intelligence, 1989.

32. Garcia Cambronero, Cristina y Gómez Moreno, Irene. *Algoritmos de Aprendizaje: knn & kmeans.* s.l. : Inteligencia en Redes de Comunicación, Universidad Carlos III de Madrid, 2006.

33. Martin Del Brio, Bonifacio y Sanz Molina, Alfredo. *Redes Neuronales y Sistemas difusos.* s.l. : Alfaomega Grupo Editor, 2002.

34. M. Mitchell, Tom. *Machine learning.* s.l. : McGraw Hill, 1997.

35. Hou, Chenping. *Learning a subspace for face image clustering via trece ratio criterion.* 2009.

36. Ding, Chris y He, Xiaofeng. *K-means clustering via principal component analysis.* s.l. : Proceedings of the twenty-first international conference on Machine learning, 2004.

37. —. *Principal Component Analysis and Effective K-Means.* s.l. : SDM, 2004.

38. Ding, Chris y Li, Tao. *Adaptive dimension reduction using discriminant analysis and means clustering.* s.l. : Proceedings of the 24th International Conference on Machine Learning (ICML-07), 2007.

39. Ordoñes Leyva, Yoanni y Avilés Vázquez, Ernesto . *Herramienta informática de Minería de Uso de la Web sobre los registros de navegación por Internet.* s.l. : Universidad de las Ciencias Informáticas, 2010.

40. Gallardo Arancibia, Jose Alberto. *Metodología para el desarrollo de proyectos en Minería de Datos: CRISP-DM.* La Habana : s.n.

41. Chapman, Pete, y otros. *CRISP-DM 1.0: Step by step data mining guide.* s.l. : SPSS, 2000.

42. Hernández Orallo, Enrique . *El Lenguaje Unificado de Modelado (UML).* s.l. : Addison Wesley, 2000.

43. Alegsa. [En línea] [Citado el: 21 de enero de 2014.] http://www.alegsa.com.ar/Dic/lenguaje de programacion.php.

44. Programación .NET. *Programación .NET.* [En línea] [Citado el: 20 de Mayo de 2014.] http://www.tuprogramacion.com/glosario/que-es-un-orm/.

45. Tamayo Valero, Herbert y Barrios Pérez, Sairenys . *Componente para el reconocimiento de rostros mediante el uso de Eigenfaces.* La Habana : Universidad de las Ciencias informáticas, 2013.

46. Arthur, David y Vassilvitskii, Sergei. *K-means++: The advantages of careful seeding.* s.l. : Society for Industrial and Applied Mathematics, 2007.

47. Arbelaitz, Olatz , y otros. *An extensive comparative study of cluster validity indices.* s.l. : Department of Computer Architecture and Technology, University of the Basque Country UPV/EHU, 2012.

48. C. Bezdek, James y R. Pal, Nikhil. *Some new indexes of cluster validity.* s.l. : IEEE Transactions on Cybernetics, 1998.

49. Kryszczuk, Krzysztof y Hurley, Paul . *Estimation of the number of clusters using multiple clustering validity indices.* Switzerland : IBM Zurich Research Laboratory.

50. Sommerville, Ian. *Ingeniería del software.* Séptima edición. Madrid : Pearson Educación, 2005. 84-7829-074-5.

51. Camacho, Erika, Cordeso, Fabio y Nuñez, Gabriel. Universidad Simón Bolivar. [En línea] [Citado el: 5 de febrero de 2014.] http://prof.usb.ve/lmendoza/Documentos/PS-6116/Guia Arquitectura v.2.pdf.

52. Delgado Hernández, Alicia. *Componente de clasificación de huellas dactilares.* La Habana : Universidad de las Ciencias Informáticas, 2013.

53. TuxNots. [En línea] [Citado el: 12 de febrero de 2014.] https://sites.google.com/site/tuxnots/home/materias-de-la-facu/metodologia-de-sistemas/patrones-grasp-patrones-gof-diferencia-entre-grasp-y-gof.

54. Nane, Silvia . *Estándares de codificación para C#.* s.l. : Infocorp, 2014.

55. Microsoft. [En línea] 24 de febrero de 2014. http://office.microsoft.com/es-es/excel-help/estimacion-lineal-HP005209155.aspx.

Bibliografía consultada

- Acharya, Tinku y K. Ray, Ajoy. 2005. *Image processing: principles and applications.* s.l. : John Wiley & Sons, 2005. 13978-0-471-71998-4.
- G. Figueroa, Roberth, J.Solís, Camilo y A.Cabrera, Armando. 2008. *Metodologías Tradicionales vs. metodologías Ágiles.* s.l. : Escuela de Ciencias en Computación. Universidad Técnica Particular de Loja, 2008.
- H. Witten, Ian y Frank, Ebie. 2005. *Data Mining: Practical machine learning tools and.* s.l. : Morgan Kaufmann, 2005.
- Hernández Orallo, José, Ramírez Quintana, María José y Ferri Ramírez, César. 2004. *Introducción a la Minería de Datos.* s.l. : Pearson Prentice Hall, 2004. 84-205-4091-9.
- Kroll, Per y Kruchten, Philippe. 2003. *The rational unified process made easy: a practitioner's guide to the RUP.* s.l. : Addison-Wesley Professional, 2003.
- M. Mitchell, Tom. 1997. *Machine learning.* s.l. : McGraw-Hill Science, 1997. 0070428077.
- Marcelo Varela, Maria Virginia. 1974. *Algebra lineal.* s.l. : Pueblo y Educación, 1974.
- Molina López, José Manuel y García Herrero, jesús. 2004. *Técnicas de análisis de datos.* s.l. : Universidad Carlos III.Madrid, 2004.
- Molina, R. 1998. *Introducción al procesamiento y análisis de imágenes digitales.* s.l. : Departamento de Ciencias de la Computaciín e IA Universidad de Granada. España, 1998.

- Phillips, Dwayne. 2000. *Image Processing in C.* s.l. : Lawrence: R & D Publications, 2000. 0-13-104548-2.
- S. Pressman, Roger. 2010. *Software Engineering: A Practitioner's Approach.* s.l. : McGraw-Hill, 2010. 978-0-07-337-597-7.
- Shashua, Ammon. 2008. *Introduction to machine learning.* s.l. : School of Computer Science and Engineering. The Hebrew University of Jerusalem, 2008.
- Sommerville, Ian. 2005. *Ingeniería del software.* s.l. : Pearson Educación, 2005.

Glosario de términos

- Agrupamiento: consiste en obtener grupos naturales a partir de los datos.
- Clasificación: agrupa a todas las herramientas que permiten asignar a un elemento la pertenencia a un determinado grupo o clase.
- Ecualización del histograma: Técnica que consiste en ajustar los niveles de gris de una imagen para obtener una nueva imagen con un histograma uniforme.
- Framework: estructura de artefactos o módulos concretos con base en la que otro proyecto de software puede ser desarrollado.
- Normalización: acción de transformar una distribución cualquiera a una distribución normal manteniendo la proporción de los datos.
- Rasgos conductuales: son aquellos que se soportan sobre características de la conducta del ser humano.

Acrónimos

- ARFF: Attribute-Relation File Format.
- ASM: Active Shape Model.
- CASE: Ingeniería de Software Asistida por Ordenador.
- CENATAV: Centro de Aplicaciones de Tecnologías de Avanzada.
- CISED: Centro de Identificación y Seguridad Digital.
- CRC: Tarjeta Clase-Responsabilidad-Colaboración.
- Crisp-DM: Cross Industry Standard Process for Data Mining.
- CVIs: Índices de Validación del Agrupamiento (por su significado en español).
- DCT: Transformada Discreta del Coseno (por su significado en español).
- GOF: Banda de los Cuatro.
- GRASP: Patrones de asignación de responsabilidades.
- HU: Historias de usuario.
- IDE: Entorno de desarrollo integrado, también conocido como entorno de diseño integrado o entorno de depuración integrada.
- IPD: Distancia interpupilar (por sus siglas en inglés).
- OpenCV: Visión por Computadora de Código Abierto.
- PCA: Análisis de Componentes Principales (por su traducción al español).
- UCI: Universidad de las Ciencias Informáticas.
- UML: Lenguaje Unificado de Modelado.
- XP: Programación Extrema.

4 Anexos

4.1 **Anexo 1:** Variables identificadas en la hipótesis de la investigación.

Variables independientes	**Dimensión**	**Indicadores**	**Escala**
Agrupamiento	Procedimiento	Técnica utilizada	No procede
	Dimensión de los datos	Cardinalidad de los datos	Número de muestras a agrupar
	Muestras	Vector característico utilizado	No procede
	Validación	1. CVIs 2. Esquema de votos	Valor numérico
Clasificación	Procedimiento	Técnica utilizada	No procede
	Tolerancia	Tasa de error	[0,100]%
Hardware especializado	Prestaciones del *hardware*	Velocidad del CPU	Megahertz (MHz)
		Capacidad de memoria RAM	Gigabyte (GB)
Variables dependientes	**Dimensión**	**Indicadores**	**Escala**
Tiempo de respuesta	Tiempo	$\Delta t = t_f - t_0$	Segundos (s)

Tabla 4.1: Descripción de las variables dependientes e independientes de la hipótesis.

4.2 **Anexo 2:** Entrevista realizada a Heydi Méndez Vázquez responsable de las investigaciones de reconocimiento facial desarrolladas en el CENATAV.

Modelo de Entrevista	
Entrevistado: Heydi Méndez Vázquez.	**Entrevistador:** Anays Gómez García. Gregorio Ferrer Cordova.
Fecha de realización: 7/10/2013	**Hora Inicio:** 10:30 am
	Hora Fin: 11:30 am
Preguntas realizadas	
Entrevistador	**Entrevistado**
¿En el centro actualmente se están realizando investigaciones sobre clasificación de imágenes faciales?	La verdad estamos trabajando en una línea de investigación similar, pero que se basa en clasificación de cualquier tipo de imagen, independientemente de si es facial o no.

¿Se ha desarrollado alguna investigación relacionada con nuestro tema?	Actualmente se está trabajando en indexación de imágenes faciales, hemos avanzado en la investigación hasta conformar un estado del arte de las principales técnicas utilizadas.
¿En ese campo de investigación en que método han indagado?	Aunque existen muchas propuestas, queremos centrarnos en un método denominado bolsa de características.
¿Algunos de los métodos investigados en el centro nos pudiera servir para lo que queremos hacer?	En realidad son campos distintos, quizás podamos proporcionarles algunas bibliografías pero no creo que puedan servirles de mucho.
Muchas gracias por su tiempo.	Fue un placer ayudarlos.

Tabla 4.2: Entrevista realizada en el CENATAV a Heydi Méndez Vázquez responsable de las investigaciones de reconocimiento facial.

4.3 **Anexo 3:** Votación del Kdnuggest del 2007 respecto a las metodologías de MD más usadas.

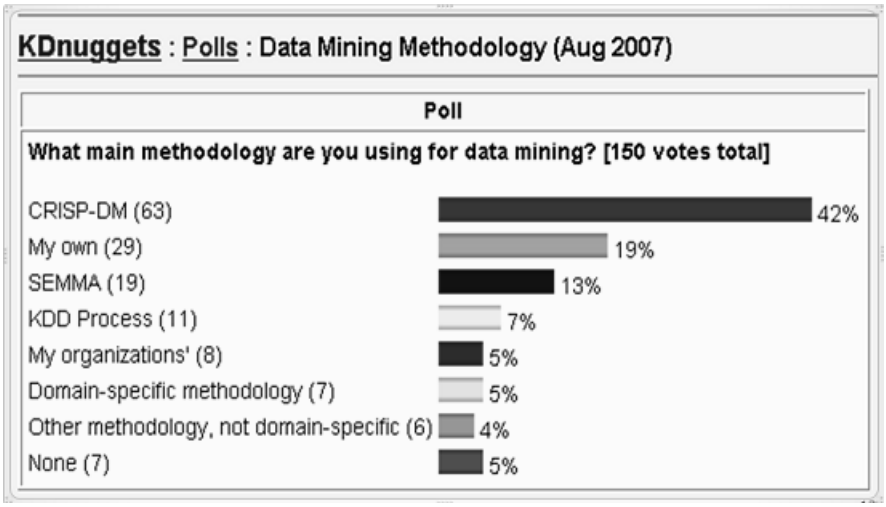


Figura 4.1: Votación entre las metodologías de Minería de Datos más utilizadas[37].

4.4 **Anexo 4:** Descripción de las historias de usuario de la Solución para la clasificación de rasgos biométricos faciales.

Historia de Usuario	
Número: HU_2	**Nombre de Historia:** Procesar imagen facial.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 3	**Iteración asignada:** 1

[37] Tomado de: http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm.

Programador responsable: Gregorio Ferrer Cordova
Descripción: La solución debe procesar una imagen facial a través de un conjunto de filtros que permitan extraer el vector característico asociado a la imagen por el método utilizado en el Sistema de reconocimiento facial.
Observaciones: Procesar imagen facial incluye: detectar rostro, normalizar imagen facial, extraer vector de características.

Tabla 4.3: HU_Procesar imagen facial.

Historia de Usuario	
Número: HU_3	**Nombre de Historia:** Generar fichero ARFF.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 2	**Iteración asignada:** 1
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe permitir convertir los vectores asociados a cada imagen facial almacenada en base de datos al formato ARFF para ser interpretado en la herramienta de Minería de datos.	
Observaciones: _____	

Tabla 4.4: HU_Generar fichero ARFF.

Historia de Usuario	
Número: HU_4	**Nombre de Historia:** Etiquetar conjunto de datos de entrenamiento.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 3	**Iteración asignada:** 2
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe permitir asignar una imagen facial a un único grupo en la base de datos.	
Observaciones:_____	

Tabla 4.5: HU_Etiquetar conjunto de datos de entrenamiento.

Historia de Usuario	
Número: HU_5	**Nombre de Historia:** Visualizar agrupamiento.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 2	**Iteración asignada:** 2
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe permitir la visualización de los resultados del proceso de agrupamiento.	
Observaciones:_____	

Tabla 4.6: HU_Visualizar agrupamiento.

Historia de Usuario	
Número: HU_6	**Nombre de Historia:** Cargar imagen facial.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 1	**Iteración asignada:** 3
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe permitir cargar una imagen facial.	
Observaciones:_____	

Tabla 4.7: HU_Cargar imagen facial.

Historia de Usuario	
Número: HU_7	**Nombre de Historia:** Clasificar imagen facial.
Usuario: Sistema Multibiométrico, Sistema de reconocimiento facial	
Prioridad en el Negocio: Alta	**Riesgo en el desarrollo:** Bajo
Puntos estimados: 3	**Iteración asignada:** 3
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe permitir la clasificación de una imagen facial en uno de los grupos obtenidos en la fase del agrupamiento.	
Observaciones:_____	

Tabla 4.8: HU_Clasificar imagen facial.

4.5 **Anexo 5:** Descripción de las Tarjetas CRC pertenecientes a la capa del negocio.

Nombre de la clase: AnalisisCaracteristicasPrincipales	
Responsabilidad	1. AnalisisCaracteristicasPrincipales 2. Kmeans 3. DistanciaEuclidea 4. ObtenerFila 5. PFA 6. GuardarPFA
Colaboradores	_____

Tabla 4.9: Tarjeta CRC correspondiente a la clase AnalisisCaracteristicasPrincipales.

Nombre de la clase: Clasificador	
Responsabilidad	1. Entrenamiento 2. Clasificacion 3. SigClase
Colaboradores	_____

Tabla 4.10: Tarjeta CRC correspondiente a la clase Clasificador.

Nombre de la clase: ModeloAgrupamientoARFF	
Responsabilidad	1. Ruta 2. NombreModelo 3. Instancias 4. Atributos 5. CantidadGrupos 6. Centroides 7. ConjuntoDatos 8. CentroideTodoConjunto 9. Grupos 10. Clases 11. ModeloAgrupamientoARFF 12. ProcesamientoModeloARFF 13. CalcularCentroides 14. DistanciaEuclideana 15. Calinski_Harabasz 16. Dunn_Index 17. Davies_Bouldin 18. NormalizacionMinMax

	19. ConvertirCentroidesCadena
Colaboradores	_____

Tabla 4.11: Tarjeta CRC correspondiente a la clase Modelo Agrupamiento ARFF.

4.6 **Anexo 6:** Diagrama de clases del diseño.

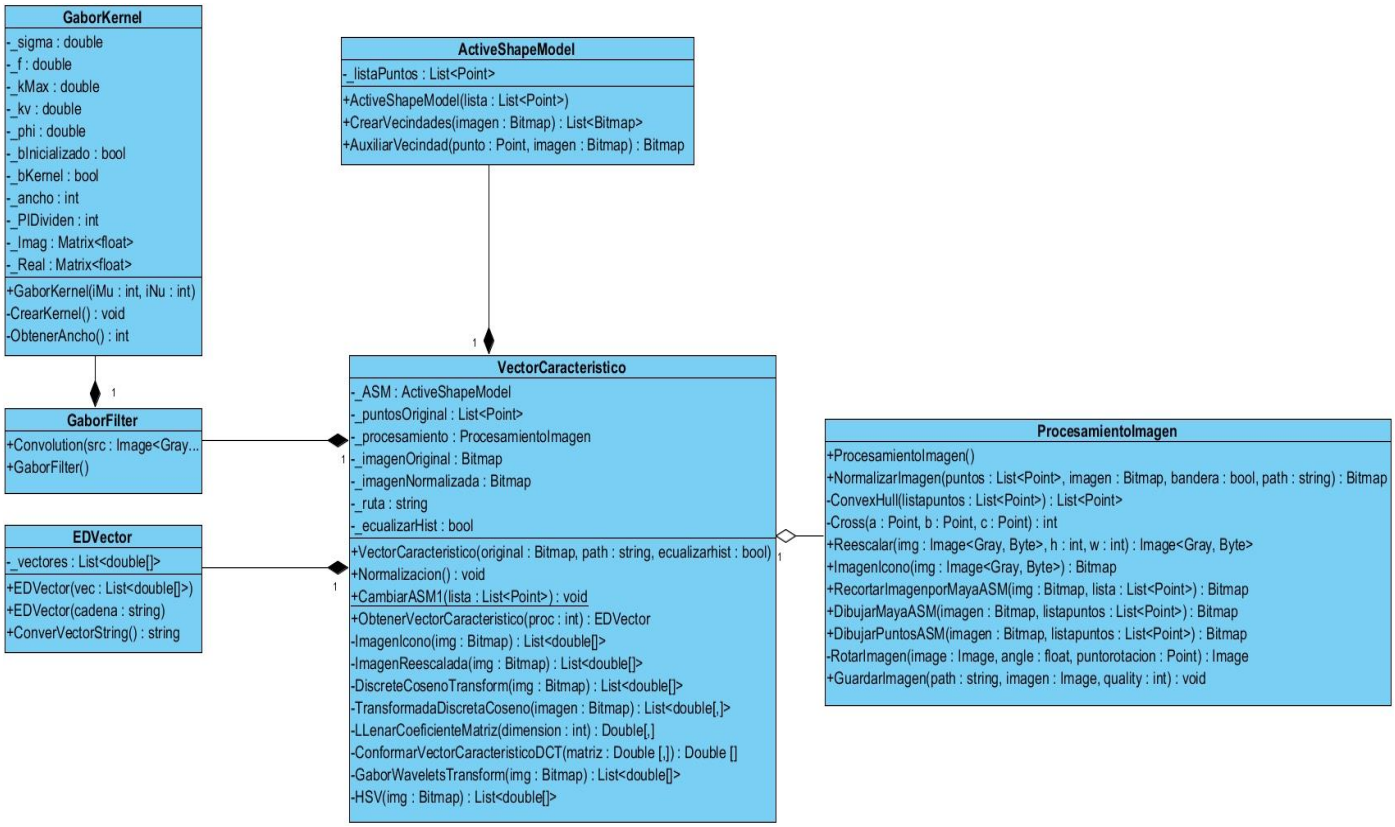


Figura 4.2: Diagrama de clases asociado al extractor de características.

4.7 **Anexo 7:** Distribución de las Tareas de ingeniería para las iteraciones 2 y 3.

Iteración 2	
Historia de Usuario	**Tareas de Ingeniería**
Etiquetar conjunto de datos de entrenamiento.	1. Crear grupos en la base de datos. 2. Asignar una imagen facial a un único grupo.
Visualizar agrupamiento.	1. Buscar en base de datos un grupo. 2. Visualizar imágenes faciales del grupo.

Tabla 4.12: Distribución de tareas de ingeniería en la iteración 2.

Iteración 3	
Historia de Usuario	**Tareas de Ingeniería**
Cargar imagen facial.	1. Cargar desde una dirección específica la imagen facial.
Clasificar imagen facial.	1. Devolver el número del grupo al que pertenece la imagen facial.

Tabla 4.13: Distribución de tareas de ingeniería en la iteración 3.

4.8 **Anexo 8:** Especificación de las Tareas ingenieriles.

Tarea de ingeniería	
Número: TI_1_HU_1	**Historia de usuario:** Cargar conjunto de datos de entrenamiento
Nombre de tarea: Cargar desde una dirección específica el conjunto de datos de entrenamiento compuesto por imágenes faciales.	
Tipo: Desarrollo	**Puntos estimados:** 1
Fecha inicio: 27/01/2014	**Fecha fin:** 3/02/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: La solución debe cargar un conjunto de imágenes faciales desde una dirección específica.	

Tabla 4.14: Tarea de Ingeniería 1 correspondiente a la HU_1.

Tarea de ingeniería	
Número: TI_1_HU_2	**Historia de usuario:** Procesar imagen facial
Nombre de tarea: Detectar el rostro en la imagen facial	
Tipo: Desarrollo	**Puntos estimados:** 1
Fecha inicio: 3/02/2014	**Fecha fin:** 10/02/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Se debe detectar el rostro en la imagen facial así como los puntos característicos generados por la librería ASM.	

Tabla 4.15: Tarea de Ingeniería 1 correspondiente a la HU_2.

Tarea de ingeniería	
Número: TI_2_HU_2	**Historia de usuario:** Procesar imagen facial
Nombre de tarea: Normalizar la imagen facial.	
Tipo: Desarrollo	**Puntos estimados:** 1
Fecha inicio: 10/02/2014	**Fecha fin:** 17/02/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Se normaliza la imagen facial a través de la IPD y la ecualización del histograma.	

Tabla 4.16: Tarea de Ingeniería 2 correspondiente a la HU_2.

Tarea de ingeniería	
Número: TI_3_HU_2	**Historia de usuario:** Procesar imagen facial
Nombre de tarea: Extraer vector de características.	
Tipo: Desarrollo	**Puntos estimados:** 1
Fecha inicio: 17/02/2014	**Fecha fin:** 24/02/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Utilizando los 68 puntos localizados a partir del algoritmo ASM, se utiliza el filtro DCT para extraer las características generando un vector característico.	

Tabla 4.17: Tarea de Ingeniería 3 correspondiente a la HU_2.

Tarea de ingeniería	
Número: TI_1_HU_3	**Historia de usuario:** Generar fichero ARFF
Nombre de tarea: Convertir los vectores de la base de datos a formato ARFF para ser interpretado por la herramienta de MD.	
Tipo: Desarrollo	**Puntos estimados:** 1
Fecha inicio: 24/02/2014	**Fecha fin:** 3/03/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Los datos asociados a cada vector característico de las imágenes faciales almacenadas en base de datos serán transformados en un fichero ARFF para ser interpretado por la herramienta de MD Weka.	

Tabla 4.18: Tarea de Ingeniería 1 correspondiente a la HU_3.

Tarea de ingeniería	
Número: TI_1_HU_4	**Historia de usuario:** Etiquetar conjunto de datos de entrenamiento.
Nombre de tarea: Crear grupos en la base de datos.	
Tipo: Desarrollo	**Puntos estimados:** 1

Fecha inicio: 3/03/2014	**Fecha fin:** 10/03/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Luego de obtener los resultados del proceso de agrupamiento, crear los grupos identificados durante el proceso en la base de datos y etiquetar cada imagen facial con el grupo al que pertenece.	

Tabla 4.19: Tarea de Ingeniería 1 correspondiente a la HU_4.

Tarea de ingeniería	
Número: TI_1_HU_5	**Historia de usuario:** Visualizar agrupamiento.
Nombre de tarea: Buscar en base de datos un grupo.	
Tipo: Desarrollo	**Puntos estimados:** 2
Fecha inicio: 10/03/2014	**Fecha fin:** 24/03/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: _____	

Tabla 4.20: Tarea de Ingeniería 1 correspondiente a la HU_5.

Tarea de ingeniería	
Número: TI_1_HU_5	**Historia de usuario:** Visualizar agrupamiento.
Nombre de tarea: Visualizar imágenes faciales del grupo.	
Tipo: Desarrollo	**Puntos estimados:** 2
Fecha inicio: 10/03/2014	**Fecha fin:** 24/03/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Al seleccionar un grupo se visualizará las imágenes faciales que pertenecen a ese grupo.	

Tabla 4.21: Tarea de Ingeniería 2 correspondiente a la HU_5.

Tarea de ingeniería	
Número: TI_1_HU_6	**Historia de usuario:** Cargar imagen facial
Nombre de tarea: Cargar desde una dirección específica la imagen facial.	
Tipo: Desarrollo	**Puntos estimados:** 1
Fecha inicio: 24/03/2014	**Fecha fin:** 31/03/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Se seleccionará desde una dirección física específica la imagen facial que se desea clasificar.	

Tabla 4.22: Tarea de Ingeniería 1 correspondiente a la HU_6.

Tarea de ingeniería	
Número: TI_1_HU_7	**Historia de usuario:** Clasificar imagen facial
Nombre de tarea: Devolver el número del grupo al que pertenece la imagen facial.	
Tipo: Desarrollo	**Puntos estimados:** 3
Fecha inicio: 31/03/2014	**Fecha fin:** 21/04/2014
Programador responsable: Gregorio Ferrer Cordova	
Descripción: Luego de cargar una imagen facial, la solución devolverá el grupo al que pertenece esa imagen.	

Tabla 4.23: Tarea de Ingeniería 1 correspondiente a la HU_7.

4.9 Anexo 9*:* Tablas asociadas al procedimiento utilizado para obtener los valores del índice DF-A.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Índices			Índices			Índices			Índices			Índices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
31	**413.6757**	**0.2114**	2.4339	**415.5456**	**0.1916**	**2.3628**	**415.4136**	0.1793	**2.3497**	**414.9507**	0.1709	2.4047	**416.3323**	0.1759	2.3678
32	402.6686	0.211	2.429	403.6208	0.1902	2.4126	403.3891	**0.1806**	2.4269	403.9055	0.1787	2.408	404.5304	**0.1829**	**2.3399**
33	392.5879	0.1896	2.4243	393.2737	0.1866	2.3817	391.9835	0.1797	2.4579	392.5764	0.1881	2.4326	394.1897	0.1813	2.3565
34	381.5032	0.1871	2.4173	382.8341	0.1854	2.3957	382.3203	0.1772	2.4333	382.8793	0.176	2.4008	383.5757	0.1724	2.386
35	371.5557	0.188	2.4197	372.565	0.1646	2.445	372.5534	0.1757	2.4353	373.5387	0.1742	**2.3741**	373.6759	0.1769	2.3812
36	363.176	0.177	2.4596	363.1668	0.1662	2.4289	363.8751	0.1787	2.398	363.3365	0.1757	2.4033	363.8829	0.1736	2.3902
37	353.9611	0.1772	2.4615	354.2902	0.1703	2.4206	355.1675	0.1777	2.4194	354.8835	0.1803	2.394	355.48	0.1813	2.3917
38	345.2663	0.1777	2.4303	346.1176	0.1808	2.4045	346.9166	0.1741	2.4049	346.2987	0.178	2.4079	347.1175	0.1816	2.3815
39	338.1654	0.1803	**2.4112**	338.2584	0.177	2.4076	338.0304	0.1798	2.4159	338.2804	**0.1904**	2.4169	338.0702	0.168	2.4303
40	329.1655	0.1858	2.4531	330.8905	0.1735	2.3694	331.5573	0.178	2.4147	331.5052	0.1798	2.4117	331.1282	0.1694	2.405
41	323.06	0.1828	2.4366	324.3089	0.1714	2.3883	324.382	0.1791	2.409	324.2814	0.1765	2.396	324.4395	0.168	2.4067

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Índices			Índices			Índices			Índices			Índices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
31	**414.51**	0.2007	2.3679	**414.3916**	0.1989	**2.3658**	**415.0547**	**0.1839**	**2.3146**	416.3606	0.1747	2.3921	**413.6359**	0.1889	2.3678
32	403.4608	**0.2019**	2.396	401.9253	0.1849	2.4219	402.6077	0.1764	2.3699	404.3421	0.1756	2.3872	402.6351	0.1796	2.3673
33	392.892	0.1962	2.3605	391.346	0.1847	2.4363	390.1088	0.1723	2.3668	393.8454	0.176	2.3538	390.472	0.1859	2.4175
34	383.7966	0.1961	**2.3355**	381.7175	0.1849	2.4059	381.5177	0.1789	2.3506	383.2876	0.1769	**2.3305**	380.4581	0.1752	2.4164
35	373.5898	0.1992	2.4088	370.6765	0.1812	2.4338	370.838	0.1741	2.4044	373.7077	0.1704	2.358	370.9712	0.1803	2.338
36	363.716	0.1927	2.4157	361.9851	0.1955	2.3914	361.8206	0.1838	2.3919	363.7988	0.1746	2.3864	362.3922	0.1792	**2.3346**
37	354.7861	0.1902	2.4387	352.1257	0.1941	2.4111	353.8011	0.1838	2.3869	354.7135	0.1717	2.4017	353.5098	0.1868	2.3482
38	346.5987	0.1918	2.4204	343.4645	0.1939	2.4041	345.7854	0.1793	2.3832	345.8262	0.176	2.405	344.8211	0.1874	2.3799
39	338.1577	0.1822	2.4171	336.2266	**0.219**	2.3985	337.1783	0.1806	2.34	337.9755	0.1763	2.4277	336.8049	**0.189**	2.3754
40	330.9046	0.1903	2.4174	329.3445	0.1745	2.4	329.7851	0.1792	2.3414	330.0808	**0.1777**	2.4439	330.4351	0.1799	2.3485
41	322.8037	0.1845	2.427	322.3151	0.1763	2.4106	322.2775	0.1813	2.3417	323.3479	0.1771	2.4346	322.4231	0.1771	2.394

Figura 4.3: Procedimiento DCT. Índices de validación sin normalizar para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Índices			Índices			Índices			Índices			Índices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
31	0.9715	0.8603	0.1879	0.9913	0.4963	0.6719	0.9899	0.2702	0.7611	0.985	0.1158	0.3867	0.9997	0.2077	0.6378
32	0.8545	0.8529	0.2212	0.8646	0.4706	0.3329	0.8621	0.2941	0.2355	0.8676	0.2592	0.3642	0.8743	0.3364	0.8278
33	0.7473	0.4596	0.2532	0.7546	0.4044	0.5432	0.7409	0.2776	0.0245	0.7472	0.432	0.1967	0.7643	0.307	0.7148
34	0.6295	0.4136	0.3009	0.6437	0.3824	0.4479	0.6382	0.2316	0.192	0.6441	0.2096	0.4132	0.6515	0.1434	0.514
35	0.5238	0.4301	0.2845	0.5345	0	0.1123	0.5344	0.204	0.1784	0.5449	0.1765	0.595	0.5463	0.2261	0.5466
36	0.4347	0.2279	0.0129	0.4346	0.0294	0.2219	0.4421	0.2592	0.4323	0.4364	0.204	0.3962	0.4422	0.1654	0.4854
37	0.3368	0.2316	0	0.3403	0.1048	0.2784	0.3496	0.2408	0.2866	0.3466	0.2886	0.4595	0.3529	0.307	0.4752
38	0.2443	0.2408	0.2124	0.2534	0.2978	0.388	0.2619	0.1746	0.3853	0.2553	0.2463	0.3649	0.264	0.3125	0.5446
39	0.1689	0.2886	0.3424	0.1699	0.2279	0.3669	0.1674	0.2794	0.3104	0.1701	0.4743	0.3036	0.1679	0.0625	0.2124
40	0.0732	0.3897	0.0572	0.0915	0.1636	0.627	0.0986	0.2463	0.3186	0.0981	0.2794	0.339	0.0941	0.0882	0.3846
41	0.0083	0.3346	0.1695	0.0216	0.125	0.4983	0.0224	0.2665	0.3574	0.0213	0.2187	0.4459	0.023	0.0625	0.373

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Índices			Índices			Índices			Índices			Índices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
31	0.9803	0.6636	0.6372	0.9791	0.6305	0.6515	0.9861	0.3548	1	1	0.1857	0.4724	0.971	0.4467	0.6378
32	0.8629	0.6857	0.4459	0.8466	0.3732	0.2696	0.8538	0.2169	0.6236	0.8723	0.2022	0.5058	0.8541	0.2757	0.6413
33	0.7506	0.5809	0.6875	0.7341	0.3695	0.1715	0.721	0.1415	0.6447	0.7607	0.2096	0.7332	0.7248	0.3915	0.2995
34	0.6539	0.579	0.8577	0.6318	0.3732	0.3785	0.6297	0.2629	0.7549	0.6485	0.2261	0.8918	0.6184	0.1949	0.307
35	0.5454	0.636	0.3587	0.5144	0.3051	0.1886	0.5161	0.1746	0.3887	0.5466	0.1066	0.7046	0.5176	0.2886	0.8407
36	0.4404	0.5165	0.3118	0.422	0.568	0.4772	0.4203	0.3529	0.4738	0.4413	0.1838	0.5112	0.4264	0.2684	0.8639
37	0.3455	0.4706	0.1552	0.3173	0.5423	0.3431	0.3351	0.3529	0.5078	0.3448	0.1305	0.4071	0.332	0.4081	0.7713
38	0.2585	0.5	0.2798	0.2252	0.5386	0.3907	0.2499	0.2702	0.533	0.2503	0.2096	0.3846	0.2396	0.4191	0.5555
39	0.1688	0.3235	0.3022	0.1483	1	0.4289	0.1584	0.2941	0.8271	0.1669	0.2151	0.2301	0.1544	0.4485	0.5861
40	0.0917	0.4724	0.3002	0.0751	0.182	0.4187	0.0798	0.2684	0.8176	0.0829	0.2408	0.1198	0.0867	0.2813	0.7692
41	0.0056	0.3658	0.2349	0.0004	0.2151	0.3465	0	0.307	0.8155	0.0114	0.2298	0.1831	0.0015	0.2298	0.4595

Figura 4.4: Procedimiento DCT. Índices de validación normalizados para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Índices			Índices			Índices			Índices			Índices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
29	**385.6363**	0.0898	2.801	**387.2365**	**0.1024**	**2.6802**	**385.8663**	0.0983	2.7356	**388.006**	0.1014	**2.6621**	**386.2314**	**0.1002**	2.7614
30	373.4084	0.0883	2.8298	375.4553	0.0996	2.7049	376.0312	0.0938	**2.6971**	375.8219	0.1018	2.6683	374.2995	0.0907	2.8124
31	363.5886	0.0932	2.7744	364.3175	0.1008	2.7159	364.9854	0.0955	2.7011	363.8293	0.1015	2.7114	363.9569	0.0892	2.7505
32	353.7855	0.0942	**2.7022**	353.6096	0.0957	2.8329	353.9061	0.0971	2.7082	352.832	0.097	2.7213	353.3638	0.0948	2.7512
33	343.0074	0.0908	2.7112	343.7036	0.093	2.8119	344.892	0.0996	2.7555	340.9464	**0.1026**	2.821	344.0581	0.0996	**2.7328**
34	334.0534	0.0882	2.7666	334.8803	0.0931	2.7969	335.3297	0.0933	2.7406	332.778	0.0935	2.6902	334.7257	0.0918	2.8085
35	324.2874	**0.096**	2.7517	325.6152	0.0985	2.7967	326.2722	0.0956	2.771	324.5178	0.0902	2.7193	327.1721	0.093	2.7492
36	315.8448	0.0922	2.7884	317.6965	0.0953	2.8122	318.809	0.0996	2.7555	316.9843	0.0944	2.7003	318.6435	0.0962	2.785
37	308.6536	0.0923	2.8111	309.7541	0.0922	2.8573	311.6589	**0.1011**	2.7343	308.8127	0.0939	2.7241	310.9186	0.0974	2.8037
38	301.1718	0.0905	2.7708	302.2636	0.0928	2.9037	304.4498	0.0982	2.7181	301.6945	0.0913	2.7591	303.1261	0.0977	2.805
39	294.2055	0.0947	2.7294	294.5991	0.0912	2.9452	297.2184	0.0999	2.7609	295.1273	0.0926	2.7378	295.4942	0.0944	2.8052

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
29	**386.3468**	0.0941	2.7418	**385.2377**	0.0947	2.7664	**385.0185**	0.0933	2.8024	**385.5262**	0.0886	2.8498	**386.2152**	0.0971	2.6982
30	375.0049	0.0957	2.786	373.9686	0.0942	2.7934	373.6027	0.098	2.8206	373.8102	0.0923	2.831	374.0547	0.0991	2.6875
31	363.7234	0.0901	2.7833	363.4281	0.0941	2.8334	362.0564	0.1004	2.8117	363.0425	0.0991	2.7934	362.488	0.0974	2.7495
32	353.1708	0.0946	2.8071	351.9906	0.0931	2.7898	351.5753	0.1025	2.8106	353.266	0.0936	2.804	351.9345	0.1005	2.7626
33	343.9315	0.0937	2.7873	342.414	0.0921	2.8319	342.0133	0.1019	2.8119	343.8574	0.0979	2.7794	342.5651	0.102	**2.6859**
34	334.9105	0.0916	2.7657	334.2773	0.0868	2.8002	333.3196	0.1016	2.8042	334.9918	0.0943	2.7758	332.0209	**0.1044**	2.7633
35	326.6265	0.0963	2.7127	325.9371	0.0848	2.8039	323.6084	**0.1038**	2.8057	325.9727	0.0955	2.8331	324.144	0.0911	2.7637
36	317.0888	0.0903	2.7354	318.2327	0.0902	2.7952	315.7753	0.0984	2.8349	317.7552	**0.0992**	2.7299	314.9651	0.0911	2.805
37	309.0819	0.0991	**2.7084**	310.2293	**0.0956**	2.7753	308.6143	0.0977	**2.7869**	309.445	0.0941	2.7404	307.4568	0.0981	2.8062
38	302.4273	**0.1039**	2.7108	302.8167	0.0955	2.7817	300.3489	0.096	2.8922	302.5444	0.0946	**2.7291**	300.729	0.0876	2.8289
39	295.5551	0.0926	2.7167	295.97	0.0934	**2.7661**	294.027	0.1013	2.8098	295.2184	0.0979	2.7771	294.4043	0.0966	2.7497

Figura 4.5: Procedimiento GWT. Índices de validación sin normalizar para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
29	0.9748	0.2551	0.5094	0.9918	0.898	0.9361	0.9772	0.6888	0.7404	1	0.8469	1	0.9811	0.7857	0.6492
30	0.8447	0.1786	0.4076	0.8665	0.7551	0.8488	0.8726	0.4592	0.8764	0.8704	0.8673	0.9781	0.8542	0.301	0.4691
31	0.7402	0.4286	0.6033	0.7479	0.8163	0.81	0.755	0.5459	0.8622	0.7427	0.852	0.8259	0.7441	0.2245	0.6877
32	0.6359	0.4796	0.8584	0.634	0.5561	0.3967	0.6372	0.6276	0.8372	0.6257	0.6224	0.7909	0.6314	0.5102	0.6853
33	0.5212	0.3061	0.8266	0.5286	0.4184	0.4709	0.5412	0.7551	0.6701	0.4993	0.9082	0.4387	0.5324	0.7551	0.7503
34	0.4259	0.1735	0.6309	0.4347	0.4235	0.5238	0.4395	0.4337	0.7227	0.4123	0.4439	0.9007	0.4331	0.3571	0.4829
35	0.322	0.5714	0.6835	0.3361	0.699	0.5245	0.3431	0.551	0.6153	0.3244	0.2755	0.798	0.3527	0.4184	0.6923
36	0.2322	0.3776	0.5539	0.2519	0.5357	0.4698	0.2637	0.7551	0.6701	0.2443	0.4898	0.8651	0.2619	0.5816	0.5659
37	0.1556	0.3827	0.4737	0.1673	0.3776	0.3105	0.1876	0.8316	0.745	0.1573	0.4643	0.781	0.1797	0.6429	0.4998
38	0.076	0.2908	0.616	0.0876	0.4082	0.1466	0.1109	0.6837	0.8022	0.0816	0.3316	0.6574	0.0968	0.6582	0.4952
39	0.0019	0.5051	0.7623	0.0061	0.3265	0	0.034	0.7704	0.651	0.0117	0.398	0.7326	0.0156	0.4898	0.4945

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
29	0.9823	0.4745	0.7185	0.9705	0.5051	0.6316	0.9682	0.4337	0.5044	0.9736	0.1939	0.337	0.9809	0.6276	0.8725
30	0.8617	0.5561	0.5623	0.8506	0.4796	0.5362	0.8467	0.6735	0.4401	0.8489	0.3827	0.4034	0.8515	0.7296	0.9103
31	0.7416	0.2704	0.5719	0.7385	0.4745	0.3949	0.7239	0.7959	0.4716	0.7344	0.7296	0.5362	0.7285	0.6429	0.6913
32	0.6293	0.5	0.4878	0.6168	0.4235	0.5489	0.6124	0.9031	0.4755	0.6303	0.449	0.4988	0.6162	0.801	0.645
33	0.531	0.4541	0.5578	0.5149	0.3724	0.4002	0.5106	0.8724	0.4709	0.5302	0.6684	0.5857	0.5165	0.8776	0.9159
34	0.435	0.3469	0.6341	0.4283	0.102	0.5122	0.4181	0.8571	0.4981	0.4359	0.4847	0.5984	0.4043	1	0.6425
35	0.3469	0.5867	0.8213	0.3395	0	0.4991	0.3148	0.9694	0.4928	0.3399	0.5459	0.396	0.3205	0.3214	0.6411
36	0.2454	0.2806	0.7411	0.2576	0.2755	0.5298	0.2314	0.6939	0.3896	0.2525	0.7347	0.7605	0.2228	0.3214	0.4952
37	0.1602	0.7296	0.8365	0.1724	0.551	0.6001	0.1552	0.6582	0.5592	0.1641	0.4745	0.7234	0.1429	0.6786	0.491
38	0.0894	0.9745	0.828	0.0935	0.5459	0.5775	0.0673	0.5714	0.1872	0.0906	0.5	0.7633	0.0713	0.1429	0.4108
39	0.0163	0.398	0.8071	0.0207	0.4388	0.6326	0	0.8418	0.4783	0.0127	0.6684	0.5938	0.004	0.602	0.6906

Figura 4.6: Procedimiento GWT. Índices de validación normalizados para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
30	**622.2329**	0.2065	2.0317	**635.8398**	**0.2401**	**1.7307**	**635.1174**	0.2643	1.7869	**640.522**	0.2322	1.8744	**634.1406**	**0.2308**	1.829
31	605.4606	0.217	1.984	616.6916	0.2353	1.7752	620.6964	0.2602	1.8105	622.7135	0.2392	**1.8563**	613.9975	0.1848	**1.8265**
32	587.4587	0.2012	2.0383	599.2916	0.2353	1.7809	606.1843	**0.2869**	**1.7828**	604.6295	0.2362	1.8933	597.8099	0.1795	1.8652
33	574.542	0.2177	1.9883	581.2145	0.2071	1.8356	588.7368	0.2502	1.8221	589.7645	0.2353	1.8796	582.0091	0.1795	1.8985
34	560.2507	0.1965	1.9902	566.5793	0.207	1.8457	578.4459	0.2503	1.8311	575.7133	0.2342	1.87	566.2087	0.1585	1.9283
35	544.8823	0.2153	1.976	553.2722	0.207	1.8435	564.9887	0.2504	1.8554	561.1681	**0.2399**	1.882	554.3539	0.1586	1.9579
36	529.5226	0.2158	1.99	548.8082	0.2093	1.821	550.3385	0.2388	1.8927	547.2703	0.2087	1.8858	540.2022	0.1582	1.9376
37	524.3126	0.2326	**1.944**	536.7564	0.2081	1.803	536.2963	0.2388	1.8951	534.2973	0.2384	1.8724	529.3681	0.1622	1.9183
38	511.7924	**0.2335**	1.9776	520.7357	0.2072	1.8076	523.168	0.2406	1.9334	522.6049	0.2368	1.8725	518.1561	0.1613	1.9284
39	496.6738	0.1934	2.0084	508.954	0.2072	1.8299	512.2449	0.2366	1.9326	511.5737	0.2384	1.8844	507.3413	0.1614	1.9569
40	487.5948	0.2023	2.0371	501.3989	0.1891	1.8579	502.9952	0.2362	1.8933	500.9922	0.2373	1.88	497.7545	0.1615	1.9432

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
30	**634.3794**	0.2274	**1.8066**	**629.5486**	**0.2324**	1.9164	**639.924**	**0.2607**	**1.8002**	**635.8788**	**0.26**	1.9492	**632.517**	0.2125	1.8699
31	616.3227	**0.2278**	1.8204	616.9205	0.2313	**1.9031**	614.7956	0.2455	1.8483	619.9023	0.2479	1.9723	620.1915	0.2131	1.8557
32	591.5894	0.1858	1.8664	600.2343	0.2138	1.9236	599.1247	0.2354	1.8555	601.9671	0.2438	1.9114	607.7403	0.2309	1.8424
33	576.9132	0.1956	1.8505	582.1153	0.2101	1.9771	582.6195	0.2044	1.879	585.8557	0.2026	1.9832	590.8802	0.2178	1.8769
34	563.0146	0.1942	1.8415	566.2507	0.2109	1.9848	566.1896	0.2037	1.8797	569.7477	0.2006	1.9877	578.513	**0.2391**	1.8457
35	548.6068	0.1955	1.8169	554.8261	0.2251	1.9926	560.9333	0.2084	1.8491	552.3618	0.1973	1.9647	564.9921	0.2253	**1.8373**
36	533.9633	0.1622	1.8879	540.0623	0.2236	1.986	547.1718	0.2082	1.8471	540.9204	0.2092	1.9323	550.5198	0.2246	1.8723
37	527.8959	0.161	1.8305	526.5877	0.224	2.0123	533.5317	0.2084	1.8539	530.5187	0.2099	1.9118	535.9212	0.2246	1.8993
38	515.4874	0.1653	1.8353	519.3302	0.223	1.9421	521.6971	0.2066	1.8803	526.9955	0.2112	**1.8852**	524.9382	0.219	1.8824
39	504.3995	0.1693	1.8378	507.5287	0.2053	1.9658	510.5555	0.2068	1.9092	514.1888	0.2132	1.8916	512.1342	0.2135	1.8695
40	491.0296	0.165	1.8901	495.5978	0.2027	1.9576	498.9078	0.2081	1.9277	502.4954	0.2116	1.89	502.3829	0.2124	1.8614

Figura 4.7: Procedimiento IC. Índices de validación sin normalizar para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
30	0.8804	0.3753	0.0215	0.9694	0.6364	1	0.9647	0.8244	0.8173	1	0.575	0.5328	0.9583	0.5641	0.6804
31	0.7707	0.4569	0.1765	0.8442	0.5991	0.8553	0.8704	0.7925	0.7406	0.8835	0.6294	0.5917	0.8266	0.2067	0.6886
32	0.653	0.3341	0	0.7304	0.5991	0.8368	0.7755	1	0.8306	0.7653	0.6061	0.4714	0.7207	0.1655	0.5627
33	0.5686	0.4623	0.1625	0.6122	0.38	0.659	0.6614	0.7148	0.7029	0.6681	0.5991	0.5159	0.6174	0.1655	0.4545
34	0.4751	0.2976	0.1564	0.5165	0.3792	0.6261	0.5941	0.7156	0.6736	0.5762	0.5905	0.5471	0.5141	0.0023	0.3576
35	0.3746	0.4437	0.2025	0.4295	0.3792	0.6333	0.5061	0.7164	0.5946	0.4811	0.6348	0.5081	0.4365	0.0031	0.2614
36	0.2742	0.4476	0.157	0.4003	0.397	0.7064	0.4103	0.6263	0.4733	0.3902	0.3924	0.4958	0.344	0	0.3274
37	0.2401	0.5781	0.3066	0.3215	0.3877	0.765	0.3185	0.6263	0.4655	0.3054	0.6232	0.5393	0.2732	0.0311	0.3901
38	0.1582	0.5851	0.1973	0.2167	0.3807	0.75	0.2326	0.6402	0.341	0.2289	0.6107	0.539	0.1998	0.0241	0.3573
39	0.0594	0.2735	0.0972	0.1397	0.3807	0.6775	0.1612	0.6092	0.3436	0.1568	0.6232	0.5003	0.1291	0.0249	0.2646
40	0	0.3427	0.0039	0.0903	0.2401	0.5865	0.1007	0.6061	0.4714	0.0876	0.6146	0.5146	0.0664	0.0256	0.3092

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
30	0.9598	0.5377	0.7533	0.9282	0.5765	0.3963	0.9961	0.7964	0.7741	0.9696	0.791	0.2897	0.9477	0.4219	0.5475
31	0.8418	0.5408	0.7084	0.8457	0.568	0.4395	0.8318	0.6783	0.6177	0.8652	0.697	0.2146	0.8671	0.4266	0.5936
32	0.68	0.2145	0.5588	0.7366	0.432	0.3729	0.7293	0.5998	0.5943	0.7479	0.6651	0.4125	0.7856	0.5649	0.6369
33	0.5841	0.2906	0.6105	0.6181	0.4033	0.199	0.6214	0.359	0.5179	0.6425	0.345	0.1791	0.6754	0.4631	0.5247
34	0.4932	0.2797	0.6398	0.5143	0.4095	0.1739	0.5139	0.3535	0.5156	0.5372	0.3294	0.1645	0.5945	0.6286	0.6261
35	0.399	0.2898	0.7198	0.4396	0.5198	0.1486	0.4796	0.3901	0.6151	0.4235	0.3038	0.2393	0.5061	0.5214	0.6534
36	0.3032	0.0311	0.4889	0.3431	0.5082	0.17	0.3896	0.3885	0.6216	0.3487	0.3963	0.3446	0.4115	0.5159	0.5397
37	0.2635	0.0218	0.6756	0.255	0.5113	0.0845	0.3004	0.3901	0.5995	0.2807	0.4017	0.4112	0.316	0.5159	0.4519
38	0.1824	0.0552	0.6599	0.2075	0.5035	0.3127	0.223	0.3761	0.5137	0.2576	0.4118	0.4977	0.2442	0.4724	0.5068
39	0.1099	0.0862	0.6518	0.1303	0.366	0.2357	0.1501	0.3776	0.4197	0.1739	0.4274	0.4769	0.1605	0.4297	0.5488
40	0.0225	0.0528	0.4818	0.0523	0.3458	0.2624	0.074	0.3877	0.3596	0.0974	0.4149	0.4821	0.0967	0.4211	0.5751

Figura 4.8: Procedimiento IC. Índices de validación normalizados para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
50	**184.9271**	**0.1623**	3.2198	**184.9115**	0.1512	3.2615	**183.631**	**0.1594**	**3.0805**	**184.4251**	0.1585	3.3074	**184.9022**	0.1564	3.2632
51	181.5733	0.1559	3.2236	181.4584	0.1525	3.2154	180.2952	0.1559	3.1161	181.1028	0.1527	3.2853	181.5211	0.153	3.2778
52	178.5216	0.1551	3.2392	178.1447	0.1547	3.2255	177.1266	0.1518	3.1076	178.2752	0.1596	3.2853	178.4784	0.1506	3.2876
53	175.4057	0.1576	3.2373	175.2667	0.1574	3.2265	174.2195	0.153	3.1142	174.8926	0.1644	3.304	175.4937	**0.1576**	**3.2606**
54	172.1624	0.1608	3.2215	172.3745	**0.1625**	3.223	171.2627	0.1522	3.1162	171.5574	0.1621	3.2987	172.329	0.147	3.2944
55	169.422	0.1437	3.1934	168.949	0.1543	3.2075	167.9365	0.1513	3.0941	168.592	0.1583	3.303	169.5756	0.1506	3.2715
56	166.4489	0.1344	3.1962	165.8275	0.1525	3.2339	165.5967	0.1525	3.1062	165.8037	0.1645	3.2892	166.7904	0.1503	3.2792
57	163.9395	0.1395	3.1991	163.7563	0.1589	**3.1798**	162.8889	0.1483	3.1309	163.9532	0.1601	3.3029	164.0615	0.134	3.3267
58	161.447	0.1427	3.1943	161.1133	0.1502	3.2118	160.4153	0.1455	3.1323	160.8994	0.1582	3.3155	161.5886	0.1558	3.2662
59	158.6624	0.146	**3.189**	158.5916	0.148	3.2305	158.1042	0.1584	3.1236	158.6785	**0.1688**	**3.2794**	159.0466	0.1513	3.2963
60	156.4362	0.1385	3.2334	156.2309	0.1517	3.2123	155.6795	0.1566	3.1055	156.0529	0.161	3.3123	156.4727	0.1469	3.278

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
50	**184.2811**	0.1556	3.1686	**184.9653**	0.207	3.2208	**184.3985**	0.1516	**3.1604**	**183.6539**	0.2017	**3.1189**	**183.8555**	0.1501	3.2605
51	180.5705	**0.1569**	3.2077	181.5845	0.2056	3.1687	180.9568	0.1561	3.1779	180.2367	0.1716	3.1753	180.8699	0.1445	3.2811
52	177.5044	0.1487	3.1992	178.3009	**0.2075**	**3.163**	177.7281	0.1526	3.1883	176.9115	**0.2058**	3.1868	177.9397	0.1489	3.249
53	174.5088	0.1538	3.1778	175.2179	0.2027	3.1782	174.4977	**0.161**	3.2007	174.1178	0.1991	3.1757	174.7087	0.1517	3.2529
54	171.4355	0.1493	3.2242	172.2308	0.2058	3.1909	171.7275	0.1488	3.2197	171.0185	0.1482	3.167	171.8372	**0.152**	3.2919
55	168.9735	0.1541	**3.1587**	169.1903	0.1995	3.1852	168.8349	0.15	3.2571	168.3977	0.1476	3.1347	169.3702	0.1477	3.2512
56	165.7811	0.1542	3.2074	166.3647	0.1577	3.1866	166.055	0.1515	3.2095	165.6215	0.148	3.1284	166.5509	0.148	3.2896
57	163.5928	0.1545	3.1875	163.7848	0.1576	3.193	163.2494	0.1401	3.2217	163.0632	0.1537	3.1346	163.8833	0.1478	3.2658
58	160.7797	0.149	3.2147	161.1264	0.1569	3.1895	160.8019	0.1418	3.2397	160.5271	0.1513	3.145	161.2768	0.146	3.2439
59	158.1266	0.1522	3.2393	158.4356	0.1467	3.2022	158.4958	0.1536	3.2258	158.4269	0.1564	3.1463	158.6373	0.1453	**3.2367**
60	155.7554	0.1467	3.2379	156.0797	0.1475	3.2161	156.0208	0.15	3.1967	155.7282	0.1561	3.1289	156.3297	0.1438	3.2717

Figura 4.9: Procedimiento IR. Índices de validación sin normalizar para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
50	0.9987	0.385	0.4342	0.9982	0.234	0.2648	0.9544	0.3456	1	0.9816	0.3333	0.0784	0.9978	0.3048	0.2579
51	0.8842	0.298	0.4188	0.8803	0.2517	0.4521	0.8405	0.298	0.8554	0.8681	0.2544	0.1682	0.8824	0.2585	0.1986
52	0.78	0.2871	0.3554	0.7671	0.2816	0.411	0.7323	0.2422	0.8899	0.7716	0.3483	0.1682	0.7785	0.2259	0.1588
53	0.6736	0.3211	0.3631	0.6688	0.3184	0.407	0.6331	0.2585	0.8631	0.6561	0.4136	0.0922	0.6766	0.3211	0.2685
54	0.5628	0.3646	0.4273	0.5701	0.3878	0.4212	0.5321	0.2476	0.855	0.5422	0.3823	0.1137	0.5685	0.1769	0.1312
55	0.4693	0.132	0.5414	0.4531	0.2762	0.4842	0.4185	0.2354	0.9448	0.4409	0.3306	0.0963	0.4745	0.2259	0.2242
56	0.3677	0.0054	0.5301	0.3465	0.2517	0.3769	0.3386	0.2517	0.8956	0.3457	0.415	0.1523	0.3794	0.2218	0.1929
57	0.282	0.0748	0.5183	0.2758	0.3388	0.5967	0.2462	0.1946	0.7953	0.2825	0.3551	0.0967	0.2862	0	0
58	0.1969	0.1184	0.5378	0.1855	0.2204	0.4667	0.1617	0.1565	0.7896	0.1782	0.3293	0.0455	0.2018	0.2966	0.2457
59	0.1019	0.1633	0.5593	0.0994	0.1905	0.3907	0.0828	0.332	0.8249	0.1024	0.4735	0.1921	0.115	0.2354	0.1235
60	0.0258	0.0612	0.379	0.0188	0.2408	0.4647	0	0.3075	0.8985	0.0128	0.3673	0.0585	0.0271	0.1755	0.1978

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
50	0.9766	0.2939	0.6422	1	0.9932	0.4301	0.9806	0.2395	0.6755	0.9552	0.9211	0.844	0.9621	0.219	0.2689
51	0.8499	0.3116	0.4833	0.8846	0.9741	0.6418	0.8631	0.3007	0.6044	0.8385	0.5116	0.6149	0.8602	0.1429	0.1852
52	0.7452	0.2	0.5179	0.7724	1	0.6649	0.7529	0.2531	0.5621	0.725	0.9769	0.5682	0.7601	0.2027	0.3156
53	0.6429	0.2694	0.6048	0.6672	0.9347	0.6032	0.6426	0.3673	0.5118	0.6296	0.8857	0.6133	0.6498	0.2408	0.2998
54	0.538	0.2082	0.4163	0.5652	0.9769	0.5516	0.548	0.2014	0.4346	0.5238	0.1932	0.6487	0.5517	0.2449	0.1413
55	0.4539	0.2735	0.6824	0.4613	0.8912	0.5747	0.4492	0.2177	0.2827	0.4343	0.185	0.7799	0.4675	0.1864	0.3067
56	0.3449	0.2748	0.4846	0.3649	0.3224	0.569	0.3543	0.2381	0.476	0.3395	0.1905	0.8054	0.3712	0.1905	0.1507
57	0.2702	0.2789	0.5654	0.2768	0.3211	0.5431	0.2585	0.083	0.4265	0.2521	0.268	0.7803	0.2801	0.1878	0.2474
58	0.1742	0.2041	0.4549	0.186	0.3116	0.5573	0.1749	0.1061	0.3534	0.1655	0.2354	0.738	0.1911	0.1633	0.3363
59	0.0836	0.2476	0.355	0.0941	0.1728	0.5057	0.0962	0.2667	0.4098	0.0938	0.3048	0.7327	0.101	0.1537	0.3656
60	0.0026	0.1728	0.3607	0.0137	0.1837	0.4492	0.0117	0.2177	0.528	0.0017	0.3007	0.8034	0.0222	0.1333	0.2234

Figura 4.10: Procedimiento IR. Índices de validación normalizados para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
25	**270.4063**	0.1165	3.6772	**271.9073**	0.1148	3.986	**273.5274**	0.1143	4.1915	**273.0059**	0.1185	4.198	**275.6922**	0.111	3.8776
26	260.7199	0.1143	3.753	263.7943	0.1159	3.9905	262.5978	0.114	4.0718	261.9543	0.1157	3.8993	264.1698	0.1106	3.8231
27	251.6253	0.1175	3.7302	253.4553	0.116	3.9473	254.1802	0.111	4.0108	252.38	0.1176	3.9472	253.9999	0.1109	3.7207
28	243.7163	**0.12**	**3.6693**	244.3344	**0.1226**	3.9704	244.5732	0.1136	3.9903	244.4574	0.1172	3.9595	245.9137	0.1115	3.765
29	236.9955	0.1185	3.7986	236.7898	0.1223	3.898	237.6056	0.1153	3.9274	235.7649	0.1165	3.8948	238.526	0.113	3.8075
30	229.2855	0.1138	3.8367	228.5567	0.1218	3.9271	229.3624	0.1162	4.0136	228.4426	0.1131	3.9626	230.2953	0.113	3.7154
31	222.1098	0.1129	3.8619	221.5246	0.1114	**3.809**	221.0616	0.1164	4.0538	220.9576	0.113	3.8679	222.5611	0.1143	3.6587
32	215.372	0.116	3.8318	215.1035	0.1126	3.8853	216.0935	**0.1176**	3.9984	215.0727	0.1095	3.9305	216.076	0.1154	3.6874
33	209.4092	0.1182	3.8771	208.9678	0.1206	3.8979	210.2338	0.1086	4.0537	209.3974	0.1184	3.8386	209.3371	0.1154	3.5997
34	203.79	0.1181	3.9137	202.9958	0.1188	3.915	203.8706	0.1097	3.9625	203.1441	**0.1187**	3.7474	202.9782	0.115	3.5333
35	198.3054	0.1153	4.0194	197.5517	0.118	3.9022	198.3037	0.1084	**3.8644**	197.208	0.1183	**3.706**	198.158	**0.116**	**3.5201**

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
25	**274.9123**	0.1102	4.0862	**276.1486**	0.1163	4.3039	**272.9541**	**0.1194**	3.8949	**271.4777**	**0.1181**	3.7056	**276.0519**	0.1206	4.4664
26	264.1899	0.1093	4.0408	266.7407	0.116	4.2766	262.2795	0.1136	3.8147	261.9325	0.1148	3.774	266.0198	0.1203	4.4291
27	254.7234	0.1105	4.0396	256.99	0.1124	4.2961	252.2635	0.1136	3.7027	253.4866	0.1175	3.7726	256.6736	0.1188	4.3849
28	246.8825	0.1095	4.0615	248.0722	0.1128	4.3251	242.7153	0.1113	3.6537	243.5281	0.1139	3.8193	247.4988	0.1172	4.3679
29	239.3788	0.1091	4.028	240.354	0.1142	4.3249	234.4212	0.1111	**3.6152**	237.715	0.1138	3.818	239.4225	0.1195	4.4158
30	231.1812	0.1084	4.0592	232.957	0.1144	4.3064	227.1908	0.1147	3.7183	229.2412	0.116	3.8435	230.9286	**0.1252**	4.3475
31	224.7857	**0.1172**	**4.0146**	225.381	0.0949	4.3509	220.7871	0.1128	3.7421	223.1744	0.1124	3.7342	225.252	0.1115	**4.2951**
32	217.8391	0.1136	4.0738	218.5352	0.0962	4.3034	215.4202	0.1078	3.6967	216.4378	0.111	3.7821	218.2059	0.112	4.3278
33	212.7121	0.1163	4.114	209.6972	0.1184	3.814	209.4891	0.1116	3.8537	210.6608	0.1155	3.7534	211.9671	0.1123	4.3768
34	206.0518	0.1114	4.174	204.2746	**0.1192**	3.802	202.4293	0.109	3.7443	204.9537	0.1144	3.7151	206.6626	0.1107	4.3863
35	200.011	0.1146	4.1972	198.2832	0.119	**3.7184**	197.461	0.1148	3.7197	198.9217	0.1148	**3.638**	200.7715	0.1103	4.4022

Figura 4.11: Procedimiento HSV. Índices de validación sin normalizar para cada corrida del K-means.

	K-medias 01			K-medias 02			K-medias 03			K-medias 04			K-medias 05		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
25	0.9273	0.7129	0.834	0.9463	0.6568	0.5077	0.9668	0.6403	0.2905	0.9602	0.7789	0.2836	0.9942	0.5314	0.6222
26	0.8046	0.6403	0.7539	0.8435	0.6931	0.5029	0.8283	0.6304	0.417	0.8202	0.6865	0.5993	0.8483	0.5182	0.6798
27	0.6893	0.7459	0.778	0.7125	0.6964	0.5486	0.7217	0.5314	0.4815	0.6989	0.7492	0.5487	0.7194	0.5281	0.788
28	0.5892	0.8284	0.8423	0.597	0.9142	0.5241	0.6	0.6172	0.5031	0.5985	0.736	0.5357	0.617	0.5479	0.7412
29	0.504	0.7789	0.7057	0.5014	0.9043	0.6007	0.5117	0.6733	0.5696	0.4884	0.7129	0.604	0.5234	0.5974	0.6963
30	0.4063	0.6238	0.6654	0.3971	0.8878	0.5699	0.4073	0.703	0.4785	0.3957	0.6007	0.5324	0.4191	0.5974	0.7936
31	0.3154	0.5941	0.6388	0.308	0.5446	0.6947	0.3022	0.7096	0.436	0.3009	0.5974	0.6325	0.3212	0.6403	0.8535
32	0.2301	0.6964	0.6706	0.2267	0.5842	0.6141	0.2392	0.7492	0.4946	0.2263	0.4818	0.5663	0.239	0.6766	0.8232
33	0.1546	0.769	0.6227	0.149	0.8482	0.6008	0.165	0.4521	0.4361	0.1544	0.7756	0.6634	0.1536	0.6766	0.9159
34	0.0834	0.7657	0.5841	0.0733	0.7888	0.5827	0.0844	0.4884	0.5325	0.0752	0.7855	0.7598	0.0731	0.6634	0.9861
35	0.0139	0.6733	0.4724	0.0044	0.7624	0.5962	0.0139	0.4455	0.6362	0	0.7723	0.8036	0.012	0.6964	1

	K-medias 06			K-medias 07			K-medias 08			K-medias 09			K-medias 10		
	Indices			Indices			Indices			Indices			Indices		
K	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB	CH	Dunn	DB
25	0.9843	0.505	0.4018	1	0.7063	0.1717	0.9595	0.8086	0.6039	0.9408	0.7657	0.804	0.9988	0.8482	0
26	0.8485	0.4752	0.4498	0.8808	0.6964	0.2006	0.8243	0.6172	0.6887	0.8199	0.6568	0.7317	0.8717	0.8383	0.0394
27	0.7286	0.5149	0.451	0.7573	0.5776	0.18	0.6974	0.6172	0.807	0.7129	0.7459	0.7332	0.7533	0.7888	0.0861
28	0.6293	0.4818	0.4279	0.6443	0.5908	0.1493	0.5765	0.5413	0.8588	0.5868	0.6271	0.6838	0.6371	0.736	0.1041
29	0.5342	0.4686	0.4633	0.5466	0.637	0.1495	0.4714	0.5347	0.8995	0.5131	0.6238	0.6852	0.5348	0.8119	0.0535
30	0.4304	0.4455	0.4303	0.4529	0.6436	0.1691	0.3798	0.6535	0.7906	0.4058	0.6964	0.6582	0.4272	1	0.1256
31	0.3493	0.736	0.4774	0.3569	0	0.1221	0.2987	0.5908	0.7654	0.3289	0.5776	0.7738	0.3553	0.5479	0.181
32	0.2613	0.6172	0.4149	0.2702	0.0429	0.1722	0.2307	0.4257	0.8134	0.2436	0.5314	0.7231	0.266	0.5644	0.1465
33	0.1964	0.7063	0.3724	0.1582	0.7756	0.6894	0.1556	0.5512	0.6475	0.1704	0.6799	0.7535	0.187	0.5743	0.0947
34	0.112	0.5446	0.309	0.0895	0.802	0.7021	0.0661	0.4653	0.7631	0.0981	0.6436	0.7939	0.1198	0.5215	0.0846
35	0.0355	0.6502	0.2845	0.0136	0.7954	0.7904	0.0032	0.6568	0.7891	0.0217	0.6568	0.8754	0.0451	0.5083	0.0678

Figura 4.12: Procedimiento HSV. Índices de validación normalizados para cada corrida del K-means.

4.10 **Anexo 10:** Resultados de la Validación del agrupamiento.

K	DF-A
50	**0.997733**
51	0.9047
52	0.889967
53	0.8248
54	0.800667
55	0.770167
56	0.563333
57	0.478867
58	0.440233
59	0.471133
60	0.430967

(a)

Estadísticas	Valor
Probabilidad	1.18E-06
Utilidad	-11.4072

(b)

	CH	Dunn	DB	Total
50	**10**	**2**	**3**	**15**
51	0	1	0	1
52	0	2	1	3
53	0	2	1	3
54	0	2	0	2
55	0	0	1	1
56	0	0	0	0
57	0	0	1	1
58	0	0	0	0
59	0	1	3	4
60	0	0	0	0

(c)

Figura 4.13: Validación del agrupamiento para el procedimiento Imagen Reescalada: tabla (a) muestra los grupos y el CVI DF-A, tabla (b) representa las estadísticas calculadas a partir del modelo de regresión, tabla (c) muestra los resultados del esquema basado en votos.

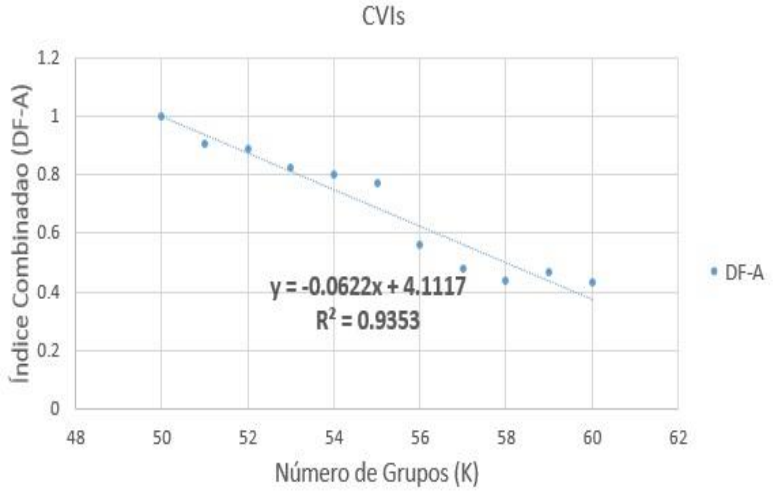


Figura 4.14: Regresión lineal asociada al procedimiento Imagen Reescalada.

K	DF-A
25	**0.894067**
26	0.824333
27	0.784367
28	0.805767
29	0.783467
30	0.748833
31	0.6488
32	0.6142
33	0.6535
34	0.635967
35	0.6135

(a)

Estadísticas	Valor
Probabilidad	1.26E-05
Utilidad	-8.5808

(b)

	CH	Dunn	DB	Total
25	**10**	**2**	**0**	**12**
26	0	0	0	0
27	0	0	0	0
28	0	2	1	3
29	0	0	1	1
30	0	1	0	1
31	0	1	3	4
32	0	1	0	1
33	0	0	0	0
34	0	2	0	2
35	0	1	5	6

(c)

Figura 4.15: Validación del agrupamiento para el procedimiento HSV: tabla (a) muestra los grupos y el CVI DF-A, tabla (b) representa las estadísticas calculadas a partir del modelo de regresión, tabla (c) muestra los resultados del esquema basado en votos.

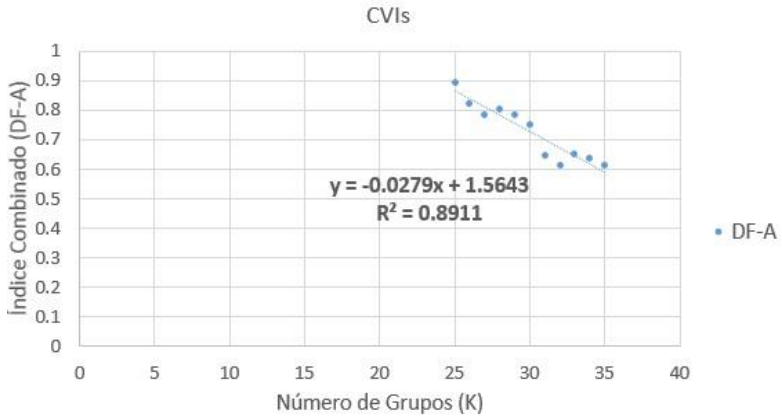


Figura 4.16: Regresión lineal asociada al procedimiento HSV.

4.11 **Anexo 11:** Código del algoritmo para entrenar el modelo.

```
public void EntrenamientoModelo() {                                                    //1
    List<Grupo> grupos = _datos.ObtenerGrupos();                                       //1
    int cantGrupos = grupos.Count;                                                     //1
    int dimensionCentroides = grupos[0].Centroide.Split(' ').Length;                   //1
    float[,] conjEntre = new float[cantGrupos,dimensionCentroides];                    //1
    float[,] clases = new float[cantGrupos, 1];                                        //1
    for (int i = 0; i < cantGrupos; i++)                                               //2
    {
        string[] aux = grupos[i].Centroide.Split(' ');                                 //3
        for (int j = 0; j < dimensionCentroides; j++)                                  //4
            conjEntre[i,j] = float.Parse(aux[j], CultureInfo.InvariantCulture.NumberFormat);  //5
        clases[i,0] = i;                                                               //6
    }
    clasificador = new Clasificador(conjEntre, clases, cantGrupos);                   //7
    _clasificador.Entrenamiento();                                                     //8
}                                                                                      //9
```

4.12 **Anexo 12:** Casos de Prueba de funcionalidad.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 2	**HU:** Cargar conjunto de datos de entrenamiento.
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para verificar que la solución carga correctamente el conjunto de datos de entrenamiento desde un archivo especificado.	
Condiciones de ejecución: El conjunto de datos de entrenamiento debe estar físicamente en la dirección que se especifique.	
Entrada/Pasos de ejecución: • Seleccionar las imágenes faciales desde una dirección física especificada. • Guardar en una carpeta. • Guardar imágenes faciales en la base de datos. • Asignar cada imagen facial a un grupo inicial.	
Resultado esperado: Se muestra una notificación de que el proceso de carga del conjunto de datos de entrenamiento se produjo correctamente.	
Evaluación de la prueba: Satisfactoria	

Tabla 4.24: Caso de Prueba de la funcionalidad: Cargar conjunto de datos de entrenamiento.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 3	**HU:** Procesar imagen facial.
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para verificar que la solución procesa las imágenes faciales correctamente.	
Condiciones de ejecución: Debe estar cargado el conjunto de datos de entrenamiento.	
Entrada/Pasos de ejecución: • Se detecta el rostro en la imagen, utilizando la biblioteca OpenCV. • Se normaliza la imagen. • Se procede a extraer los vectores de características utilizando los procedimientos Imagen Icono, Imagen Reescalada, DCT, HSV y GWT. • Se almacenan los vectores de características extraídos en la base de datos.	
Resultado esperado: Son añadidos los vectores de características asociados a cada imagen en la base de datos.	
Evaluación de la prueba: Satisfactoria	

Tabla 4.25: Caso de Prueba de la funcionalidad: Procesar imagen facial.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 4	**HU:** Generar fichero ARFF.
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para verificar que la solución genera correctamente el fichero ARFF.	
Condiciones de ejecución: Para generar el fichero ARFF de las características principales deben existir más datos que la dimensionalidad de los vectores característicos.	
Entrada/Pasos de ejecución: • Cargar vectores característicos. • Generar fichero ARFF.	
Resultado esperado: Se genera un fichero ARFFF que puede ser interpretado por el Weka.	
Evaluación de la prueba: Satisfactoria	

Tabla 4.26: Caso de Prueba de la funcionalidad: Generar fichero ARFF.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 2	**HU:** ·Etiquetar conjunto de datos de entrenamiento
Responsable: Anays Gómez García	

Descripción: Prueba de funcionalidad para verificar que la solución etiqueta los datos de entrenamiento correctamente.
Condiciones de ejecución: Deben generarse los ficheros ARFF y estar realizado el proceso de agrupamiento.
Entrada/Pasos de ejecución: • Cargar modelo ARFF. • Interpretar el modelo. • Etiquetar la base de datos.
Resultado esperado: Se asigna a cada imagen facial un grupo, almacenándose esta información en la base de datos.
Evaluación de la prueba: Satisfactoria

Tabla 4.27: Caso de Prueba de la funcionalidad: Etiquetar conjunto de datos de entrenamiento.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 2	**HU:** Visualizar agrupamiento
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para comprobar que la solución visualice los grupos correctamente.	
Condiciones de ejecución: Debe estar realizado el proceso de agrupamiento.	
Entrada/Pasos de ejecución: • Seleccionar grupo a visualizar. • Buscar en la base de datos dicho grupo. • Visualizar grupo.	
Resultado esperado: Son mostrados las imágenes faciales asociados a un grupo seleccionado.	
Evaluación de la prueba: Satisfactoria	

Tabla 4.28: Caso de Prueba de la funcionalidad: Visualizar agrupamiento.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 2	**HU:** Cargar imagen facial
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para comprobar que la solución carga la imagen facial correctamente.	
Condiciones de ejecución:	
Entrada/Pasos de ejecución: • Cargar imagen facial desde un archivo especificado.	
Resultado esperado:	
Evaluación de la prueba: Satisfactoria	

Tabla 4.29: Caso de Prueba de la funcionalidad: Cargar imagen facial.

Caso de Prueba de funcionalidad	
Código Caso de Prueba: 2	**HU:** Clasificar imagen facial
Responsable: Anays Gómez García	
Descripción: Prueba de funcionalidad para verificar que la solución clasifica una imagen facial.	
Condiciones de ejecución: Debe realizarse el entrenamiento del modelo.	
Entrada/Pasos de ejecución: • Cargar imagen facial desde una dirección física especificada. • Clasificar la imagen facial.	
Resultado esperado: Se muestra en una notificación el grupo al que pertenece la imagen facial.	
Evaluación de la prueba: Satisfactoria	

Tabla 4.30: Caso de Prueba de la funcionalidad: Clasificar imagen facial.