BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Reconocimiento automático de escarabajos (Insecta: Coleoptera) usando imágenes digitales

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Caleb Jimenez Herrera

TUTORAS:

Dra. María de Lourdes Sandoval Solís

Dra. Hortensia Carrillo Ruiz

Puebla, Pue, Agosto 2017

NAASÓN JOAQUÍN GARCÍA APÓSTOL DE JESUCRISTO, SIN EL NADA DE ESTO HABRÍA SIDO POSIBLE; QUE DIOS PAGE POR LA VIDA, EL AMOR, EL CORAJE, EL TRABAJO Y EL EJEMPLO, TODOS ELLOS NECESARIOS PARA MATERIALIZAR CUALQUIER IDEA; HOY PUEDO VER QUE TODO VALIÓ LA PENA. TODAS LAS PALABRAS NOBLES QUE CONOZCO NO PUEDEN EXPRESAR NI EN UNA MÍNIMA CANTIDAD EL RESPETO QUE LE TENGO. ME HA ENTREGA, SU ESFUERZO, SU ENTUSIASMO, SU PERSEVERANCIA Y SU FORTALEZA. CON TODO SU AFECTO, ALIENTO, CONFIANZA, TERNURA Y COMPRESIÓN HA SIDO, ES Y SERÁ UNA COLUMNA EN MI VIDA. LA RESPONSABILIDAD QUE HA MOSTRADO POR ENSEÑARME HACER LO CORRECTO Y NO LO MÁS FÁCIL, ES SIN DUDA TAREA DE DISCIPLINA Y CONVICCIÓN QUE ESPERO COMPRENDE Y PRACTICAR. LA PASIÓN QUE ME HA DEMOSTRADO POR TODO LO QUE SE PROPONE ME INSPIRA PARA ENFRENTAR MIS MIEDOS Y LIMITANTES. SUS PROPÓSITOS ME EMOCIONAN Y DESAFÍAN A BUSCAR CADA DÍA, SER MEJOR Y ÚTIL, QUIERO IR DETRÁS DE EL SOY PARTE DE SU HISTORIA SOY DE NJG.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción.

Los escarabajos son insectos que pertenecen al orden Coleoptera, son organismos clave para el funcionamiento de los ecosistemas en los cuales habitan, debido a que se alimentan de animales muertos, estiércol, desechos vegetales, así como de tallos, hojas y ramas, lo que los convierte por un lado en recicladores naturales, ya que degradan e incorporan la materia orgánica al suelo, acelerando así la circulación de la energía almacenada en los desechos orgánicos; y por otro lado son capaces de consumir el follaje con lo que la planta responde con la producción de nuevo follaje y nuevas ramas. Además, debido a su diversidad, abundancia y valor nutritivo, son fuente importante de alimento para murciélagos, aves y otros artrópodos [5].

Es uno de los grupos de insectos más estudiados en el mundo, desde distintas áreas de la Biología, entre las que se encuentra la Sistemática. Los sistemáticos elaboran esquemas de clasificación con el propósito de que sirvan como un esquema general de referencia, es decir, que organicen a los diferentes organismos, con base en sus caracteres heredados, para poder ser estudiados. Actualmente, la diversidad biológica está organizada en categorías (p.e. Reino, Phylum, Clase, Orden, Familia, etc.) dentro de un sistema de clasificación jerárquico [25] .

Con estos sistemas de clasificación, los biólogos pueden identificar las especies que habitan en los diferentes ecosistemas, es decir, se establecen relaciones de identidad entre un organismo particular y una categoría a la cual pertenece de acuerdo a un esquema de clasificación establecido previamente. La forma más común de identificar especímenes es utilizar una clave. Las claves son dispositivos para identificar grupos de acuerdo a una secuencia ordenada de dilemas o disyuntivas, donde cada dilema sucesivo va planteando dilemas cada vez más restringidos y finalmente se llega al nombre del grupo o taxón al que pertenece [25]. Esto representa un arduo trabajo por parte de los biólogos que desean identificar sus ejemplares colectados. Muchos pasan horas incluso días empleando las claves de identificación las cuales son impresas y constan de más de 200 hojas.

Es por esta razón que en este trabajo se planteó generar por un lado, la automatización de las claves de identificación desarrollado por Morón *et al.* [23], lo cual es de gran ayuda para especialistas en el área y por otro lado, se exploró la posibilidad de identificar a estos organismos a partir de un sistema de reconocimiento basado en imágenes digitales, en donde se generó un sistema que emplea el contorno del cuerpo de estos insectos y permite identificarlos a nivel específico (especie), sin tener que recurrir a las claves de identificación empleadas por los biólogos.

Dado que la diversidad de Scarabaeoidea es muy grande, en este trabajo se utilizaron a las especies de dos familias, Scarabaeidae e Hybosoridae (Scarabaeoidea), que se distribuyen en la localidad del Rancho El Salado, en el municipio de Jolalpan en el estado de Puebla, México [29].

1.1. Planteamiento del problema.

La Sistemática es una disciplina de la Biología, la cual estudia científicamente la diversidad de organismos y sus relaciones evolutivas, entre sus actividades se encuentra la elaboración de esquemas de clasificación y la identificación de especímenes usando estos esquemas. Así, la clasificación biológica es el ordenamiento de los organismos en grupos o taxones sobre la base de sus relaciones. Un esquema de clasificación se basa en un sistema de jerarquías; durante los siglos XVII y XVIII se desarrolló la jerarquía o categoría Linneana que consiste en siete niveles principales: Reino, Phylum, Clase, Orden, Familia, Género y Especie. Actualmente este número de jerarquías o categorías ha aumentado y por lo tanto los sistemas se vuelven más complejos [25].

Un taxón es un grupo de cualquier rango que es considerado suficientemente distinto para ser reconocido formalmente como una categoría determinada y recibir un nombre, por ejemplo, algunos taxones son Animalia (Categoría Reino), Mammalia (Categoría Clase), Primates (Categoría Orden) y *Homo sapiens* (Categoría Especie) [25].

Otra actividad que realizan los sistemátas o taxónomos es la identificación o determinación de especies, es decir, el establecimiento de relaciones de identidad entre un organismo particular y el taxón al cual pertenece de acuerdo a una clasificación establecida previamente [19]. Por ejemplo, si alguien ha colectado un escarabajo en el campo, y de vuelta al laboratorio, emplea literatura apropiada como claves de identificación y lo compara con especímenes de una colección puede, decidir que se trata de un *Dynastes hyllus* de la familia Dynastinae, está identificando al organismo. El proceso de identificación requiere de mucha habilidad en el manejo de claves de identificación, las cuales son la forma común que tienen los biólogos para identificar especímenes.

Estas claves deben ser elaboradas de manera clara y deben incluirse características morfológicas que puedan observarse sin técnicas o equipos especiales, sin embargo, esto en la realidad no sucede y el tiempo que se invierte pueden ser horas incluso hasta días empleados en identificar un ejemplar [25][19].

De acuerdo con [3] las claves son instrumentos imperfectos, hacer claves sin errores es virtualmente imposible, además utilizan lenguaje que solo los especialistas del grupo pueden comprender. El trabajo de los biólogos desde el área de la sistemática es poder incrementar el conocimiento de la biodiversidad que habita en los diferentes ecosistemas. Dado que día con día se destruyen los hábitats a una gran velocidad, se hace urgente determinar qué especies habitan en estas regiones, determinar su importancia dentro del ecosistema y prevenir la destrucción de su hábitat. Los insectos, se han convertido en excelentes indicadores de la salud de los ecosistemas, ya que existen grupos que sostienen la estructura de los mismos. Tal es el caso de los escarabajos, quienes son considerados por algunos autores como especies indicadoras ambientales [33] y es urgente lograr identificar qué especies son y en qué números se encuentran en nuestra entidad. Por esta razón, se vuelve importante lograr disminuir el tiempo invertido en la identificación de estos organismos, y es aquí en donde el desarrollo de herramientas, desde el área de la computación se vuelven indispensables para lograr alcanzar este objetivo, el cual no es exclusivo solo para los insectos sino para la diversidad en general.

1.2. Objetivos.

La identificación de escarabajos es un proceso el cual es abordado por la taxonomía, la cual para muchos autores constituye una subdisciplina dentro de la Sistemática [19]. Entre las tareas de la taxonomía biológica, se encuentra el reconocimiento de especies con base en características morfológicas (diferencias y similitudes). Los taxónomos para esto, elaboran dispositivos denominados claves de identificación y por medio de éstos, pueden llevar a cabo el reconocimiento de especies en los diferentes niveles jerárquicos de un sistema de clasificación [25]. Identificar especies por lo tanto, requiere tiempo y conocimiento de los grupos con los cuales se trabaja, muchas veces se hace necesario desarrollar técnicas especializadas de preparación de los ejemplares a identificar montaje en seco, disecciones de genitalia, disecciones de piezas bucales, etc., razones por las cuales en este proyecto se planteó como objetivo general desarrollar un sistema computacional que permita realizar esta tarea en un menor tiempo, ver figura 1.1.

Figura 1.1: Ejemplar de *Phanaeus mexicanus*, recolectado en la región de Jolalpan y montado en seco con alfiler entomológico y con etiquetas de colecta e identificación.

A continuación se mencionan los objetivos desarrollados en este proyecto:

1.2.1. Objetivo general.

Desarrollar herramientas automatizadas que disminuyan el tiempo de identificación de las especies de escarabajos que se distribuyen en el rancho El Salado, ubicado en la localidad Jolalpan en el estado de Puebla, México.

1.2.2. Objetivos particulares.

1. Automatizar las claves de identificación para estas especies y disminuir el tiempo de identificación para los taxónomos.

2. Elaborar un sistema computacional el cual logre identificar a las especies de escarabajos con base en las imágenes digitales, sistema que se pretende sea útil no solo a los taxónomos sino al público en general.

1.3. Estructura de la tesis.

Este trabajo está dividido en 6 capítulos.

Capítulo 1. Se da una introducción de la importancia de los escarabajos en los ecosistemas en los que habitan, y la importancia de poder clasificarlos e identificarlos para su estudio. Se presentan la forma en la que actualmente la Biología los identifica con ayuda de esquemas elaborados por taxomatas. Se plantea el problema de identificación de estos insectos y todo el trabajo que conlleva. Se mencionan los objetivos que se alcanzaron en este trabajo.

Capítulo 2. Se desarrolló un sistema computacional llamado "Automatización de claves de las especies de dos familias, *Scarabaeidae e Hybosoridae (Scarabaeoidea)*". Se presenta la metodología para automatizar las claves de clasificación, donde se presenta el análisis, el diseño, implementación y validación del sistema computacional.

Capítulo 3. Presenta una alternativa para la identificación de escarabajos utilizando imágenes digitales de las familias *Scarabaeidae e Hybosoridae (Scarabaeoidea)*, recolectadas en la región de Jolalpan en el estado de Puebla.

Capítulo 4. Se describe el diseño y la implementación del sistema computacional de identificación de escarabajos utilizando imágenes digitales, de una manera detallada a través de diagramas.

Capítulo 5. Se muestran los resultados de los experimentos de la identificación de escarabajos con imágenes digitales empleando tres clasificadores diferentes *SVM*, *Naive Bayes* y *Random Forest*. Se menciona el tipo de características extraídas y la variación entre ellas especialmente en los número de armónicos aplicados, debido a que se busca el número de armónico óptimo para la clasificación.

Capítulo 6. Por último se mencionan las conclusiones y el trabajo futuro.

Capítulo 2

Automatización de claves de las especies de dos familias, Scarabaeidae e Hybosoridae (Scarabaeoidea).

Por esta razón, se planteó trasladar del sistema de claves de identificación para las especies de Puebla, los dilemas que permiten la identificación de las especies que se distribuyen en la región de Jolalpan a un sistema automatizado. Este sistema tiene la función de ser una herramienta de investigación que agiliza la clasificación de las especies de escarabajos disminuyendo el tiempo empleado por los taxónomos o bien por los estudiantes del área.

En México hasta el momento, no existe una aplicación que supla las necesidades que se tienen dentro del área biológica en este sentido, específicamente dentro del área de la Sistemática.

Entre las actividades principales de los biólogos que se especializan en el área de la Sistemática, se encuentra identificar taxones, de ahí su nombre de taxónomos [19]. Para identificar especies adecuadamente, una de las tareas más básicas es la búsqueda de la literatura específica para el grupo en cuestión. Libros, guías de campo, revisiones y otros trabajos generales pueden ayudarnos en la búsqueda exploratoria inicial, pero se requiere consultar literatura especializada como claves de identificación. Las claves de identificación se basan en una diagnosis, esto es, en un enunciado breve de los caracteres que permiten identificar a los ejemplares dentro de cada una de las categorías del sistema de clasificación (Orden, Familia, Subfamilia, Género y Especie). Así, las diagnosis diferenciales señalan específicamente cómo la especie difiere de otras relacionadas o con las cuales prodría confundirse [25].

Así, los especímenes recolectados en campo o bien depositados en colecciones biológicas, se identifican utilizando claves de identificación,

las cuales se elaboran utilizando los sistemas de clasificación disponibles y analizando los caracteres provistos por especímenes de referencia, esto hace que por un lado la elaboración de las claves sea una actividad de varios años y por otro lograr la identificación de los organismos empleando estas claves, requiere de una experiencia básica que consiste en tener cierta familiaridad con aspectos de la morfología, empleo de conceptos dentro de la sistemática y conocer las técnicas de preparación y observación bajo microscopio [24].

De este modo, las claves taxonómicas son la forma más común para identificar especies de los diferentes grupos biológicos. Son herramientas para reconocer taxones de acuerdo con una secuencia ordenada de disyuntivas, en donde cada dilema va planteando disyuntivas cada vez más acotadas hasta que se llega al nombre del taxón [25].

Desde un punto de vista biológico las claves son artificiales, ya que no se busca que plasmen las relaciones evolutivas sino más bien que se logre identificar el taxón. Por lo tanto, estas herramientas son perfectibles, es decir, conforme se van empleando se pueden detectar errores dado que toda la variación de los organismos de una especie no puede ser englobada dentro de estos dilemas que conforman la clave.

Para los coleopteros escarabaeoideos que se distribuyen en el estado de Puebla, la elaboración de claves de identificación inició con los primeros estudios del biólogo Federico Islas Salas entre los años 1941 y 1942 [23], a partir de ahí las claves se han ido perfeccionando, actualmente se cuenta con la clave impresa y publicada por Miguel Ángel Morón, para identificar las especies de coleóptera Scarabaeoidea del estado de Puebla, la cual consta de aproximadamente 304 dilemas que incluyen aproximadamente 600 opciones [23]. Tan solo para identificar una especie que se distribuya en el estado, la tarea consume varias horas de trabajo.

2.1. Antecedentes.

Acerca de las claves de identificación automatizadas, exite la página web Generic Guide to New World Scarab Beetles de la Universidad de Nebraska, que implementa las claves para la Superfamilia Scarabaeoidea basadas en el esquema de clasificación de Lawrence y Newton [15]. En esta página se presentan diferentes opciones como Guide Home, Taxa Map, Keys to Taxa, Catalogs, Gallery y Search. En la sección de Keys to Taxa es donde se encuentran implementadas las claves para la Superfamilia Scarabaeoidea, la forma en la que se presentan es la siguiente: dilema compuesto de sus dos opciones; diagnosis de caracteres en cada opción; finalmente algunas figuras que ilustran las características que describe a la especie. Debido a que estas claves fueron implementadas para la Superfamilia Scarabaeoidea, la primera opción define a esta superfamilia y por ende tiene que cumplirse, en la segunda opción es donde empiezan los dilemas que se recorren hasta llegar a su especie. Una vez que se llega a la especie, se muestra una imagen o dibujo que la representa y la opción para más información. Es importante mencionar que este trabajo se inició el 6 de marzo del 2002 y su última modificación respecto a las claves fue el 15 de noviembre del 2005, por lo que ya tiene

12 años sin actualizar y los esquemas de clasificación se han actualizado, además se limita a su uso con acceso a internet o bien se pueden imprimir las claves lo que deja en la misma situación de uso [17].

Para poder llevar a cabo este proyecto fue necesario conocer las claves elaboradas por los sistématas y taxónomos, su estructura y funcionamiento. En México se emplea para la clasificación de escarabaeoideos, el sistema propuesto por Endrödi [7] el cual reconoce 13 familias entre las que se encuentran Scarabaeidae e Hybosoridae. Es por esto, que las claves de identificación del estado de Puebla se basan en este sistema.

Posteriormente para lograr interpretar las características en las diagnosis de las especies, se conocieron los conceptos englobados en los dilemas de la clave, esto por medio del "Curso-Taller de taxonomía y Ecología de escarabajos (Coleoptera: Scarabaeoidea)" impartido por la doctora Hortensia Carrillo-Ruiz, en este curso se determino la forma en la que los biólogos realizan la identificación de sus escarabajos empleando microscopios esteresocópicos y las claves de identificación. Así, se logra determinar la manera en que se identifican las 16 especies de las dos familias empleadas en este proyecto.

2.2. Metodología y análisis del sistema.

Modelo de ingeniería de Software. Para el desarrollo del software se utilizó el modelo en cascada también conocido como lineal secuencial, este modelo contiene etapas bien definidas y en cada etapa se fijan objetivos. No se puede avanzar a la siguiente etapa sin tener los objetivos de la etapa actual completamente cubiertos [32] ver figura 2.1.

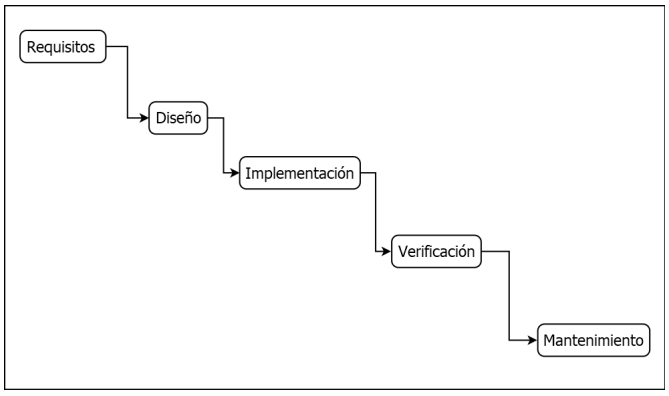

Figura 2.1: Modelo en cascada que muestra el proceso de desarrollo del sistema.

Las etapas del modelo en cascada reflejan las actividades que se tienen que desarrollar, a continuación se menciona cada etapa de forma particular:

- Requisitos: En esta etapa se analiza las necesidades del usuario, para definir las funcionalidades del software, establecer restricciones, servicios y metas [32].

- Diseño: Describe la estructura del proyecto, se define el funcionamiento del software, el rendimiento y se proponen las interfaces del sistema, esto implica identificar y describir de forma clara las bases del sistema de software y sus relaciones ayudados de ilustraciones y diagramas [32].
- Implementación: En esta etapa es tiempo de programar es decir traducir el diseño a una forma legible para la máquina, si el diseño fue correcto la implementación es directa [32].
- Verificación: Para validar el software se ingresan datos de los cuales se conocen la salida, es importante verificar el manejo de errores. Después de esta validación el software es liberado [27].
- Mantenimiento: En esta etapa se contemplan cambios que el cliente requiera o corrección de errores que no se consideraron en la verificación, también el cliente puede pedir ampliaciones funcionales en el sistema [27].

2.3. Aplicación del modelo al problema de estudio.

2.3.1. Requisitos:

Los biólogos taxónomos requieren un software en el que se presente de manera automatizada las claves que ocupan para identificar a los escarabajos. Dado que la diversidad es muy grande y no se cuenta con todos los ejemplares de insectos este software se limitó a dos familias, Scarabaeidae e Hybosoridae (Scarabaeoidea), que se distribuyen en la localidad del Rancho El Salado, en el municipio de Jolalpan en el estado de Puebla [23].

Otro requisito que se considera es seguir las claves de forma precisa pues el trabajo que se tiene que realizar es automatizar este método de identificación de insectos.

Un requisito indispensable es el Diagrama de Esquemas de claves de las especies de dos familias, Scarabaeidae e Hybosoridae (Scarabaeoidea) que se muestra en la figura 2.2.

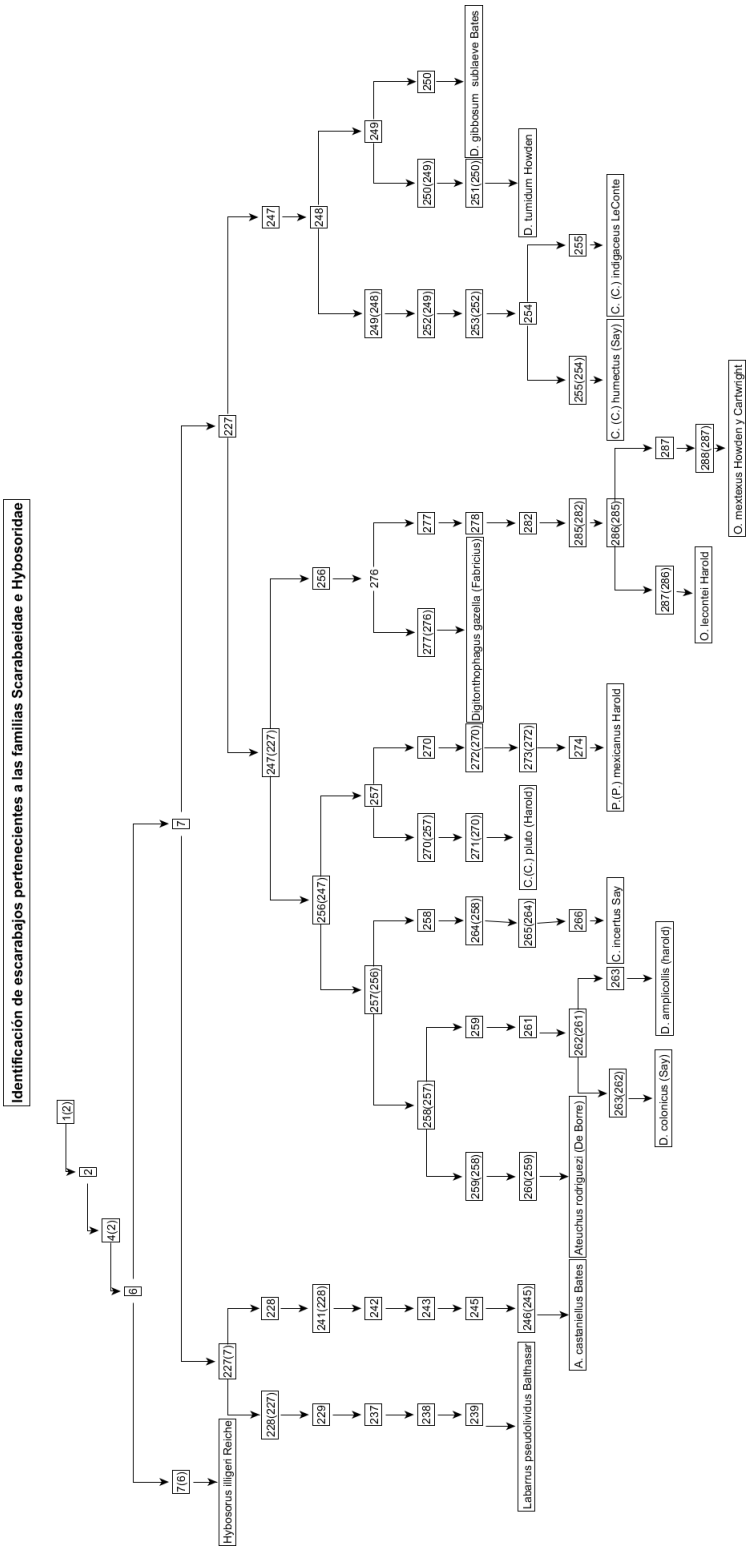

Figura 2.2: Diagrama de esquemas de claves de las especies de dos familias, Scarabaeidae e Hybosoridae (Scarabaeoidea).

2.3.2. Diseño.

Para el diseño del software, se cuenta únicamente con un actor y siete casos de uso ver figura 2.3.

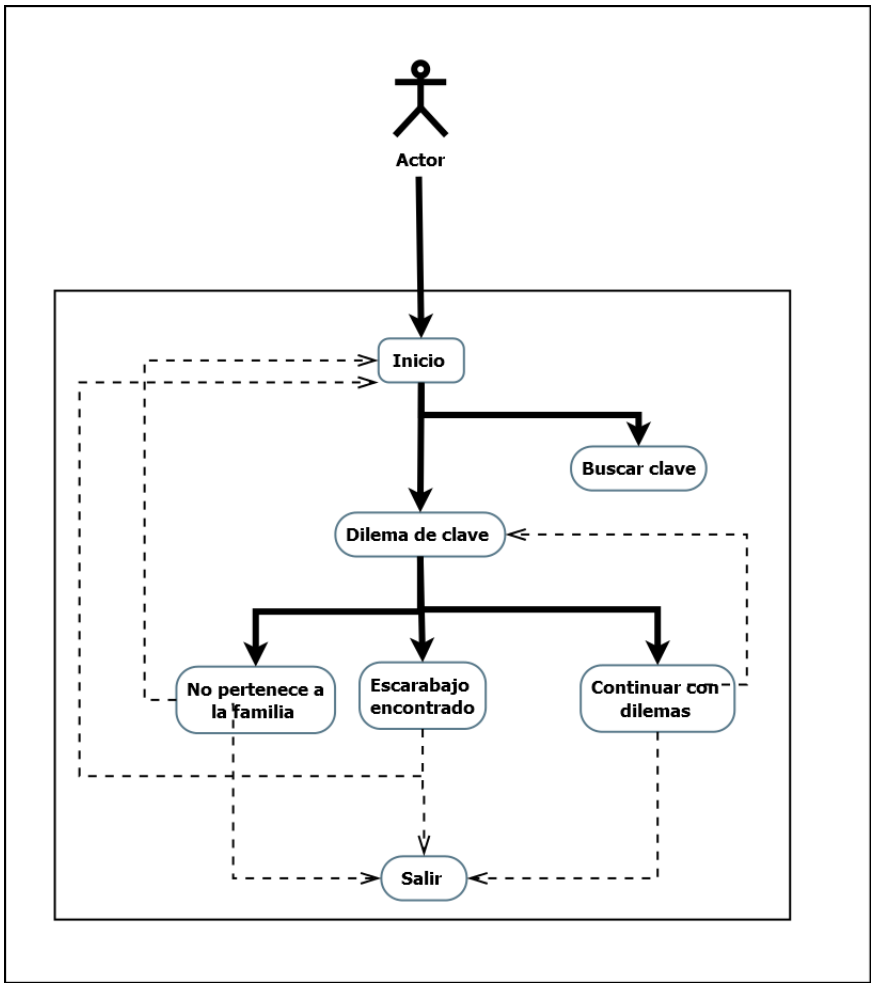

Figura 2.3: Diagrama de casos de uso, se emplean en total 7 casos de uso.

El único actor que se presenta en el sistema es el biólogo taxónomo, para cada espécimen se tiene que realizar el mapeo de todos los dilemas correspondientes que conforman el esquema.

2.3.2.1. Descripción de los casos de uso:

a). **Caso de uso Inicio:**

1) **Actores:** Biólogo taxónomo.
2) **Descripción:** El usuario ingresa al sistema, verifica si cumple o no con el dilema y da click en aceptar.
3) **Flujo principal:**
   - El usuario verifica si se cumplen con los dilemas.
     - Selecciona sí o no dependiendo de su espécimen.
     - Da click en aceptar.

b). **Caso de uso Dilema de clave:**

1) **Actores:** Biólogo taxónomo.
2) **Descripción:** El usuario ingresa al sistema, verifica si se cumplen los dilemas y da click en aceptar.
3) **Flujo principal:**
   - El usuario verifica si se cumplen con los dilemas.
     - Selecciona sí o no dependiendo de su espécimen.
     - Da click en aceptar.

c). **Caso de uso No pertenece a la familia:**

1) **Actores:** Biólogo taxónomo.
2) **Descripción:** Muestra un mensaje No pertenece la familia y nos presenta la opción de nueva búsqueda o salir del sistema.
3) **Flujo principal:**
   - • El usuario decide nueva búsqueda y acepta.
     - El usuario decide salir y se termina el proceso.

d). **Caso de uso escarabajo encontrado:**

1) **Actores:** Biólogo taxónomo.
2) **Descripción:** Muestra el nombre de la especie una breve descripción, una imagen muestra de nuestro escarabajo y nos presenta opción de nueva búsqueda o salir del sistema.
3) **Flujo principal:**
   - • El usuario decide nueva búsqueda y acepta.
     - El usuario decide salir y se termina el proceso.

e). **Caso de uso continuar con dilemas:**

1) **Actores:** Biólogo taxónomo.
2) **Descripción:** Es un paso intermedio entre los dilemas siempre que se cumplan y no se recorra por completo el esquema de clasificación, nos presenta la opción de una clave anterior, nueva búsqueda o salir.
3) **Flujo principal:**
   - El usuario decide nueva búsqueda y acepta.
   - El usuario decide salir y se termina el proceso.
   - El usuario regresa a un dilema anterior ó continua con el siguiente dilema.

f). **Caso de uso buscar clave:**

1) **Actores:** Biólogo taxónomo.
2) **Descripción:** El usuario puede buscar una clave sin tener que recorrer las claves anteriores, el sistema lo posiciona en la clave deseada.
3) **Flujo principal:**
   - El usuario decide nueva búsqueda y acepta.
   - El usuario decide salir y se termina el proceso.
   - El usuario regresa a un dilema anterior ó continua con el siguiente dilema.
4) **Excepción E1:**
   - La clave no existe intente de nuevo

2.3.2.2. Interfaz:

El objetivo principal es que la interacción entre usuarios y aplicación cubra los siguientes requisitos:

- Que sea muy intuitivo (fácil de usar). Con ayuda de botones y cuadro de texto se orienta y describe el funcionamiento del sistema, además, se presenta información necesaria en cada evento evitando la carga de información.
- Que el diseño sea agradable a la vista. El tamaño de letra se determinó de acuerdo a la cantidad de información que se presenta y los elementos gráficos para la interacción se dimensionaron para no exponer al usuario a un tamaño en donde le afecte o le cueste trabajo apreciar los mensajes. Los colores empleados fueron elegidos con base en [13] que nos menciona que el color blanco es el que mejor refleja la luz y por lo tanto posee la mayor sensibilidad frente al usuario. Es la suma o síntesis de todos los colores (colores luz).
- Minimizar los errores. Se contemplan los posibles errores para tener control sobre ellos y no tener salidas inesperadas en el sistema.

2.4. Implementación.

En esta fase se codifica cada requerimiento que se obtuvo del análisis hasta tener el software en un funcionamiento operacional completo. Cubriendo las necesidades de los usuarios.

2.4.1. Herramientas de desarrollo:

Se eligieron herramientas de software libre que proporcionan confiabilidad y facilidad para la elaboración del proyecto: Para el análisis, se utilizó el software Dia [9] es un editor de diagramas versátil y fácil de usar, que permite crear y modificar diagramas.

El lenguaje de programación en el que se desarrolla la estructura es en java versión 8.0.2 lo que permite programar orientado a objetos, es multiplataforma completamente gratuito y cuenta con un entorno de desarrollo integrado libre NetBeans que facilita realizar interfaz humano computadora, tiene un entorno de desarrollo muy completo, puede trabajar con las últimas tecnologías, constantemente tiene mejoras y existen muchos foros de desarrollo [30].

2.4.2. Programación del árbol binario.

Una vez que se conoce la forma en la que se realiza la identificación y se cuenta con la estructura, se procede a la programación. En cuanto al modelo se implementó un árbol binario que es una estructura de datos, la cual cuenta con un nodo llamado raíz que da origen a la identificación, a partir de este nodo se desglosan dos nodos más y estos pasan a un segundo nivel. De esta forma se cumplen las propiedades que los taxónomos establecieron en su clave de identificación, que son las siguientes: tienen un orden jerárquico y solo se cuenta con dos opciones para cumplir la propiedad que se presentan en dilemas. Si el escarabajo que se desea identificar, no pertenece a ninguna de las dos familias se llegará a un punto donde no existen argumentos suficientes para poder identificarlo. De esta manera se ubica en un nodo que indica que no pertenece a ninguna de las dos familias. Si el escarabajo a identificar pertenece a alguna de las dos familias, en cada dilema se presentan argumentos para continuar hasta determinar la especie. En este punto existe un nodo más, el cual proporciona el nombre de la especie a la que pertenece el escarabajo y una imagen icono.

2.4.3. Funcionamiento del sistema.

Se muestra el flujo del software a través de imágenes de pantalla con una descripción en cada imagen de lo que se realiza en esa pantalla.

a. **Pantalla principal:**

En la figura 2.4 muestra la pantalla principal del sistema, en donde se puede notar el dilema 1 y 1(2), que determina la dirección en la búsqueda en el árbol. Como segundo punto, se presenta una descripción de cada una de las características con la pregunta siguiente: **¿Cumple con las características?** y las opciones con **sí** o **no**, más el botón de confirmación.

La pantalla principal también incluye un apartado de búsqueda que ayuda a posicionarse en un nodo conocido sin tener que recorrer todos los nodos anteriores y a partir de este nodo conocido continuar con la identificación.

También cuenta con una imagen de un escarabajo original (Dibujado por Carrillo-Ruiz,H.) como logotipo; un botón que regresa a un nivel anterior y finalmente un botón para salir del sistema ver figura 2.4.

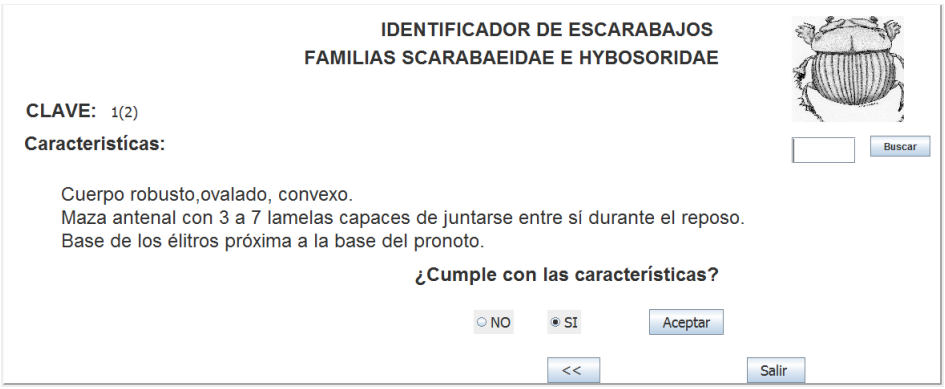

Figura 2.4: Pantalla principal del sistema IDENTIFICADOR DE ESCARABAJOS FAMILIAS SCARABAEIDAE E HYBOSORIDAE.

b. **Pantalla del escarabajo identificado:** En esta pantalla aparece el nombre de la especie a la que pertenece el espécimen, después se muestra una imagen de la especie ver figura 2.5.

Cuenta con los botones para terminar y nuevo, la opción terminar finaliza el programa; mientras que la opción nueva reinicia el programa para una nueva identificación.

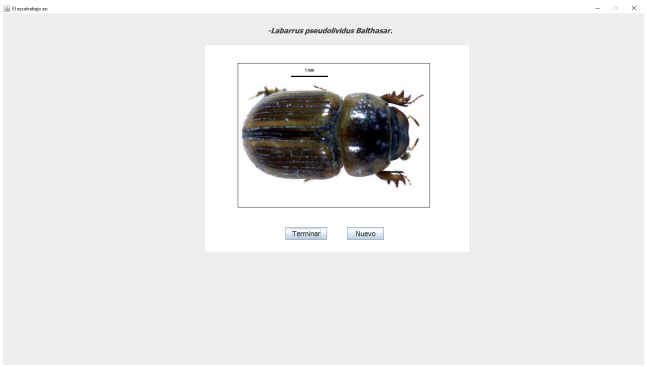

Figura 2.5: Pantalla de la identificación de *Labarrus pseudolividus* en donde se muestra una imagen de la especie.

c. **Pantalla Escarabajo no pertenece a las familias:** En esta pantalla aparece el mensaje de: No pertenece a ninguna de las dos familias, como resultado de que el espécimen analizado no cumple con las características ver figura 2.6.

Cuenta con los botones para terminar y nuevo, la opción terminar finaliza el programa; mientras que la opción nueva reinicia el programa para una nueva identificación.

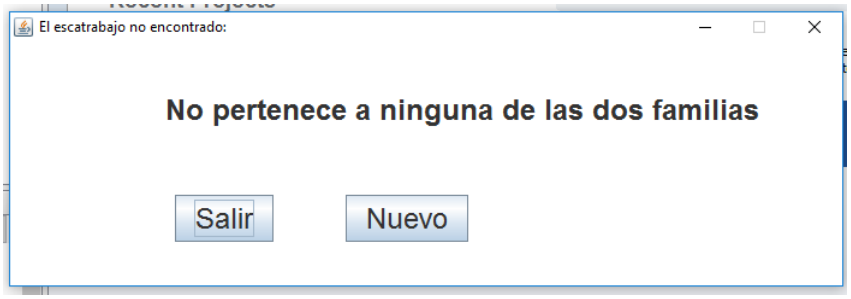

Figura 2.6: Pantalla de no pertenencia a ninguna de las dos familias.

d. **Pantalla nodo inicial:** Es una pantalla emergente de alerta que avisa que se encuentra en el nodo de inicio de la clasificación ver figura 2.7.

Solo contiene el botón **ok** que garantiza que el usuario está enterado que llego al punto de origen.

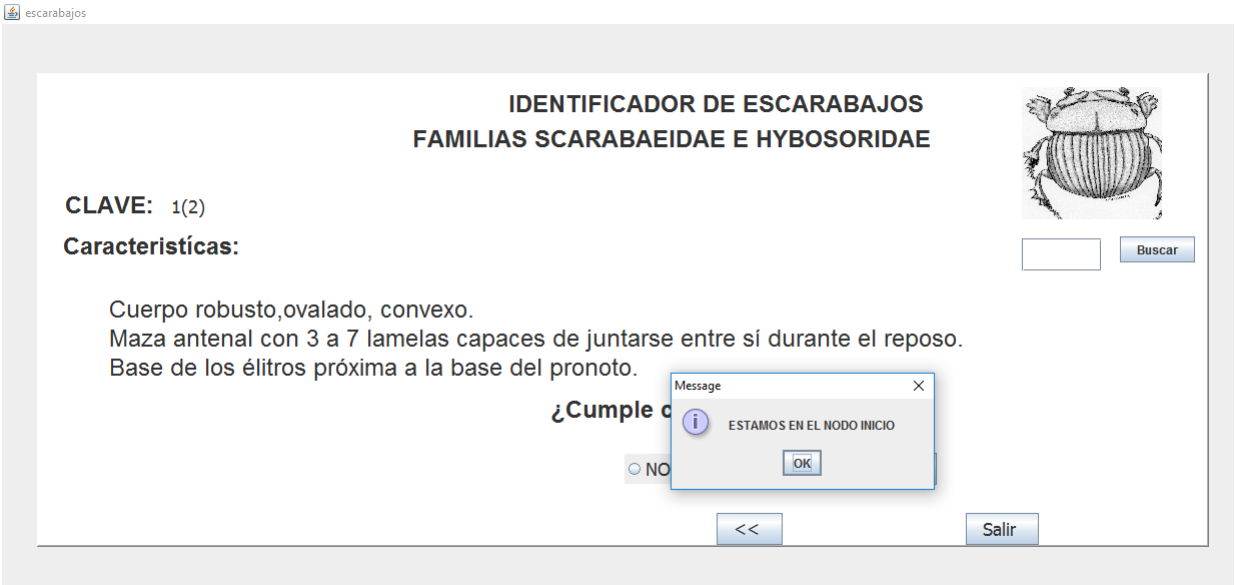

Figura 2.7: Pantalla emergente de alerta de inicio de la clasificación.

2.4.4. Verificación.

Para validar el sistema fue probado por 75 estudiantes de Biología de los cuales identificaron 5 especies cada uno, dando un total de 375 de las cuales 322 identificaron al escarabajo correctamente y 53 obtuvieron a un escarabajo diferente, este error se presentó por que los alumnos al observar los escarabajos no identificaron sus cualidades morfológicas correctamente, ayudados y orientados por la doctora Hortensia Carrillo lograron la identificación correctamente. Los resultados eran esperados pues la aplicación representa las claves que utilizan los biólogos y los errores que se presentan es por interpretaciones incorrectas de las claves y no por un mal funcionamiento del sistema.

2.5. Conclusiones.

El software cumple con los objetivos fijados, es una herramienta que facilita la identificación de escarabajos de las familias Scarabaeidae e Hybosoridae, permite al usuario centrar su atención en su espécimen y no en la búsqueda de las claves, otro objetivo alcanzado es que se redujo el tiempo empleado para esta tarea, para llegar a esta conclusión se tomaron los tiempos de 10 parejas de estudiantes con conocimiento de la nomenclatura de las claves, para realizar la identificación de escarabajos uno con las claves impresas y otro estudiante identificando al mismo escarabajo pero con ayuda del software y el tiempo empleado por estos últimos se redujo en un promedio de 54 %.

Capítulo 3

Clasificador usando imágenes digitales

En esta sección se presenta un método para identificar escarabajos de manera automática con ayuda de técnicas de inteligencia Artificial empleando como atributos, medidas estadísticas y la transformada elíptica de Fourier, extraídas de las imágenes digitales de los escarabajos. Se utilizaron 48 imágenes de 8 especies diferentes de la colección del Laboratorio de Entomología de la Facultad de Biología de las familias Scarabaeidae e Hybosoridae (Scarabaeoidea), recolectadas en la región de Jolalpan en el estado de Puebla.

Se entrenaron tres clasificadores diferentes usando las imágenes proporcionadas por los biólogos, esta calibración se realizó de la siguiente forma: 1) Adquisición de imágenes 2) Pre-procesamiento 3) Extracción de características 4) Clasificadores (NaiveBayes, RandomForest, SVM) ver figura 3.1.

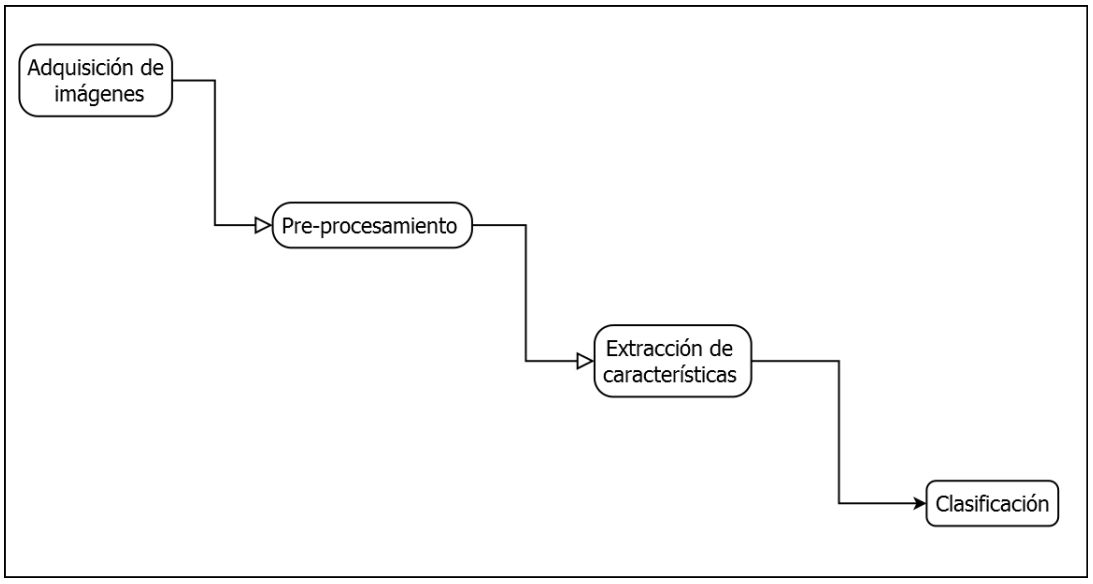

Figura 3.1: Método de procesamiento digital de imágenes.

3.1. Antecedentes.

Adams, *et al.* en [1] plantean un enfoque denominado Morfología geométrica. Este enfoque se basa en la digitalización de contornos o puntos clave, cuyas conformaciones espaciales son analizadas matemáticamente. Dentro de este grupo de técnicas, los métodos de contornos se basan en la digitalización de los puntos alrededor de un contorno para ajustarlos a una función matemática, generalmente derivada del análisis de Fourier. Posteriormente los coeficientes de éste, pueden emplearse en estudios comparativos.

Los descriptores elípticos de Fourier, propuestos por Kuhl y Giardina usado en [16], pueden delinear cualquier forma con un contorno cerrado bidimensional. Estos han sido aplicados al análisis de varias formas biológicas. Así, Rohlf y Archie en [28] presentan un análisis del contorno de las alas de 127 especies de mosquito del norte de México y comparan diferentes descriptores de Fourier, normalizado, con coordenadas polares, entre otros y menciona que los descriptores de la trasformada elíptica de Fourier prometen resultados satisfactorios para la diferenciación entre especies. Por otra parte Furuta *et al.* en [10] analiza la forma de las hojas del frijol de soya empleando los coeficientes de los armónicos de la transformada elíptica de Fourier, emplea 20 armónicos normalizados y por lo tanto 77 descriptores por imagen de hoja. Con ellos realiza análisis de componentes principales y encuentra que la contribución acumulativa del quinto componente es del 96 %. Concluye que los coeficientes de Fourier proporcionan una medida cuantitativa poderosa para evaluar la forma de las hojas del frijol de soya. Zhan y Wang en [37] utilizan la transformada elíptica de Fourier para analizar la forma del contorno del ala de cinco especies de Antlion (hormiga león parecida a la libélula). Ellos analizaron variaciones de la forma del perfil de las alas de 5 especies: considerando un total, 98 alas anteriores y 98 alas posteriores. Utilizan los veinte primeros armónicos de Fourier que resume a través de un análisis de componente principal y consideran los primeros 8 componentes principales de la variación de forma para realizar pruebas estadísticas (análisis de varianza multivariable, análisis de variables canónicas y análisis de conglomerados). Concluyen que sus resultados del análisis de Fourier de contorno de alas de las 5 especies están de acuerdo con el sistema taxonómico actual.

Recientemente los descriptores elípticos de Fourier, han sido empleados para proponer la construcción de herramientas para la identificación automatizada de especies como lo mencionan Singh, et al. en [31] quienes han desarrollado un sistema automatizado de reconocimiento de especies de bambú basado en las características de forma de la vaina de Culm de bambú usando momentos de Fourier y de Legendre. Concluyen que el momento de Fourier tiene resultados significativamente mejores que el momento de Legendre, obteniendo un 100 % de exactitud en la clasificación. El autor comenta que su sistema puede eliminar la necesidad de métodos laboriosos de reconocimiento humano que requieran un taxónomo de plantas.

Los resultados obtenidos muestran una considerable precisión de reconocimiento, demostrando que las técnicas utilizadas son adecuadas para ser implementadas con fines comerciales.

Yang *et al.* en [36] presentan un trabajo para la identificación de insectos basado en el borde de sus alas reportando una exactitud media para la identificación de especies que varia entre el 90 % y 98 %. Utilizan 120 ejemplares de 7 especies distintas. Emplean la transformada elíptica de Fourier con 30 armónicos y consideran 117 coeficientes para obtener las características del borde de las alas de estos insectos. Para la clasificación emplean máquinas de soporte vectorial. Finalmente los descriptores elípticos de Fourier invariantes, simétricos, asimétricos y estandarizados, también se ha utilizados para clasificar granos en Mebatsion *et al* [22]., evalúan la forma de cuatro diferentes tipos de grano: cebada, avena centeno y trigo. Utilizan 100 imágenes por cada grano, los autores mencionan que los descriptores elípticos de Fourier ofrecen buenos resultados considerando la variabilidad de la forma de los granos y que podrían utilizarse para realizar clasificación no supervisada.

Hasta el momento no se tiene información de la utilización de los descriptores elípticos de Fourier para el estudio de las formas de los escarabajos.

3.2. Fundamentos teóricos.

3.2.1. Procesamiento de imágenes digitales.

El ser humano cuenta con un mecanismo de procesamiento de imágenes sin duda el que más utiliza, debido a que es capaz de detectar analizar, identificar y almacenar una gran cantidad de imágenes. El procesamiento digital de imágenes es un área muy amplia, de manera muy general es la manipulación y análisis de imágenes por computadora. Como objetivos se puede mencionar, la extracción de información, la captura de imágenes digitales, resaltar áreas de interés, segmentar, medir, identificar y visualizar formas de interés.

Actualmente existen una gran variedad de áreas en las que se aplica el procesamiento digital de imágenes, como es la medicina, el medio ambiente, la industria, la seguridad, la gestión, las comunicaciones y control de procesos industriales [34]. Un modelo de procesamiento digital de imágenes es el siguiente: 1) Adquisición de imágenes, 2) Pre-procesamiento (Recortar imagen digital, convertirla a escala de grises, binarizar imagen, eliminar ruido, identificar borde y definir borde), 3) Extracción de características. Para el desarrollo de este trabajo se utilizó el modelo anterior y se agregaron los módulos clasificación e identificación: 4) Entrenamiento de clasificadores 5) Identificación de especies de escarabajos, como se puede observar en la figura 3.2.

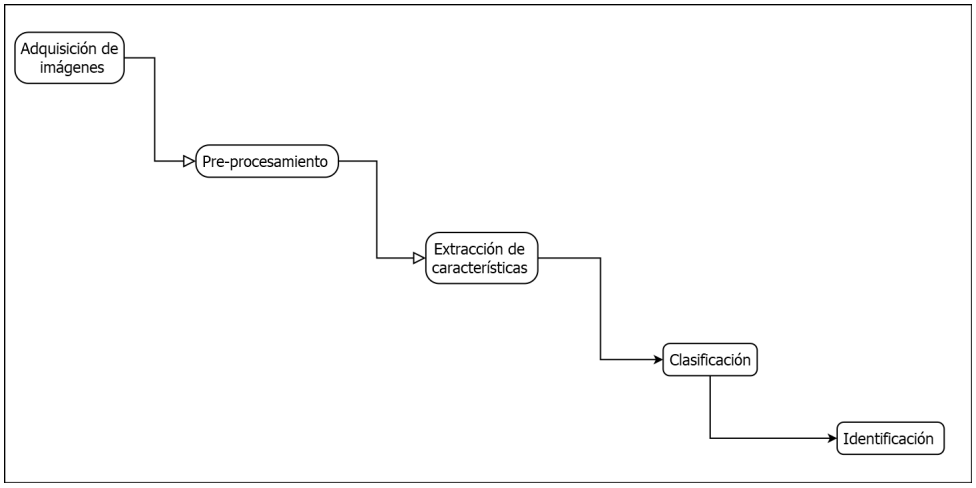

Figura 3.2: Modelo de procesamiento digital de imágenes con clasificador e identificador.

Cada una de las etapas del procesamiento de imágenes digitales tiene un objetivo en particular, y existen una gran variedad de métodos definidos para cumplir con estos objetivos, a continuación, se definen los que se ocuparon en la metodología de desarrollo de este trabajo:

3.2.2. Adquisición de imágenes:

Para obtener imágenes digitales se consideran algunos puntos importantes, como la cantidad de luz, el escenario adecuado de ser posible, una cámara fotográfica digital y un tripie. Cuando se obtienen imágenes digitales para procesamiento digital de imágenes es importante que se considere el área de interés que desea procesar y de esto depende la distancia a la que debe colocarse la cámara, la cantidad de luz, si se puede usar el tripie si es necesario utilizar el flash de la cámara. En la adquisición de imágenes existen factores físicos difíciles de controlar, entre los que se pueden citar los niveles bajos de iluminación, la reflexión sobre los objetos y el ruido aleatorio, que hacen que las imágenes no presenten siempre una buena calidad para su utilización [26].

Las imágenes fueron obtenidas de un grupo de ejemplares de colección provenientes de la región de Jolalpan en el estado de Puebla, estos ejemplares fueron recolectados durante un año en esta zona, su captura no es sencilla por lo que son pocos los representantes empleados en este estudio [29]. En total se emplearon 67 especímenes conservados en seco, los cuales fueron previamente identificados de manera tradicional empleando las claves de identificación para las especies de escarabajos del estado de Puebla [23] y pertenecen a dos familias, tres subfamilias, seis tribus, 12 géneros y 16 especies como se muestra en la Tabla 3.1.

Figura 3.3: Escarabajo *Canthon Canthon) humectus* en posición dorsal y escala de referencia de diez milímetros.

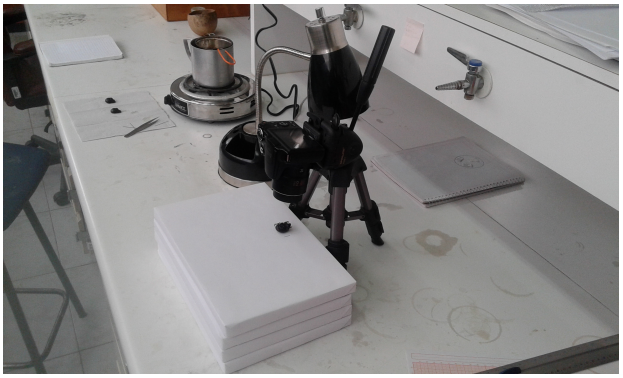

Figura 3.4: Obtención de imágenes de los ejemplares estudiados.

Para obtener las imágenes, cada uno de los ejemplares se reblandece con vapor de agua para poder retirar el alfiler entomológico que se emplea para conservarlos. Posteriormente se colocaron en posición dorsal sobre una placa de unicel forrada con papel blanco, sobre la cual se dibuja una escala de referencia en milímetros ver figura 3.3. Finalmente se hace la fotografía con una cámara canon 1068C001AA montada sobre un tripie figura 3.4.

Tabla 3.1: Ejemplares de escarabajos empleados en el estudio, se muestran los nombres que les corresponden dentro de cada una de las categorías taxonómicas a partir de familia.

Familia	Subfamilia	Tribu	Género, subgénero y especie	No. ejemplares
Scarabaeidae	Scarabaeinae	Scarabaeini	*Canthon (Canthon) humectus*	6
			Canthon (Canthon) indigaceus	6
			Deltochilum gibbosum sublaeve	6
			Deltochilum tumidum	5
		Coprini	*Copris incertus*	2
			Dichotomius colonicus	6
			Dichotomius amplicollis	1
			Ateuchus rodriguezi	6
		Phanaeini	*Phanaeus mexicanus*	7
			Coprophanaeus (Coprophanaeus) pluto	3
		Onthophagini	*Onthophagus lecontei*	1
			Onthophagus mextexus	2
			Digitonthophagus gazella	6
	Aphodiinae	Aphodiini	*Labarrus pseudolividus*	2
		Eupariini	*Ataenius castaniellus*	3
Hybosoridae	Hybosorinae		*Hybosorus illigeri*	5

3.2.3. ¿Qué es una imagen digital?

Es una representación de un objeto real como una función bidimensional, f(x, y) donde x e y son una secuencia de coordenadas espaciales planas, en un sistema de color RGB se conoce como imagen en color porque consta de tres imágenes de componentes individuales (rojo, verde y azul). Por esta razón existen imágenes monocromáticas e imágenes a color. Una imagen puede ser tratada como una matriz cuyos índices de renglón y columna determinan un punto en la imagen y el correspondiente valor del elemento de la matriz identifica el nivel de intensidad de luz en ese punto, los elementos de tal arreglo digital son llamados elementos de imagen, elementos de pintura, pixels o pels. Un píxel es la unidad mínima de visualización de una imagen digital. Si se aplica el zoom sobre ella se observa que está formada por una parrilla de puntos o píxeles. Las cámaras digitales y los escáneres capturan las imágenes en forma de cuadrícula de píxeles ver figura 3.5 [34].

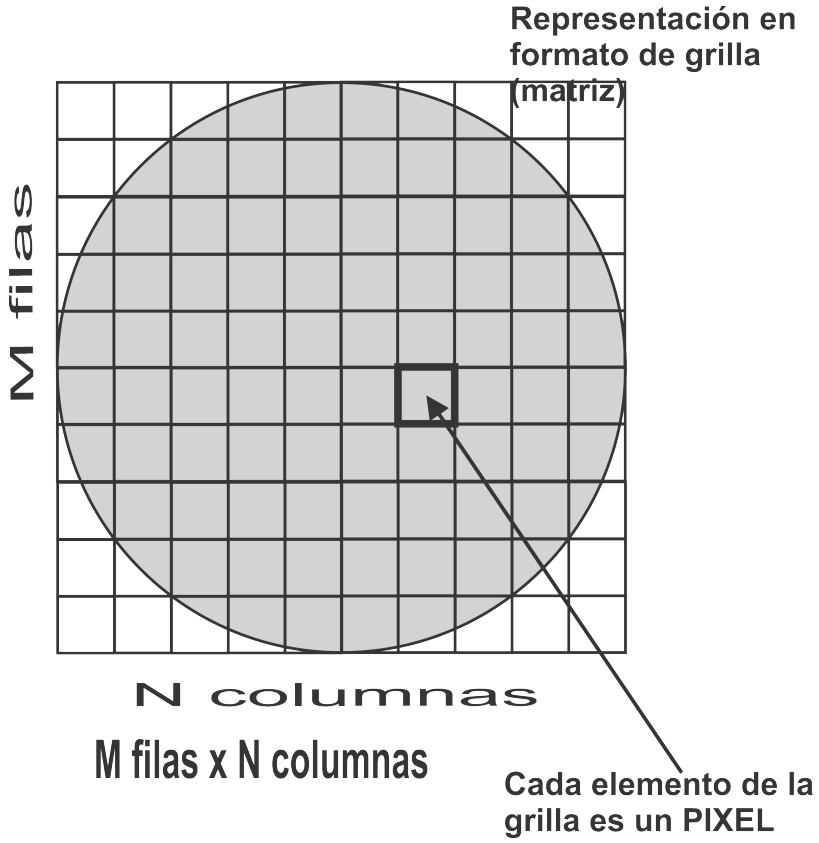

Figura 3.5: Representación de píxel.

3.2.4. Pre-procesamiento:

En este proceso se involucran el análisis de una imagen digital para aplicarle una serie de funciones para el mejoramiento de la imagen y extracción de características de una sección o área de la imagen digital, por esto entre las técnicas que se utilizan esta la segmentación, la de recorte, las técnicas de reducción de ruido, técnicas de realce y contraste [6].

3.2.4.1. *Convertir la imagen RGB a escala de grises.*

La conversión del modelo RGB a una escala de grises de blanco y negro obedece a la respuesta cromática de los sensores en el ojo humano. En el ojo humano hay aproximadamente 6 millones de receptores de color (conos) y 134 millones de receptores en blanco y negro (bastones). En total se promedian 140 millones de receptores, mediante la fórmula siguiente [12]:

$$Gris = (Rojo * 0.299) + (verde * 0.587) + (Azul * 0.114)$$

Con estos porcentajes se puede obtener nuestra imagen en escala de grises, eliminando la información de matiz y saturación mientras se conserva la iluminación (fiura 3.6).

Figura 3.6: Pre-procesamiento de las imágenes de un ejemplar de la especie *Canthon (Canthon) humectus*: convertido a escala de grises.

3.2.4.2. *Binarización de la imagen.*

Esta operación consiste en tener únicamente dos valores 0 y 1 que serían blanco y negro esta técnica es muy útil cuando se desea identificar dos regiones la primera regularmente es el fondo y la segunda es la región de nuestro interés, el proceso se basa en una imagen que se encuentre en escala de grises para poder trabajar con su histograma y de aquí tomar un valor umbral para identificar lo que es blanco y lo que es negro [12] .Para esta entrada, el método calculará la tonalidad de gris de cada pixel de la imagen de entrada, si la tonalidad de gris supera el umbral ese pixel será verdadero, es decir, un pixel blanco en la imagen destino, en caso contrario, el pixel será negro en la imagen destino a continuación se muestra la función umbral (3.1) que nos da como resultado la imagen de la figura. 3.7.

$$f(x,y) = \begin{cases} 1 & \text{si f(x,y) } x > T \\ 0 & \text{si f(x,y) } x \leq T \end{cases} \tag{3.1}$$

Para binarizar la imagen se contempló el umbral de 0.5 por la cualidad de la imagen que su fondo es completamente blanco ver figura 3.7.

Figura 3.7: Pre-procesamiento de las imágenes de un ejemplar de la especie *Canthon (Canthon) humectus:* Imagen binarizada.

3.2.4.3. *Segmentación.*

La segmentación tiene como propósito subdividir una imagen en sus objetos constituyentes que la conforman, hasta un nivel de subdivisión en el que se aíslen las regiones u objetos de interés, cada objeto es considerado una región y esta debe permitir obtener información que busca el observador, en la mayor parte de los casos tener una solución correcta depende directamente de la segmentación, por lo que se tiene que poner el mayor esfuerzo en esta etapa [21]. En la segmentación una región es definida como la continuidad de píxeles que tienen características en común como intensidad, textura y vecindad, en conjunto deben tener la capacidad de definir objetos aislados. Existen dos áreas en las que se dividen los métodos de segmentación que son discontinuidad y similitud, se utilizó un método de cada área en este proyecto.

***Similitud*:** Se buscan zonas en donde los pixeles contengan valores similares, considerando criterios que se fijan de acuerdo al interés, este método se utilizara más adelante.

Crecimiento de regiones: Segmentación de imágenes mediante crecimiento de regiones, es un procedimiento que agrupa los píxeles o subregiones de la imagen en regiones mayores basándose en un criterio prefijado. La forma en que trabaja es aumentar pixeles, que inician con un conjunto de pixeles como parámetro de forma que a partir de estos se agreguen más píxeles, para que las regiones crezcan, cada uno de estos píxeles que estén próximos tienen que cumplir con propiedades similares como: nivel de gris textura, color, etc.

Para la segmentación en este trabajo se utilizó el crecimiento de regiones y se fijó como parámetro la continuidad de píxeles que tienen como característica en común 1 esto es todos los pixeles blancos conectados en la misma región ver figura 3.8.

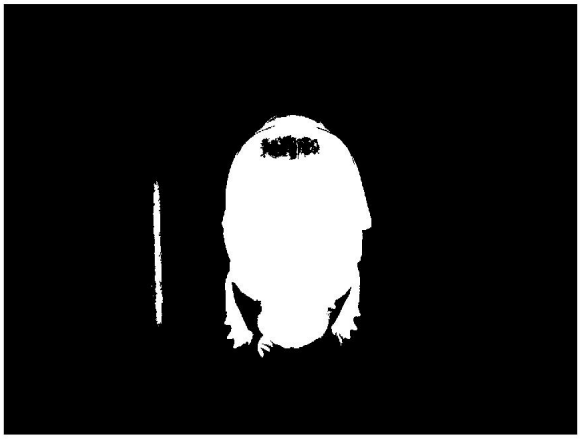

Figura 3.8: Pre-procesamiento de las imágenes de un ejemplar de la especie *Canthon (Canthon) humectus:* Imagen con regiones definidas.

3.2.4.4. *Eliminación de ruido.*

Se considera como ruido a los efectos que pueden presentarse en la imagen que no estén relacionados con el objeto que se quiere digitalizar y este ruido puede producirse por el efecto de captura, el medio en que fue tomada la foto, la intensidad de la luz, entre otros factores y para esto existen diferentes algoritmos para reducir el ruido como filtros lineales, filtro gaussiano por mencionar algunos [12].

En este trabajo se considera como ruido a las regiones menores a 100 pixeles, debido a que nos interesa conocer el borde de nuestro escarabajo, la métrica de referencia tiene 104 pixeles por lo que es posible eliminar las regiones menores a 100 pixeles garantizando que se queda con la referencia y el escarabajo ver figura 3.9.

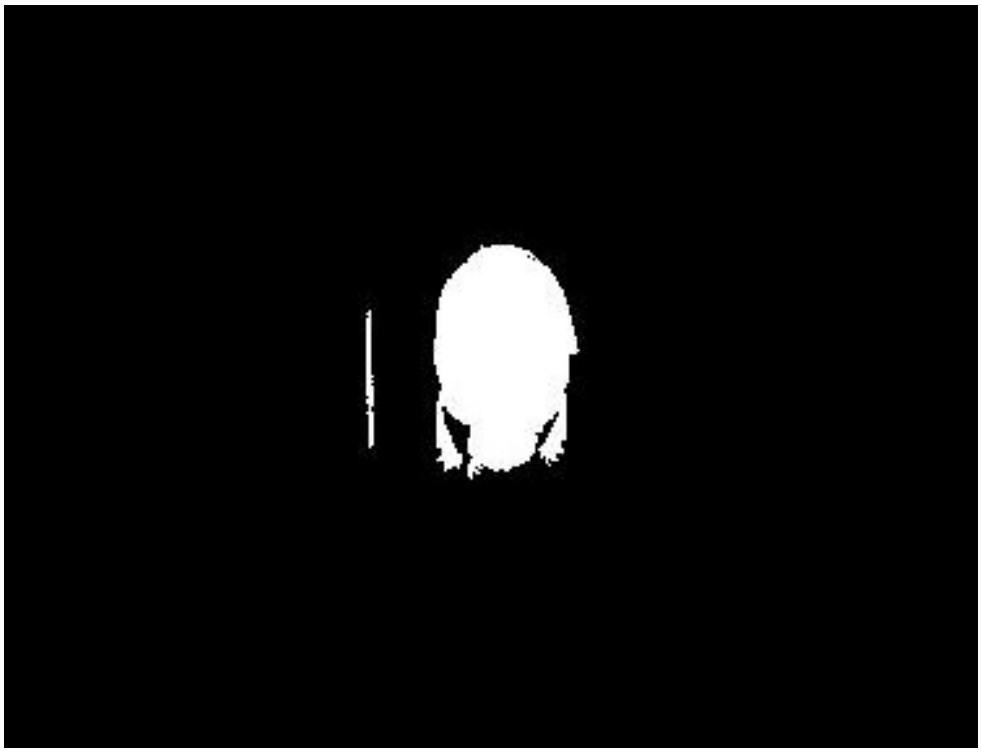

Figura 3.9: Pre-procesamiento de las imágenes de un ejemplar de la especie *Canthon (Canthon) humectus:* Imagen después de eliminar regiones menores de 100 pixeles

3.2.4.5. *Recorte de región de interés.*

La segmentación permite conocer la posición y el número de regiones que aparecen en la imagen digital, esto facilita el siguiente paso que es recortar la región de interés. La forma de recortar la imagen se realiza calculando el valor mínimo en la coordenada del eje X, el valor mínimo en el eje Y, el ancho y la altura de la región de interés ver figura 3.10, [12] la imagen es la que tiene por nombre A y tiene una región llamada B la cual se considera la región de interés ver figura 3.11 que es el resultado de recortar la región B de la imagen A.

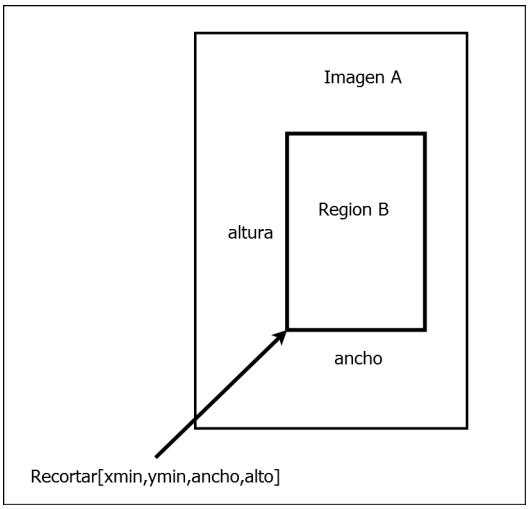

Figura 3.10: Ejemplo de imagen A que tiene una región B que será recortada

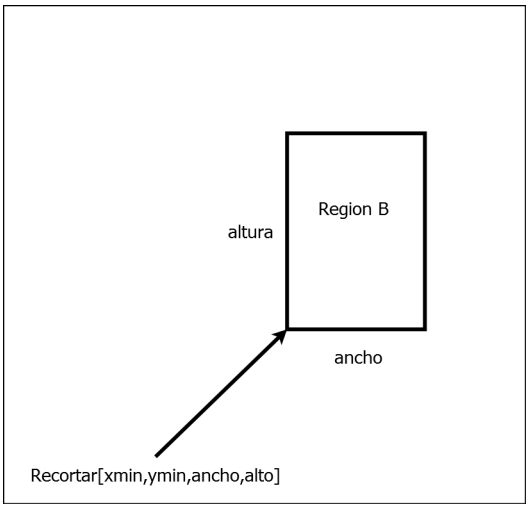

Figura 3.11: Región B recortada de la imagen A.

Para recortar el área de interés que es el cuerpo del escarabajo, se toma el parámetro de la región más grande, debido a que en este punto del procesamiento se tiene la referencia y el cuerpo del escarabajo como regiones de la imagen digital. Evidentemente la región del cuerpo del escarabajo es más grande que la región de la referencia, de esta manera se recorta el escarabajo rodeado por un margen de 15 pixeles, para no afectar el borde del cuerpo del escarabajo ver figura 3.12.

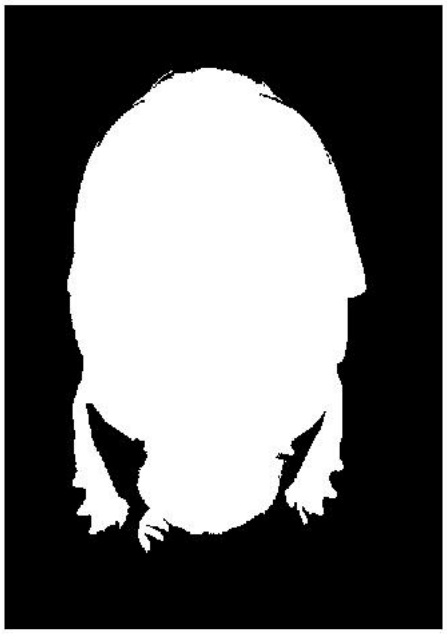

Figura 3.12: Pre-procesamiento de las imágenes de un ejemplar de la especie Canthon (Canthon) humectus:) Imagen recortada con un margen de 15 pixeles en cada lado.

En esta etapa del proyecto se utiliza la discontinuidad que es la otra área de la segmentación, que se mencionó anteriormente.
Discontinuidad: Divide la imagen con base a los cambios bruscos de nivel de intensidad de los pixeles, y el método utilizado es detección de bordes Sobel [21].

3.2.4.6. *Sobel.*

Es un operador para la deteccion de bordes y suavizado de la imagen de tal manera que se elimine el ruido de la imagen si es que lo tiene, este operador representa una aproximación imprecisa del gradiente de la imagen, pero suficiente para la deteccion de bordes. Esto debido a que solamente utiliza valores de intensidad en una region de 3X3 alrededor de cada píxcel[21], las máscaras de operaciones de Sobel se muestran en la figura 3.13.

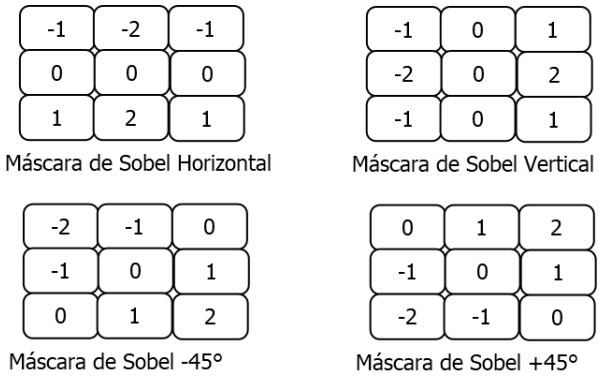

Figura 3.13: Máscaras de operaciones de Sobel.

Posteriormente se procede a detectar el borde de las imágenes aplicando la máscara de Sobel vertical. Así, se realiza la convolución de dicho operador (máscara Sobel vertical), con los puntos de la imagen y define el borde de la imagen ver figura 3.14.

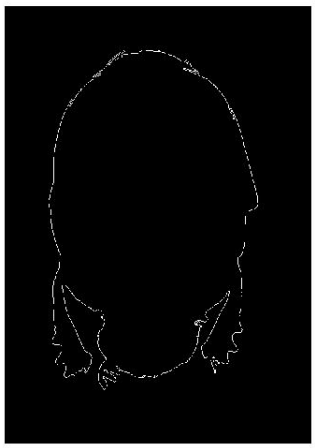

Figura 3.14: Imagen de un ejemplar de la especie *Canthon (Canthon) humectus* aplicación de mascara sobel.

3.2.4.7. *Dilatación (operación morfológica en imágenes binarias).*

Se basa en teoría de conjuntos al aplicar esta operación se conservan las principales características de la forma de los objetos, este operador y su composición, permite que las formas subyacentes sean identificadas y reconstruidas morfológicamente a partir de sus formas distorsionadas y ruidosas. Esta operación juega un papel muy importante en el pre-procesamiento de imágenes digitales, suprimiendo ruido o simplificando formas, también se puede destacar la estructura de los objetos ampliando o reduciendo el grosor de los bordes por mencionar un ejemplo [12]. Dada una imagen A, y un elemento estructural B, ambas imágenes binarias con fondo blanco, la dilatación de A por B se define como:

$$A \oplus B = X \quad talque \quad \tilde{B}_x \cap A \neq 0$$

La máscara morfológica de dilatación utilizada en el proyecto se muestra en la figura 3.15.

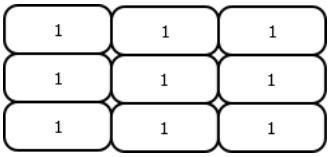

Figura 3.15: Máscara de dilatación.

Este operador resulta ser muy útil debido a que cuando se hace detección de bordes con Sobel se pierden algunos pixeles del borde de la imagen, este operador nos ayuda a recuperar el contorno cerrado de nuestro insecto ver figura 3.16.

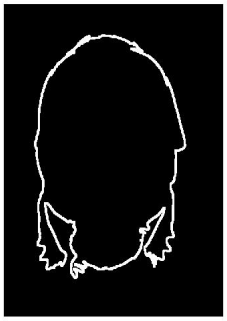

Figura 3.16: Imagen de un ejemplar de la especie *Canthon (Canthon) humectus)* aplicación de la operación de dilatación.

3.2.5. Extracción de características:

Las coordenadas (x, y) obtenidas en el paso anterior se utilizaron para calcular descriptores elípticos de Fourier. Después del pre-procesado de las imágenes, se aplica la transformada elíptica de Fourier, para trabajar con este método se considera el resultado de la imagen pre-procesada, que muestra claramente el borde del escarabajo, se emplea como un conjunto de datos importantes que ofrecen información importante para la clasificación. Realizar este método involucra un método intermedio conocido como, código de cadena de Freeman que realiza lo siguiente:

3.2.5.1. Método de Freeman.

El método de Freeman codifica un movimiento como un número entero entre 0 y 7. Estos códigos se muestran en la figura 3.17, donde 0 representa un movimiento hacia la derecha y 2 representa un movimiento hacia abajo [14].

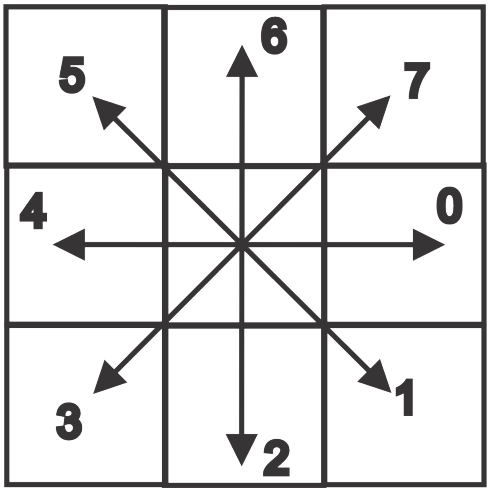

Figura 3.17: Código de Freeman.

Se considera la imagen binaria de la figura 3.18 . Hay una sola forma en esta imagen y se puede describir eligiendo un punto de partida y luego trazando los movimientos alrededor del contorno de la forma. La figura 3.18 muestra los movimientos alrededor del contorno de la forma indicando el punto de partida.

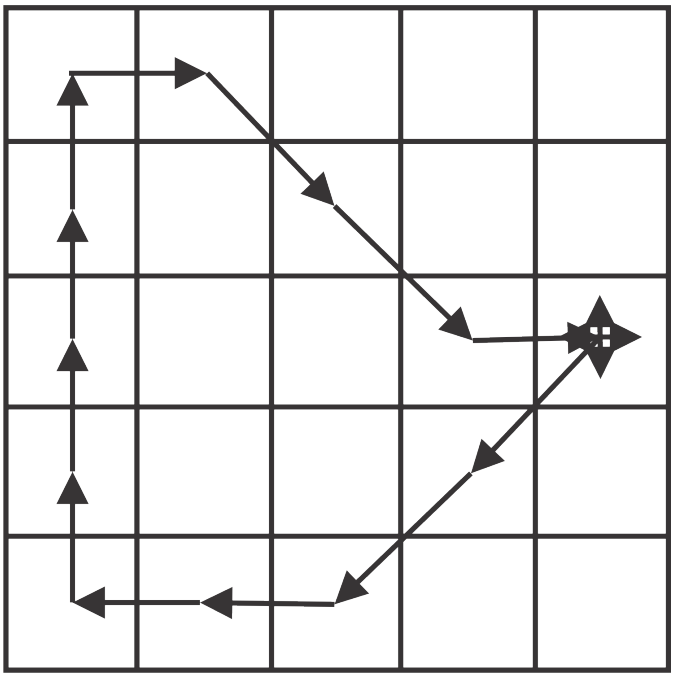

Figura 3.18: Ejemplo de cadena Freeman.

Puesto que cualquiera de los píxeles límite puede usarse como punto de partida, hay muchas maneras diferentes de representar una forma. Aquí hay una manera de describir la forma de la figura 3.18, usando el código de la figura 3.17.

Punto de partida: (4, 2) Cadena generada: [334466660110]

3.2.5.2. Serie de Fourier.

En general, una serie de Fourier es una forma de aproximar una función periódica, lo interesante de esta aproximación es que se puede hacer arbitrariamente buena, es decir el error entre la aproximación y la función real puede ser tan pequeño como se necesite [8]. La serie de Fourier $f(x)$ se describe por la ecuación (3.2), se observa que la serie de Fourier está parametrizada enteramente por sus coeficientes a_n y b_n.

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}[a_n cos(nx) + b_n sen(nx)] \tag{3.2}$$

Esta aproximación es arbitrariamente buena debido a que se aproxima bien a la función real en cualquier intervalo finito. Si se reemplaza la suma infinita por una finita, de $n = 1$ a k, esto se llama el k-ésimo armónico; Cuanto mayor sea el valor de k, mejor será la aproximación [8]. Observe que la función del armónico está escrita en color rojo en la leyenda del gráfico. Es increíble lo rápido que convergen a una buena aproximación ver figura 3.19.

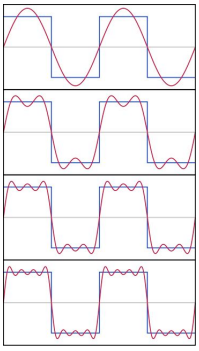

Figura 3.19: Describe hasta 6 armónicos para la serie de Fourier.

La cadena Freeman de codificación para el contorno de la que se habló anteriormente, se separa en sus proyecciones x e y. Por lo que se define x_p y y_p, la proyección de los primeros p enlaces en la cadena de codificación son la suma de las diferencias entre todos los enlaces anteriores. Las ecuaciones (3.3) y (3.4) nos dicen como calcular los coeficientes a_n y b_n.

$$X_p = \sum_{i=1}^{p} \triangle X_i \tag{3.3}$$

$$Y_p = \sum_{i=1}^{p} \triangle Y_i \tag{3.4}$$

Las ecuaciones anteriores son muy parecidas en la dirección x e y por tanto solo se resuelve para x y se tiene que resulta lo mismo para y, ecuación (3.5).

$$t_p = \sum_{i=1}^{p} \triangle t_i \tag{3.5}$$

Se considera la derivada temporal de x. Cuando se dice "tiempo" que en realidad significa la longitud de la cadena, por ejemplo, la contribución "tiempo"de cualquier enlace horizontal o vertical es $\sqrt{2}$. El "tiempo"de los p's enlaces es solo la suma de todos los tiempos de los enlaces anteriores ecuación (3.6).

$$x'(t) = \sum_{n=1}^{\infty} [\frac{2np_i}{T} b_n cos(\frac{2np_i t}{T}) - \frac{2np_i}{T} a_n sin(\frac{2np_i t}{T})] \tag{3.6}$$

De esta forma se implican los coeficientes que se quieren encontrar a_n y b_n. Entonces se puede resolver para a_n y b_n de la siguiente forma ecuaciónes (3.7) y (3.8):

$$a_n = \frac{T}{2n^2p_i^2}\sum_{p=1}^{k}\frac{\triangle X_p}{\triangle t_p}[cos(\frac{2np_it_p}{T}) - cos(\frac{2np_it_{p-1}}{T})] \tag{3.7}$$

$$b_n = \frac{T}{2n^2p_i^2}\sum_{p=1}^{k}\frac{\triangle X_p}{\triangle t_p}[sen(\frac{2np_it_p}{T}) - sen(\frac{2np_it_{p-1}}{T})] \tag{3.8}$$

Se sabe cómo calcular a_n y b_n. T es el periodo (el intervalo), n es el número del armónico que se observa, p es el índice del eslabón de la cadena, k es el número total de eslabones de la cadena, $\frac{x_p}{y_p}$ es la pendiente de cada eslabón, t_p y t_{p-1} son las longitudes de la cadena en el p's enlaces. Lo mismo sucede exactamente en la dirección y, por lo que se llama valor a_n en la dirección y, c_n y el valor b_n en la dirección y, d_n. Los coeficientes calculados se normalizaron para hacerlos invariantes al tamaño, a la rotación, y punto de inicio. Después de la normalización los primeros tres coeficientes del primer armónico son constantes y por lo tanto no considerados como atributos. El resto de coeficiente se utilizó como atributos en el proceso de clasificación.

3.2.5.3. Descriptor elíptico de Fourier.

Se utilizan estos coeficientes para modificar una elipse; después de todo, básicamente se quiere una elipse modificada que se aproxime a un contorno cerrado [36].

A continuación, se muestra la ecuación (3.9) que da el enésimo Descriptor Elíptico de Fourier y se parece mucho a la ecuación de una elipse.

$$\frac{(d_n^2 + c_n^2)X_n^2 + (a_n^2 + b_n^2)Y_n^2 - 2X_nY_n(a_nc_n + b_nd_n)}{(a_nd_n + b_nc_n)^2} = 1 \tag{3.9}$$

3.2.5.4. Medidas estadísticas.

Por otra parte, también se calcularon medidas estadísticas, considerando las coordenadas x e y, estas fueron: a) media, b) mediana, c) máximo, d) mínimo, e) suma, f) desviación estándar, g) varianza, h) curtosis, i) oblicuidad.

a). **Media:** Es el resultado de sumar todos los valores de los datos y dividir este valor entre el número total de los datos y se denota con ecuación (3.10) [35].

$$\bar{X} = \sum_{i=1}^{n}\frac{x_i}{n} = \frac{x_1, x_2, ..., x_n}{n} \tag{3.10}$$

b). **Mediana:** Es el valor que ocupa el lugar de enmedio cuando los datos están ordenados de menor a mayor. Si el número de los datos es impar la media es el número que ocupa la posición central. Si el número de datos es par la mediana es la media de los dos valores que ocupan las posiciones centrales, y se denota con la ecuación (3.11) [35].

$$\bar{X} = \begin{cases} x_{\frac{n+1}{2}} & \text{si n es impar} \\ \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}) & \text{si n es par} \end{cases} \tag{3.11}$$

c). **Máximo:** Es el máximo valor que puede tomar el conjunto de datos [18].

d). **Mínimo:** Es el valor mínimo que puede tomar el conjunto de datos [18].

e). **Suma:** Es la suma del vector de valores del conjunto de datos.

f). **Varianza:** Es la media aritmética del cuadrado de las desviaciones respecto a la media [35]: y se denota con la ecuación (3.12).

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{3.12}$$

g). **Desviación estándar:** Es una medida de dispersión que mide cuanto se separan los datos de la media [35] y se denota con s, es la raíz cuadrada positiva de s^2, es decir ecuación (3.13).

$$s = \sqrt{s^2} \tag{3.13}$$

h). **Curtosis:** También es conocido como concentración central, mide el grado de apuntamiento o aplastamiento de una distribución normal o gaussiana, una mayor concentración de datos respecto al promedio harán que la forma sea más alargada (Leptocúrtica) y si los datos tienen una mayor dispersión la forma será más aplastada (Platicúrtica) como se muestra en la figura 3.20. y se denota en la siguiente ecuación (3.14) [18].

$$k = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{ns^4} - 3 \tag{3.14}$$

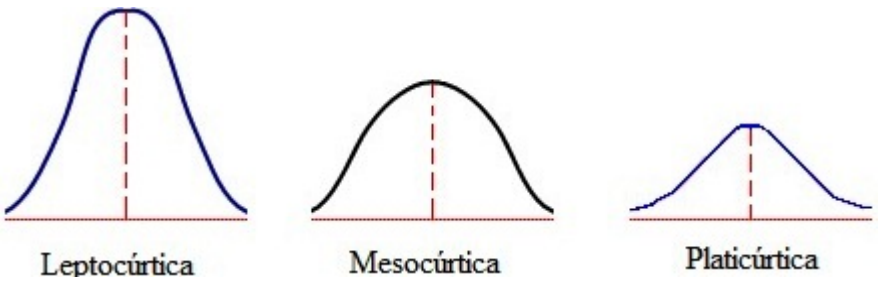

Figura 3.20: Concentraciones centrales Leptocúrtica, Mesocúrtica y Platicúrtica.

i). **Oblicuidad:** Es la medida de que tan simétrica es una distribución alrededor de su media y se denota por la ecuación (3.15) [18].

$$\vartheta = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^3}{\vartheta} \tag{3.15}$$

De forma que se tienen 18 medidas estadísticas, 9 por X y 9 más por Y.

3.2.5.5. Vector característico.

En resumen, el número de características para cada imagen digital son: 18 características estadísticas, más las características obtenidas del número de k's armónicos que determinan la mejor aproximación, para este trabajo se emplearon 8 números armónicos que ofrecen una buena aproximación, dando como resultado 29 características, con esto se obtiene una cadena de 47 características para cada imagen digital que se muestran a continuación:

Las características estadísticas se muestran en la tabla 3.2

Tabla 3.2: Características estadísticas.

Medidas estadísticas para el eje x_n	Medidas estadísticas para el eje y_n
Media:	Media:
Mediana:	Mediana:
Máximo:	Máximo:
Mínimo:	Mínimo:
Suma:	Suma:
Varianza:	Varianza:
Desviación estándar:	Desviación estándar:
Curtosis:	Curtosis:
Oblicuidad:	Oblicuidad:

Las características resultantes de los 8 armónicos se muestran en la tabla 3.3:

Tabla 3.3: Las características resultantes de los 8 armónicos.

a_n	b_n	c_n	d_n
const	const	const	$1d_n$
$2a_n$	$2b_n$	$2c_n$	$2d_n$
$3a_n$	$3b_n$	$3c_n$	$3d_n$
$4a_n$	$4b_n$	$4c_n$	$4d_n$
$5a_n$	$5b_n$	$5c_n$	$5d_n$
$6a_n$	$6b_n$	$6c_n$	$6d_n$
$7a_n$	$7b_n$	$7c_n$	$7d_n$
$8a_n$	$8b_n$	$8c_n$	$8d_n$

Los primeros tres coeficientes del primer armónico son constantes para todas las imágenes y por lo tanto no considerados como atributos.

3.2.6. Clasificadores:

Clasificar consiste asignar un objeto a una de las clases existentes, las clases tienen que estar bien definidas, por características, en las que se contempla la forma y tamaño. Se tienen que definir fronteras para diferentes clases, estas fronteras se calculan mediante el proceso de entrenamiento, en el que se consideran las características de un conjunto de objetos que pertenecen a la misma clase. Clasificar un objeto desconocido es buscar el mayor grado de pertenencia de acuerdo a las características que definen a una clase a continuación se mencionan tres clasificadores:

3.2.6.1. Support Vector Machines (Máquinas de Soporte Vectorial)

SVM por su nombre en inglés *Support Vector Machines:* Es una herramienta de clasificación poderosa, que mapea los puntos de entrada a un espacio de características de una dimensión mayor, por ejemplo si los puntos de entrada están en R^2 estos pueden ser mapeados por la *SVM* a R^3. Las *SVM* determinan un hiperplano que sirve para separar clases de forma que se maximice este margen ver figura 3.21 [2].

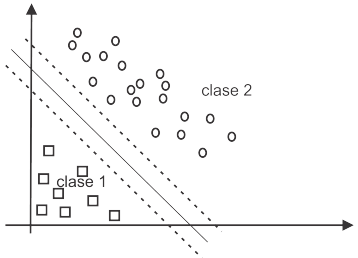

Figura 3.21: La frontera de decisión debe estar tan lejos de los datos de ambas clases como sea posible.

1). **Modelo linealmente separable:** Un conjunto de vectores etiquetados nos ayuda en determinar cuándo una muestra pertenece a una de las clases existentes [2]. El objetivo es encontrar un único hiperplano que proporcione el mayor margen de separación entre las clases con el mínimo error como se muestra en la figura 3.22. Se dice que un conjunto S es linealmente separable si existe (w, b) para la ecuación (3.16):

$$\begin{cases} (w * z_{i+b}) \geq 1, & y_i = 1 \\ (w * z_{i+b}) \leq 1, & y_i = -1 \end{cases} \tag{3.16}$$

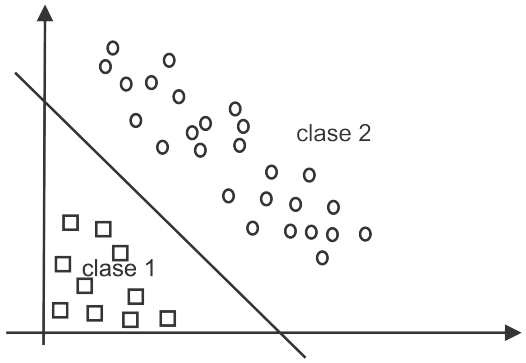

Figura 3.22: Objetos linealmente separables.

2). **Modelo no linealmente separable:** En estos casos el conjunto de vectores no es posible separarlos de una clase de otra con un hiperplano como se puede notar en la figura 3.23 y en esta situación se dice que el conjunto es no separable.

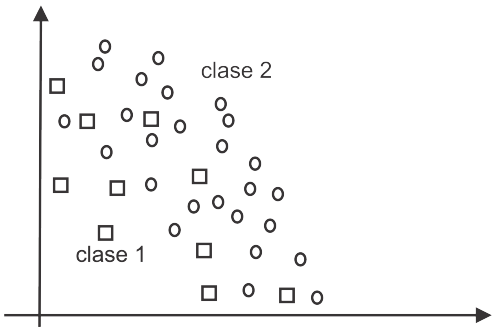

Figura 3.23: Caso no linealmente separable.

En el caso de que las clases no sean linealmente separables lo que se hace es proyectar los casos a un espacio de dimensión superior donde sí sean linealmente separables como se muestra en la figura 3.24.

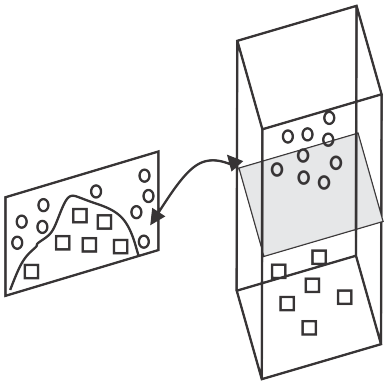

Figura 3.24: Idea del uso de un kernel para transformación del espacio de los datos.

Existen las funciones Kernel que ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de la máquina de aprendizaje lineal. Es decir, se mapea el espacio de entradas X a un nuevo espacio de características de mayor dimensionalidad, para realizar este proceso se introducen algunas variables no negativas $\varepsilon_i \geq 0$. De forma que la ecuación (3.16) se modifique la ecuación (3.17) [2].

$$\Big\{ y_i(w * z_i + b) \geq 1 - \varepsilon_i, \quad i = 1, ..., l \tag{3.17}$$

La función matemática utilizada para la transformación se conoce como función kernel y se pueden mencionar los siguientes tipos de kernel:

a). **Lineal.**

b). **Polinómico.**

c). **Función de base radial (RBF).**

d). **Sigmoide.**

Ventajas:

1). El entrenamiento es fácil.

2). No hay óptimo local.

3). Se escalan relativamente bien para datos en espacios dimensionales altos.

4). El compromiso entre la complejidad del clasificador y el error puede ser controlado explícitamente.

5). Datos no tradicionales como cadenas de caracteres y árboles pueden ser usados como entrada a la SVM.

Desventaja:

1). Se necesita una buena función kernel, eficiente para sintonizar los parámetros de inicialización de la *SVM*.

2). En este trabajo se emplea el kernel lineal utilizando la estrategia uno contra el resto, consiste en entrenar un solo clasificador por clase, con las muestras de esa clase como muestras positivas y todas las demás como muestras negativas. Esta estrategia reduce el problema de clasificación multi clase a múltiples problemas de clasificación binaria.

3.2.6.2. Naive Bayes (Bayes Ingenuo)

Es un clasificador probabilístico basado en el teorema de Bayes, se debe calcular que un elemento pertenece a una clase determinada, no requiere de muchos datos de entrada, es altamente escalable y ofrece buenos resultados. Un aspecto importante es que asume independencia entre sus características, considera que cada característica contribuye independientemente a la probabilidad de pertenecer a una clase. Se muestran los siguientes pasos para realizar una clasificación con *Naive Bayes* [20].

1). **Dado un conjunto de elementos para ser clasificados, representados por n característica $X = (x_1, x_2, \ldots, x_n)$ se deben calcular las probabilidades.**

$$p(C_1|x_1, x_2, ..., x_n), p(C_2|x_1, x_2, ..., x_n), ..., p(C_k|x_1, x_2, ..., x_n) \tag{3.18}$$

De que el elemento $X = (x_1, x_2, ..., x_n)$ pertenezca a una de las k clases C_i.

2). **Asignar el elemento a la clase con mayor probabilidad.**

3). **Para calcular las probabilidades se utiliza el Teorema de Bayes** (3.19)**.**

$$p(C_k|x) = \frac{p(C_k)p(X|C_k)}{p(x)} \tag{3.19}$$

Donde:

a). $p(C_k|X)$ Probabilidad que dado un cierto $X = (x_1, x_2, ..., x_n)$ pertenezca a la clase C_k.

b). $p(C_k)$ Probabilidad de que pertenezca a la clase C_k.

c). $p(X|C_k)$Probabilidad que dada una cierta clase C_k sus características sean $X = (x_1, x_2, ..., x_n)$.

d). $p(X)$ Probabilidad de que sea el conjunto de características $= (x_1, x_2, ..., x_n)$.

e). $p(C_k|X)$ Es la probabilidad a **posteriori**, y es la que se quiere calcular.

f). $p(C_k)$ Es la probabilidad a **priori** y se calcula como el porcentaje de cada clase.

g). $p(X)$ Es la evidencia y al ser la probabilidad de escoger un cierto elemento X, se puede asumir que es igual a todas las constantes para $\frac{1}{n}$.

h). $p(X|C_k)$ Es la verosimilitud y puede ser reescrita en base a cada una de las características por separado ecuación (3.20).

$$p(C_k|x_1, x_2, ..., x_n) = p(C_k)p(x_1|C_k)p(x_2|C_k)p(x_3|C_k), ..., = p(C_k)\prod_{i=1}^{k} p(x_i|C_k) \tag{3.20}$$

Se asume que las variables son independientes entre si

i). • El modelo de clasificación queda como se muestra en la ecuación (3.21).

$$\widetilde{y} = \underset{k\varepsilon(1...k)}{argmax}\{p(C_k)\prod_{i=1}^{k} p(x_i|C_k)\} \tag{3.21}$$

4). **Ventajas:**

1). **Es fácil de implementar.**

2). **Obtiene buenos resultados en gran parte de los casos.**

5). **Desventajas:**

1). **Asumir que las variables tienen independencia condicional respecto a la clase lleva a una falta de precisión.**

2). **En la práctica, existen dependencias entre las variables.**

En este trabajo se empleó el clasificador *Naive Bayes* simple, que se describe anteriormente, para calcular las probabilidades se supone que las variables X_i cumplen una distribución normal.

3.2.6.3. Random forest (Bosques aleatorios)

El bosque aleatorio es una herramienta de conjunto que toma un subconjunto de observaciones y un subconjunto de variables para construir árboles de decisión. Construye múltiples árboles de decisión y los fusiona para obtener una predicción más precisa y estable. Al crear varios árboles de decisión como un solo modelo, cada árbol tiene el mismo peso en el proceso final de toma de decisiones, es por eso que la mayoría determina el resultado [11].

En *Random Forest*, se generan varios árboles en lugar de solo un árbol de decisión. Para clasificar un nuevo objeto basado en atributos, cada árbol da una clasificación y se dice que el árbol voto para esa clase. El bosque elige la clasificación que tiene más votos sobre todos los árboles del bosque. Cada árbol se crea de la siguiente manera:

a). Se considera que el número de casos en el conjunto de entrenamiento es N. Entonces, la muestra de estos N casos se toma al azar, pero con reemplazo. Esta muestra será el conjunto de entrenamiento para generar el árbol.

b). Si hay M variables de entrada, se especifica un número $m \leq M$ de tal manera que en cada nodo se seleccionan al aleatoriamente al azar m variables de las M. La mejor división de estos m se utiliza para dividir el nodo. El valor de m se mantiene constante mientras se generan árboles, Brieman sugiere tres valores posibles para m: $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$, y $2\sqrt{m}$ [4].

c). Cada árbol se genera en la mayor extensión posible y no hay poda.

d). Para predecir nuevos datos se agregan las predicciones de los n árboles del bosque con votos de la mayoría para la clasificación y el promedio para la regresión.

e). **Ventajas:**

1). Puede manejar miles de variables de entrada e identificar las variables más significativas por lo que se considera como uno de los métodos de reducción de la dimensionalidad.

2). Para muchos conjuntos de datos produce un resultado certero.

3). Poder manejar cientos de variables entrantes sin excluir ninguna.

4). Dar estimados de qué variables son importantes en la clasificación.

5). Tener un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida.

6). Calcular los prototipos que dan información sobre la relación entre las variables y la clasificación.

7). Ofrecer un método experimental para detectar las interacciones de las variables.

f). **Desventajas:**

1). Se tiene muy poco control sobre lo que hace el modelo.

2). Si los datos contienen grupos de atributos correlacionados de relevancia similar para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes.

En este trabajo se utilizó *Random Forest* con la cantidad de 100 árboles de decisión, utilizando 7 características de 47, la elección de estas 7 características es de manera aleatoria y se pueden repetir las muestras en diferentes árboles.

Capítulo 4

Diseño e implementación del sistema.

Se describe el diseño y la implementación del sistema computacional de identificación de escarabajos utilizando imágenes digitales, de una manera detallada a través de diagramas. En la figura. 4.1, muestra el diseño de forma general como se desarrolló el proyecto, se conforma por 5 módulos principales 1) Adquisición de imágenes, 2) Pre-procesamiento, 3) Extracción de características, 4) Clasificación, 5) Identificación, en cada uno de ellos se desarrollan sub-módulos con objetivos dependientes para conseguir un objetivo principal, que a continuación se desarrollan.

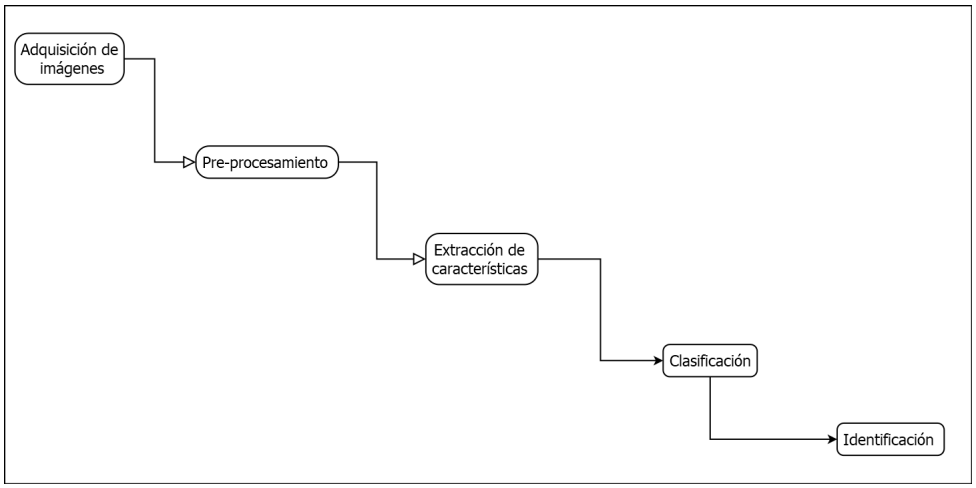

Figura 4.1: Diseño general de Reconocimiento automático de escarabajos (Insecta: Coleoptera) usando imágenes digitales.

4.1. Diagrama de adquisición de imágenes digitales.

En la figura. 4.2, se desglosa el módulo Adquisición de imágenes, que se describe en el **Capitulo 3.2.2 Adquisición de imágenes.**

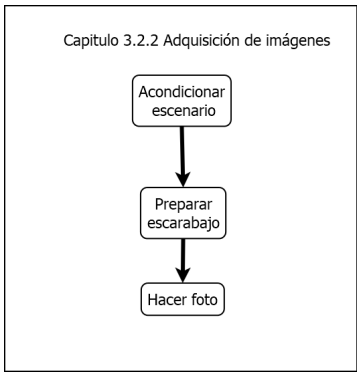

Figura 4.2: Diagrama de Adquisición de imágenes.

4.2. Diagrama de pre-procesamiento de imágenes digitales.

En la figura. 4.3, se muestra el pre-procesamiento por el que se someten las imágenes obtenidas, para posteriormente extraer sus características, en el **Capitulo 3.2.4 Pre-procesamiento.** se describe cada una de ellas.

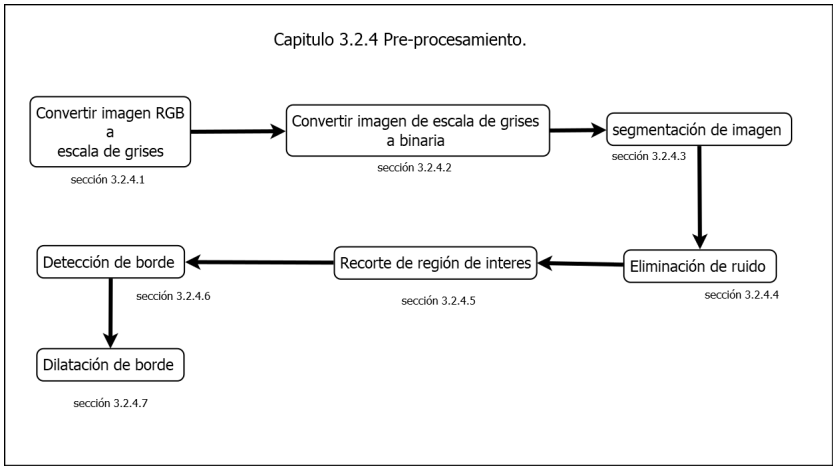

Figura 4.3: Diagrama de Pre-procesamiento de imágenes.

4.3. Diagrama de extracción de características de imágenes digitales.

En la figura. 4.4, se muestra las dos formas de extracción de características para formar un vector característico de cada foto.

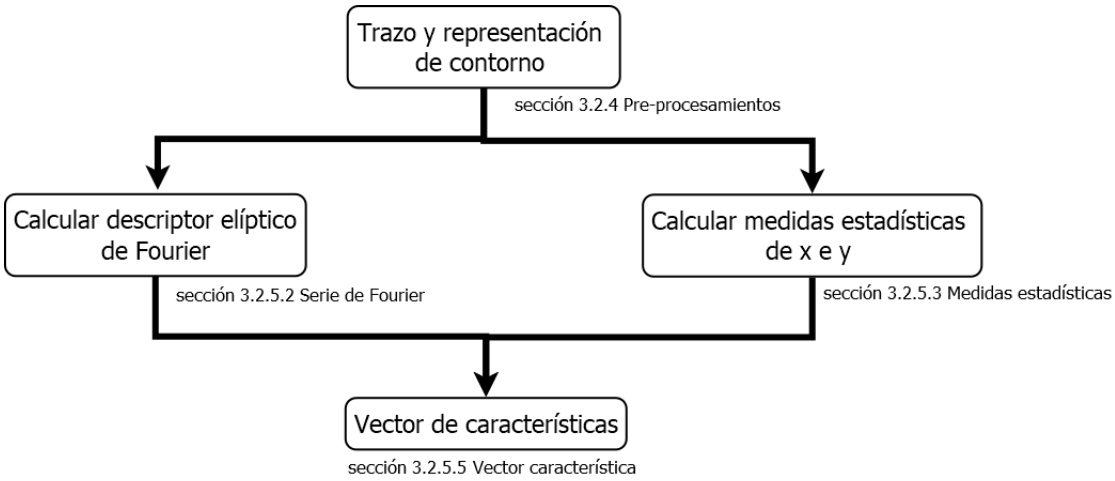

Figura 4.4: Diagrama de extracción de características.

4.4. Diagrama de entrenamiento de clasificadores.

Con los vectores característicos del paso anterior se forma un archivo de entrenamiento para los clasificadores *SMV*, *Naive Bayes* y *Random Forest*, ver figura. 4.5.

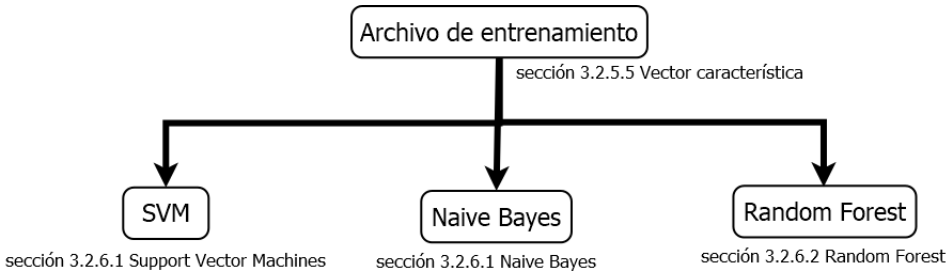

Figura 4.5: Diagrama de entrenamiento de los clasificadores *SMV*, *Naive Bayes* y *Random Forest*.

4.5. Diagrama de validación para determinar que la especie pertenece o no a una clase del clasificador.

Con los clasificadores entrenados se puede ingresar un nuevo ejemplar de validación para determinar que la especie pertenece ó no es posible identificar la muestra con los datos característicos, ver figura. 4.6.

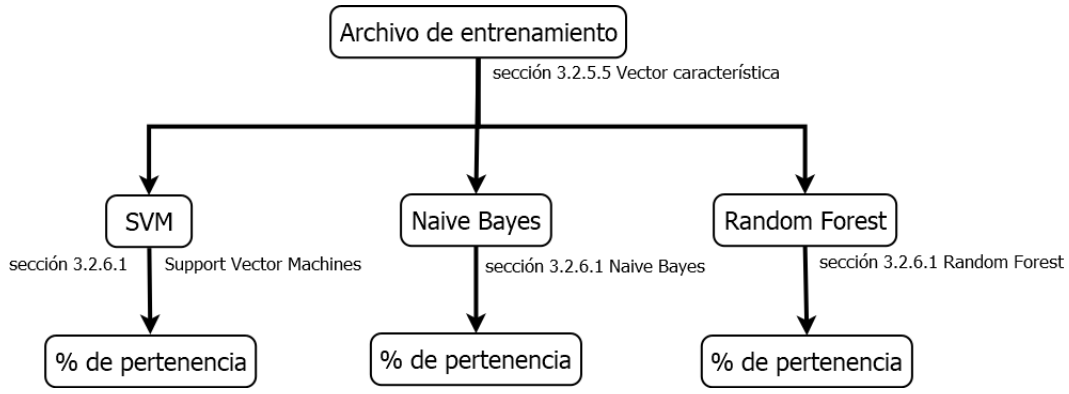

Figura 4.6: Diagrama para determinar el grado de pertenencia de un ejemplar a una especie conocida.

4.6. Diagrama de validación del método propuesto

El clasificador *Random Forest* presenta una exactitud de 0.87 que es mayor a la de los otros dos clasificadores por esta razón el método propuesto trabaja con *Random Forest* y el diagrama , ver figura 4.7 presenta la forma de validación donde se ingresa una forma al clasificador y se determina el porcentaje de pertenencia. Si tenemos nueve especies entonces la probabilidad de pertenencia es $z = \frac{1}{9} = 0.11$. Si al ingresar una imagen al sistema de identificación de escarabajos su porcentaje de pertenencia es mayor o igual a 0.11, se dice que pertenece a una especia de la familia en caso contrario no pertenece a ninguna especie de la familia.

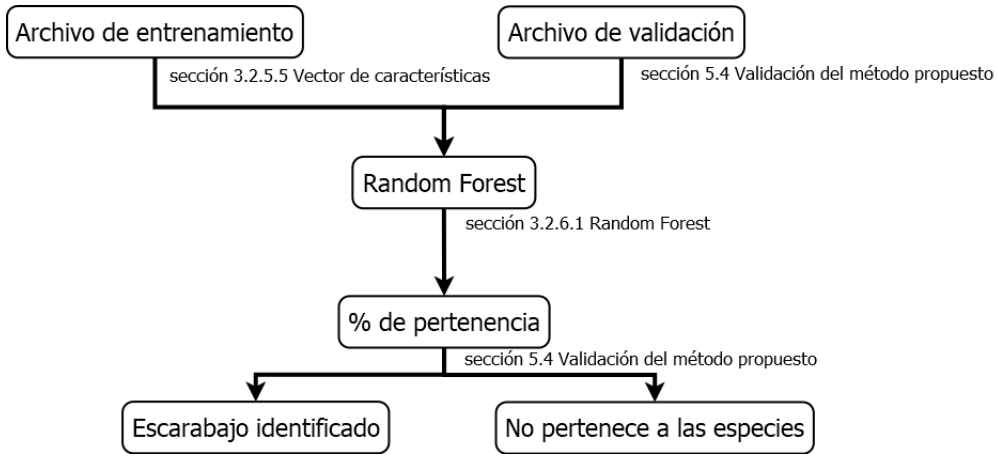

Figura 4.7: Diagrama de validación del método propuesto.

Capítulo 5

Pruebas.

En este capítulo se presentan los resultados de los experimentos realizados calculando características del borde del cuerpo de los escarabajos con la transformada elíptica de Fourier y medidas estadísticas, para posteriormente clasificar usando los algoritmos *Random Forest, SVM* y *Naive Bayes*. Primero se presenta la clasificación utilizando la transformada elíptica de Fourier con el número de armónicos 2, 4, 8 y 16. Con 30 , 50 y 100 puntos equidistantes, considerados del borde de cada escarabajo. Se consideran los diferentes tamaños de números de armónicos y puntos equidistantes para comparar y encontrar el menor número que contenga suficiente información que permita distinguir entre escarabajos y otras formas. Ademas se comparan diferentes cantidades de puntos equidistantes para utilizar medidas estadísticas que permitan identificar el menor número de puntos para distinguir a los escarabajos. Tambien se combinan estos dos métodos, considerando el menor número de atributos que aportan información para identificar cada escarabajo. Por último, se genera un modelo para cada individuo y se calcula la probabilidad de pertenencia a la clase.

Las imágenes se analizaron en posición vertical con la masa antenal en la parte inferior izquierda, considerando esta orientación se calcularon las medidas estadísticas mencionadas en el capítulo 3. Todos los experimentos se realizaron en la misma maquina con las siguientes características, procesador Intel Core i3-3517U a 2.40 GHz con 8 GB de memoria ram. Se utilizaron los clasificadores implementados en WEKA 3.7 para todos nuestros experimentos.

Para realizar los experimentos se tomaron en cuenta únicamente las especies de la familia Scarabaeidae, que tienen 5 ejemplares en cuenta o más para un total de 53 imágenes como se muestra en la Tabla 5.1. Misma que se recolectaron en un año, como se mencionó en el capítulo 3.

Tabla 5.1: Ejemplares de escarabajos empleados en el estudio, se muestran los nombres que les corresponden dentro de cada una de las categorías taxonómicas a partir de familia.

Familia	Subfamilia	Tribu	Género, subgénero y especie	No. ejemplares
Scarabaeidae	Scarabaeinae	Scarabaeini	*Canthon (Canthon) humectus*	6
			Canthon (Canthon) indigaceus	6
			Deltochilum gibbosum sublaeve	6
			Deltochilum tumidum	5
		Coprini	*Dichotomius colonicus*	6
			Ateuchus rodriguezi	6
		Phanaeini	*Phanaeus mexicanus*	7
		Onthophagini	*Digitonthophagus gazella*	6
Hybosoridae	Hybosorinae		*Hybosorus illigeri*	5

5.1. Clasificación de contorno con transformada elíptica de Fourier

La cantidad de atributos varia dependiendo del número de armónicos que se utilizá y éstos se muestran en la tabla 5.2. Después de aplicar el método de pre-procesamiento, se entrenaron tres clasificadores *Random Forest, SVM* y *Naive Bayes*, con los datos obtenidos al aplicar el método transformada Elíptica de Fourier, con 2, 4, 8, 16 armónicos y 30, 50, 100 puntos equidistantes ver Tabla 5.3.

Tabla 5.2: Atributos dependiendo del número de armónicos.

No.armonico	Total de características
2	5
4	13
8	29
16	61
32	125

Tabla 5.3: Resultados de la exactitud para las imágenes de la familia Scarabaeidae caracterizadas con los coeficientes de los armónicos de la transformada elíptica de Fourier.

No.armonico	No.puntos	Random Forest	SVM	Naive Bayes
2	30	0.47	0.12	0.47
2	50	0.47	0.12	0.47
2	100	0.47	0.12	0.47
4	30	0.41	0.22	0.52
4	50	0.41	0.22	0.52
4	100	0.41	0.22	0.52
8	30	0.35	0.39	0.41
8	50	0.35	0.39	0.41
8	100	0.35	0.39	0.41
16	30	0.18	0.29	0.29
16	50	0.18	0.29	0.29
16	100	0.18	0.29	0.29

Para realizar este experimento solo se utilizó los datos de la transformada elíptica de Fourier. Cabe destacar que en este experimento la cantidad de puntos equidistantes que se utilizan para aplicar el método Elíptico de Fourier no influye en la exactitud, pero la cantidad de armónicos si afecta el grado de exactitud que oscila dependiendo del número de armónicos que se ocupe como se ve en la figura. 5.1. Se puede observar que el mayor grado de exactitud es de 0.52 y se obtiene con cuatro números de armónicos, que generan 13 atributos.

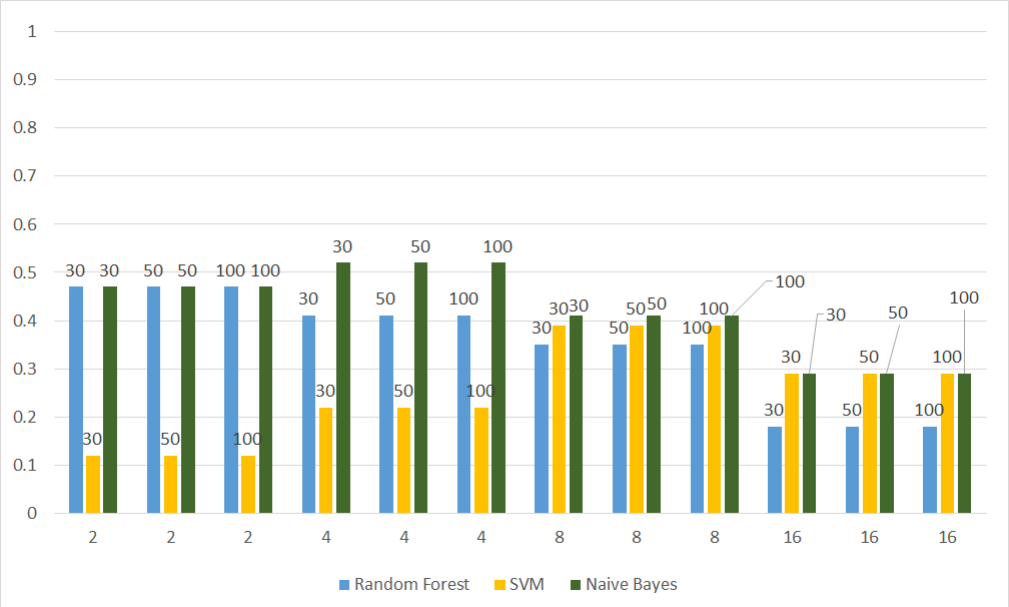

Figura 5.1: Clasificación con transformada Eliptica de Fourier.

En cada clasificador el comportamiento es similar y los tres tienen la misma tendencia, logran la mayor exactitud con cuatro números armónicos, sin importar el número de puntos equidistantes que en este experimento se consideraron 30, 50 y 100 puntos.

5.2. Clasificación de contorno con medidas estadísticas

En esta sección se entrenaron tres clasificadores *Random Forest*, *SVM* y *Naive Bayes*, los mismos del experimento anterior (Clasificación de contorno con transformada elíptica de Fourier), pero utilizando solo puntos equidistantes que son los necesarios para obtener las siguientes medidas estadísticas: Media, mediana, máximo, mínimo, suma, varianza, desviación estándar, curtosis y oblicuidad. Se consideran cantidades diferentes de puntos equidistantes que son 30, 50, 100, 250 y 500. La cantidad de atributos siempre son 18 a diferencia del experimento anterior, donde la cantidad de atributos varía dependiendo del número de armónicos que se considere, Tabla 5.4.

Tabla 5.4: Resultados de la exactitud para las imágenes de la familia Scarabaeidae caracterizadas con medidas estadísticas.

No. puntos	Random Forest	SVM	Naive Bayes
30	0.83	0.77	0.83
50	0.83	0.83	0.75
100	0.83	0.83	**0.85**
250	0.81	0.79	0.79
500	0.79	0.83	0.83

En este experimento solo se utilizaron medidas estadísticas, se puede notar que cada clasificador tiene un comportamiento diferente dependiendo de la cantidad de puntos considerados para calcular las medidas estadísticas. Es importante mencionar que los tres clasificadores obtienen una mayor exactitud con 100 puntos, *Random Forest* y *SVM* obtiene una exactitud de 0.83 y *Naive Bayes* de 0.85 como se muestra en la figura 5.2

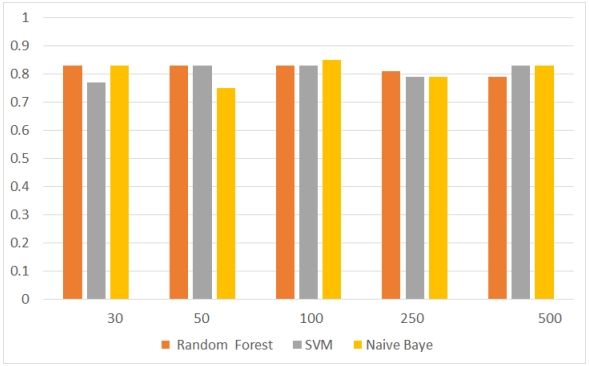

Figura 5.2: Medidas estadísticas

5.3. Clasificación de contorno con transformada elíptica de Fourier y medidas estadísticas

En este experimento se consideran los resultados obtenidos de los dos experimentos previos, donde se observa que la cantidad de puntos equidistantes para la transformada Elíptica de Fourier ofrece los mismos resultados, por esta razón se experimenta únicamente con 30 puntos equidistantes, estos puntos son considerados para obtener los 18 atributos que son fijos para las medidas estadísticas y los atributos que se generan dependiendo del número de armónicos para la transformada Elíptica de Fourier, con esto se garantiza un menor tiempo de procesamiento Tabla 5.5.

Tabla 5.5: Resultados de la exactitud para las imágenes de la familia Scarabaeidae caracterizadas con los coeficientes de los armónicos de la transformada elíptica de Fourier.

No. puntos Estadistica	No.puntos armónicos	No.puntos descriptor	Random Forest	SVM	Naive Bayes
30	2	30	0.85	0.81	0.79
30	4	30	0.87	0.83	0.87
30	8	30	0.89	0.81	0.83
30	16	30	0.81	0.75	0.79

En este experimento se combinan los experimentos anteriores manteniendo los 4 números armónicos y los resultados alcanzan la exactitud de 0.87 para *Random Forest y Naive Bayes* y 0.83 para *SVM* que es mayor a la que obtienen de forma independiente, se puede observar que el comportamiento de cada clasificador es muy diferente pero mantienen una relación con los experimentos previos pues la exactitud oscila dependiendo de la cantidad de los números armónicos, *Random Forest* y *Naive Bayes* obtiene una mayor exactitud que *SVM* con 8 números armónicos lo que nos indica que se tiene una mayor exactitud con un mayor número de armónicos esto implica un mayor número de atributos respecto a los dos experimentos anteriores figura 5.3.

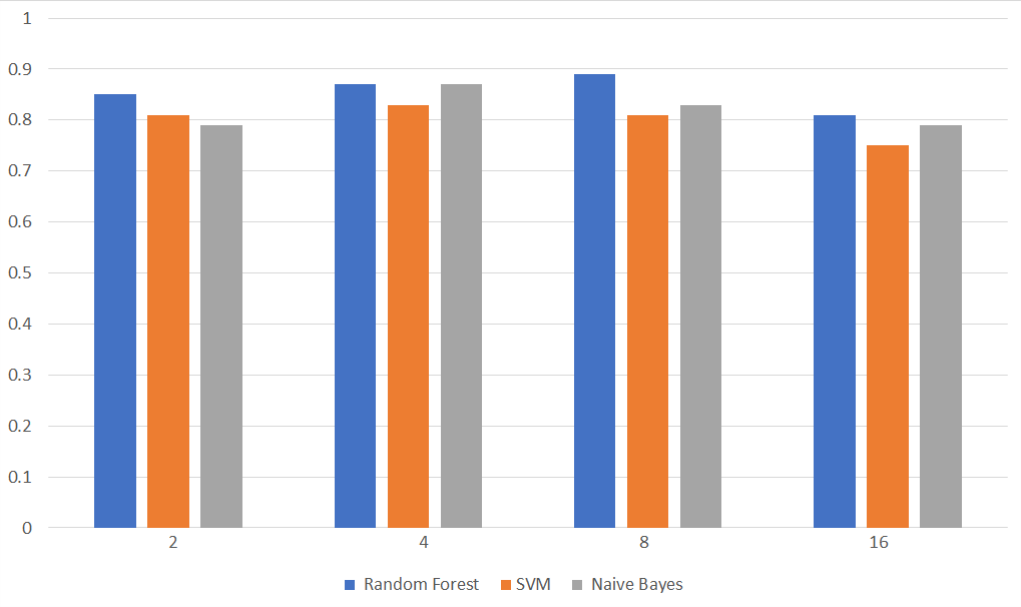

Figura 5.3: Combinación de tranformada elíptica de Fourier y datos estadísticos.

En la Tabla 5.6 se muestra la matriz de confusión para estos resultados donde puede observarse que el mayor número de errores se da entre las especies *Deltochilum gibbosum sublaeve y Deltochilum tumidum* que de acuerdo a la Tabla 5.1 pertenecen a la misma tribu.

Tabla 5.6: Resultados de la exactitud obtenida para las imágenes de la familia Scarabaeidae caracterizadas con los coeficientes de los armónicos de la transformada elíptica de Fourier.

a	b	c	d	e	f	g	h	i	<—classified as
6	0	0	0	0	0	0	0	0	a= *Ateuchus rodriguezi*
0	6	0	0	0	0	0	0	0	b= *Canthon(Canthon)humectus*
0	0	6	0	0	0	0	0	0	c= *Canthon(Canthon)indigaceus*
0	0	0	5	1	0	0	0	0	d= *Delthochilum gibbosum sublaeve*
0	0	0	3	2	0	0	0	0	e= *Deltochilum tumidum*
0	0	0	0	0	6	0	0	0	f= *Dichotomius colonicus*
0	0	0	0	0	0	6	0	0	g= *Digitonthophagus gazella*
0	0	0	0	0	1	0	6	0	h= *Phanaeus mexicanus*
0	0	0	0	0	0	0	0	6	h= *Hybosorus illigeri*

En la figura 5.4 se pueden observar un ejemplar de cada especie y ver que de manera natural tienen ciertos atributos en común.

Deltochilum Tumidum

Delthochilum Gibbosum Sublaeve

Figura 5.4: Especies *Deltochilum gibbosum sublaeve vs Deltochilum tumidum.*

A continuación se definen algunos términos que se utilizan como parámetros de evaluación:

Exactitud: es la proporción de resultados verdaderos tanto verdaderos positivos(VP) como verdaderos negativos (VN) entre el número total de casos examinados (verdaderos positivos, falsos positivos(FP), verdaderos negativos, falsos negativos(FN), ecuación (5.1).

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN} \tag{5.1}$$

Precisión: Es la fracción de observaciones clasificadas correctamente como positivas, sobre todas las predicciones clasificadas como positivas, ecuación (5.2).

$$Precision = \frac{VP}{VP + FP} \tag{5.2}$$

Recuerdo: Es la fracción de observaciones clasificadas correctamente como positivas, sobre todas las observaciones positivas, ecuación (5.3).

$$Recuerdo = \frac{VP}{VP + FN} \tag{5.3}$$

Medida F: Es el significado armónico entre la precisión y el recuerdo, ecuación (5.4).

$$MedidaF = \frac{(1 + \beta^2)(2 * Presicion * Recuerdo)}{(\beta^2 * Presicion) + Recuerdo} \tag{5.4}$$

En la Tabla 5.7 se presentan los resultados para la precisión, recuerdo y medida F donde puede observarse que el promedio de la precisión es de 0.90 y de medida F de 0.89.

Tabla 5.7: Precisión, recuerdo y medida F obtenidos.

Especie	Precisión	Recuerdo	Medida F
Ateuchus rodriguezi	1.00	1.00	1.00
Canthon (Canthon) humectus	1.00	1.00	1.00
Canthon (Canthon) indigaceus	1.00	1.00	1.00
Deltochilum gibbosum sublaeve	0.63	0.83	0.71
Deltochilum tumidum	0.67	0.40	0.50
Dichotomius colonicus	0.86	1.00	0.92
Digitonthophagus gazella	1.00	1.00	1.00
Phanaeus mexicanus	1.00	0.86	0.92
Hybosorus illigeri	1.00	1.00	1.00
Promedio	0.90	0.89	0.89

En la figura 5.5, la línea roja se aproxima al contorno del escarabajo *Canthon (Canthon) humectus* los asteriscos verdes son los 30 puntos que se toman para las medidas estadísticas y se ilustran 2, 4, 6 y 8 armónicos para aproximar el contorno de nuestro escarabajo. Observe que para 30 puntos y conforme aumenta el número de armónicos, la serie elíptica de Fourier (línea roja) se aproxima más al contorno del escarabajo, se resalta en inciso **D)** con 30 puntos, y 8 armónicos que nos dan 29 atributos para los descriptores de Fourier; además los 30 puntos verdes son los que se consideran para calcular las 18 medidas estadísticas. También se probó para 16 armónicos pero el costo computacional crece y la información obtenida no aporta mejoras significativas a los resultados.

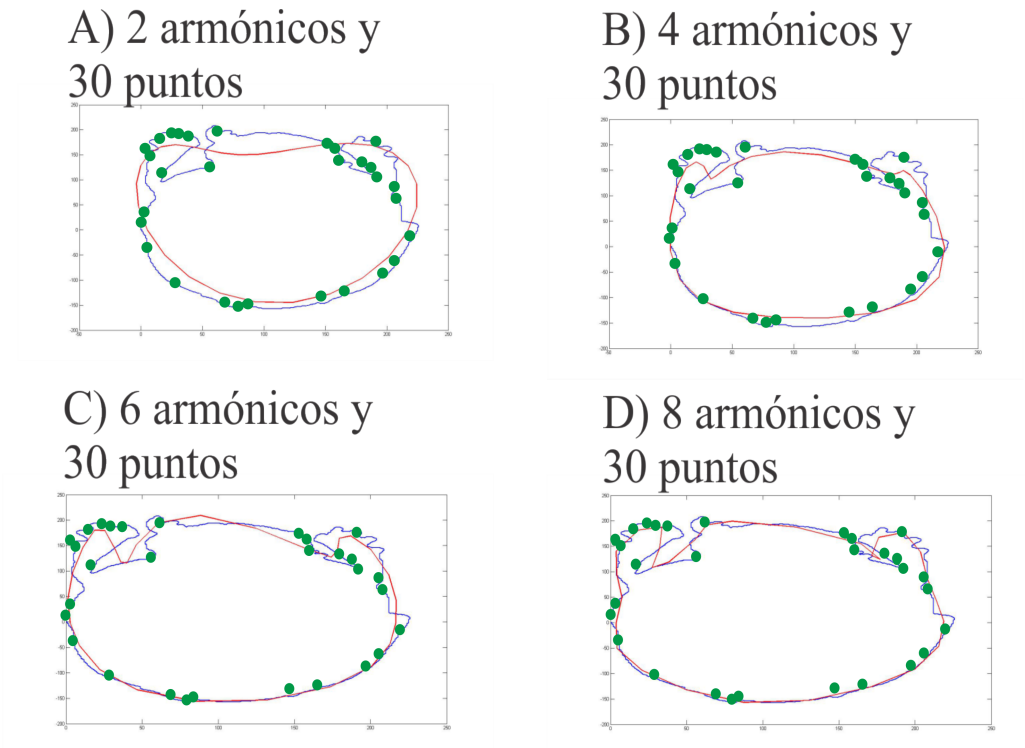

Figura 5.5: Aproxima al contorno del escarabajo *Canthon (Canthon) humectus* los 30 puntos verdes que se toman para las medidas estadísticas y se ilustran 2, 4, 6 y 8 armónicos.

5.4. Validación del método propuesto

Para validar nuestro método se creó un modelo con 5 escarabajos de cada especie teniendo un total de 45 imágenes, con este modelo se obtiene la probabilidad de la pertenencia de cada escarabajo a su especie. Teniendo la probabilidad de pertenencia de cada escarabajo se calcula el promedio y la desviación estándar para conocer el comportamiento de nuestro modelo Tabla 5.8.

Tabla 5.8: Tabla de pertenencia de cada escarabajo a su especie

Escarabajo	Fam1	Fam2	Fam3	Fam4	Fam5	Fam6	Fam7	Fam8	Fam9
Escarabajo1	0.87	0.54	0.50	0.30	0.55	0.67	0.32	0.32	0.58
Escarabajo2	0.76	0.55	0.53	0.47	0.44	0.69	0.39	0.54	0.52
Escarabajo3	0.76	0.71	0.76	0.65	0.32	0.55	0.47	0.50	0.26
Escarabajo4	0.77	0.24	0.54	0.45	0.44	0.36	0.26	0.57	0.58
Escarabajo5	0.76	0.46	0.47	0.54	0.49	0.56	0.39	0.51	0.74
promedio	0.78	0.50	0.56	0.48	0.45	0.57	0.37	0.49	0.54
desviacionestandar	0.05	0.17	0.12	0.13	0.08	0.13	0.08	0.10	0.17

Se calcula la cota del Azar que es 1 entre el número de familias de nuestro modelo y se obtiene que la cota del azar es 0.1111 esto es que si la probabilidad de pertenecía es menor a 0.1111 el clasificador está clasificando al azar. En la figura 5.6 la línea azul muestra que la probabilidad de pertenencia está por arriba de la cota del azar lo que significa que el clasificador está funcionando correctamente y la línea naranja muestran el margen de holgura en la que puede estar la probabilidad de pertenencia.

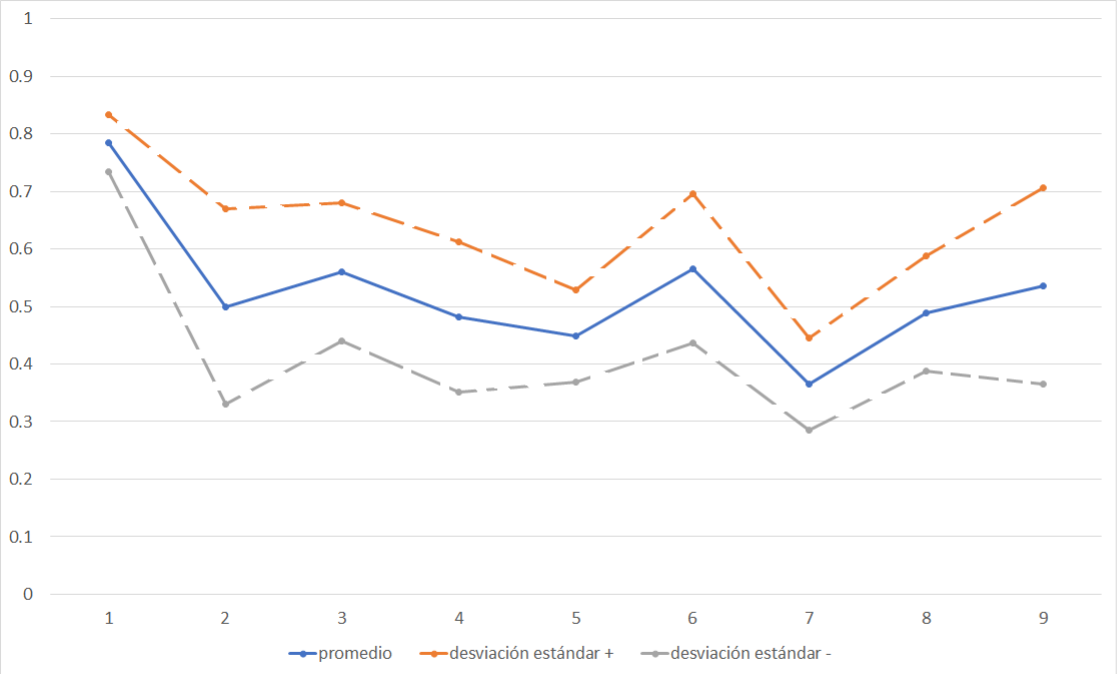

Figura 5.6: Promedio y desviación estandar de pertenencia de cada escarabajo

Para fijar un parámetro de pertenencia se agregan imágenes de formas diferentes al escarabajo y escarabajos que no pertenezcan a las especies del modelo Tabla 5.9.

Tabla 5.9: Tabla de pertenencia de cada figura con el modelo

Escarabajo	Fam1	Fam2	Fam3	Fam4	Fam5	Fam6	Fam7	Fam8	Fam9
fig1	0.04	0.08	0.03	0	0	0	0.22	0.62	0.01
fig2	0.30	0.02	0.02	0	0	0.05	0.20	0.07	0.34
fig3	0.34	0.02	0	0	0.01	0.07	0.13	0.08	0.35
fig4	0.28	0.10	0	0	0.01	0.09	0.15	0.06	0.31
fig5	0.32	0.06	0.01	0	0	0.07	0.13	0.07	0.34
promedio	0.26	0.06	0.01	0	0.00	0.06	0.17	0.18	0.27
desviación estándar	0.12	0.04	0.01	0	0.01	0.03	0.04	0.25	0.15

En la figura 5.7 la línea azul muestra que la probabilidad de pertenencia está cerca de la cota del azar pero muy por debajo de la probabilidad de pertenencia que tiene cada escarabajo con su especie y la naranja muestran el margen de holgura en la que puede estar la probabilidad de pertenencia para cada forma.

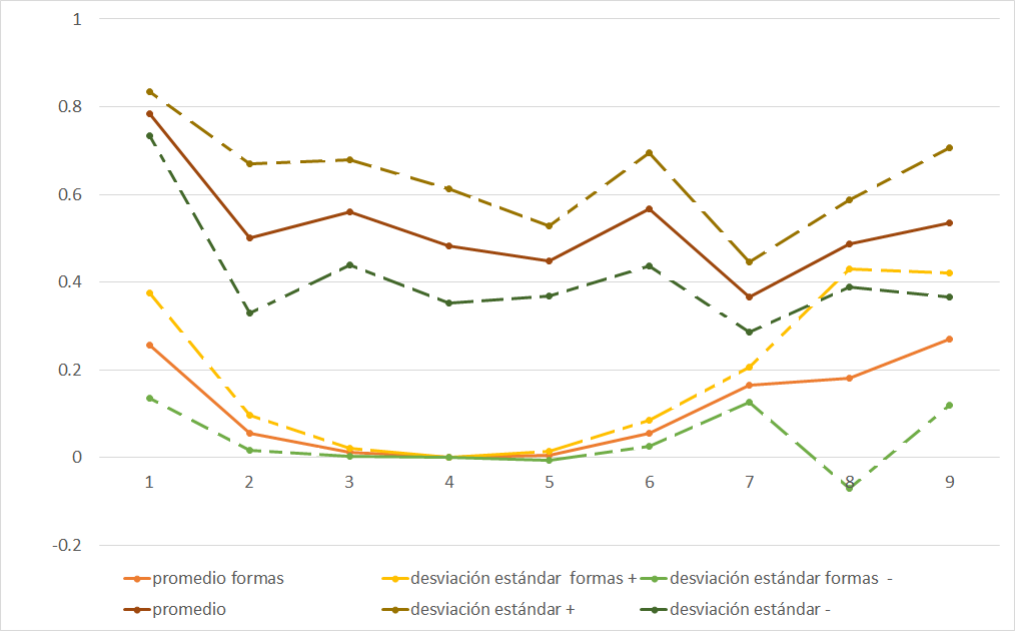

Figura 5.7: Promedio y desviación estandar de pertenencia de cada figura

Capítulo 6

Conclusiones y trabajo futuro.

Las herramientas digitales desarrolladas en este trabajo para el reconocimiento de escarabajos contribuyen y facilitan esta tarea, el sistema "Automatización de claves de las especies de dos familias, Scarabaeidae e Hybosoridae (Scarabaeoidea)", es una primera herramienta que resulta de gran utilidad para los estudiantes y maestros en el área de Biología, su importancia radica en la eliminación de las claves impresas. Este sistema reduce el tiempo empleado en un 54 % para la identificación de algún escarabajo perteneciente a alguna de las dos familias, como trabajo futuro se agregarán más familias al sistema. A pesar de que este sistema es funcional y se ha incorporado como una herramienta indispensable en el área de Biología, solo dió pauta para realizar reconocimiento automático de estas especies utilizando imágenes digitales.

Para la parte de reconocimiento de escarabajos con imágenes digitales, el contorno del escarabajo proporciona información útil para esta tarea. La forma de extraer información del contorno del escarabajo es por medio de la transformada Elíptica de Fourier, esta ofrece resultados por abajo de la cota del azar de un 0.52 de exactitud pero no es suficiente para poder realizar un reconocimiento de forma confiable. Es por esta razón que se utiliza medidas estadísticas como una segunda forma de obtener información del contorno del escarabajo, el resultado de las medidas estadísticas es de 0.85 de exactitud, se tiene una exactitud de 0.85 y se presenta una confusión con 4 especies de escarabajos por lo que este método tampoco es suficiente para poder realizar un reconocimiento de forma confiable. Al analizar la confusión de las especies en el método transformada Elíptica de Fourier, se nota que dos de las especies de escarabajos que presentan confusión con el método de las medidas estadísticas no presentan este problema. Por lo que se propone una combinación de estos dos métodos con la finalidad de aumentar la exactitud y reducir el número de especies que presentan confusión, con esto se puede concluir que la transformada Elíptica de Fourier y las medidas estadísticas consideradas en este trabajo ofrecen información suficiente para alcanzar una exactitud del 0.89 y se reduce la confusión a dos especies.

El mayor número de errores en la clasificación se dan entre ejemplares de la misma tribu, lo que implica que de manera natural tienen ciertos atributos en común, así que estos errores deben disminuir al contar con más ejemplares. En cuatro especies la exactitud y el recuerdo son iguales a 1, con el 100 por ciento de ejemplares clasificados correctamente. Como trabajo futuro se experimentará agregando nuevos ejemplares y otros atributos como el color, y líneas o manchas específicas de los ejemplares.

- Se automatizo un sistema de claves.
- Se estudiaron modelos de clasificación.
- Se creó un modelo de clasificación de escarabajos con diferentes atributos.
- Se propuso una nueva forma de identificación de escarabajos.
- Se determinó parámetros de pertenencia a cada especie.
- Se creó una herramienta computacional que permite identificar escarabajos utilizando imágenes digitales.

Referencias

[1] Dean C Adams, F James Rohlf, and Dennis E Slice. Geometric morphometrics: ten years of progress following the 'revolution'. *Italian Journal of Zoology*, 71(1):5–16, 2004.

[2] Gustavo A Betancourt. Las máquinas de soporte vectorial (svms). *Scientia et technica*, 1(27), 2005.

[3] Richard E Blackwelder. Taxonomy; a text and reference book. 1967.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Rubén Castañeda-Osorio, Hortensia Carrillo-Ruiz, Sombra Patricia Rivas-Arancibia, and Marcela Sánchez-Carrillo. Melolonthidae y cetoniidae (coleoptera: Scarabaeoidea) en el rancho el salado, jolalpan, puebla, méxico. *Dugesiana*, 22(2):227–241, 2015.

[6] Reyna CASTILLO, Juan M HERNÁNDEZ, Everardo INZUNZA, and Juan P TORRES. Procesamiento digital de imágenes empleando filtros espaciales.

[7] Rodríguez del Bosque, Luis Ángel Morón Ríos, Miguel Ángel, Rodríguez del Bosque, and Miguel Ángel Morón. *Plagas del suelo*. Number 632.90972 P5. 2010.

[8] Dennis Denis Ávila. Aplicación de las funciones elípticas de fourier para la descripción de la forma de los huevos de las aves. *Revista de Biología Tropical*, 62(4), 2014.

[9] Dia is free software available under the terms of the GNU GNU General Public License, the GPLv2. http://dia-installer.de/, 23 de Mayo de 2014.

[10] Naoya Furuta, Seishi Ninomiya, Nobuo Takahashi, Hiroshi Ohmori, and Ukai Yasuo. Quantitative evaluation of soybean (glycine max l. merr.) leaflet shape by principal component scores based on elliptic fourier descriptor. *Japanese Journal of Breeding*, 45(3):315–320, 1995.

[11] Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006.

[12] Rafael C Gonzalez and Richard E Woods. Processing, 2002.

[13] Eva Heller. *Psicología del color Cómo actúan los colores sobre los sentimientos y la razón*. Gustavo Gili, 1 edition, 2008.

[14] Arturo Hinojo-Ruelas and Hermilo Sánchez-Cruz. Análisis de técnicas para extracción de características en imágenes binarias. 12.

[15] Mary Liz Jameson. Phylogenetic analysis of the subtribe rutelina and revision of the rutela generic groups (coleoptera: Scarabaeidae: Rutelinae: Rutelini). *Bulletin of the University of Nebraska State Museum*, 14(1997):1–184, 1998.

[16] Frank P Kuhl and Charles R Giardina. Elliptic fourier features of a closed contour. *Computer graphics and image processing*, 18(3):236–258, 1982.

[17] LAWRENCE, J. F. and A. F. NEWTON. `http://www.museum.unl.edu/research/entomology/Guide/Guide-introduction/index.html`, 8 de Julio de 2002.

[18] Seymour Schiller John J Cortiñas Lipschutz et al. *Introducción a la probabilidad y estadística*. Number 519.5 L56. 2000.

[19] Jorge Llorente Bousquets and Jorge Llorente Bousquets. *La búsqueda del método natural*. 2002.

[20] Constantino Malagón Luque. Clasificadores bayesianos el algoritmo naïve bayes. *Mayo*, 2003.

[21] Marcos Martın. Técnicas clásicas de segmentación de imagen. 2002.

[22] HK Mebatsion, J Paliwal, and DS Jayas. Evaluation of variations in the shape of grain types using principal components analysis of the elliptic fourier descriptors. *Computers and electronics in agriculture*, 80:63–70, 2012.

[23] MA Morón, Agustín Aragón-García, and Hortensia Carrillo-Ruiz. Fauna de escarabajos del estado de puebla. *Coordinación Editorial Miguel Ángel Morón Ríos, Coatepec, Veracruz, México.*, 2013.

[24] MA Morón Ríos and Roberto Terrón. *Entomología práctica: una guía para el estudio de los insectos con importancia agropecuaria, médica, forestal y ecológica de México*. Number QL 477. M67 1988. 1988.

[25] JJ Morrone. Sistemática. fundamentos, métodos, aplicaciones. *Revista de la Sociedad Entomol*, 2, 2013.

[26] Gianfranco Passariello. *Imagénes médicas. Adquisición, Análisis.* Equinoccio, 1999.

[27] Roger S. Pressman. *Ingeniería del software un enfoque practico.* Mc Graw Hill, 7 edition, 2010.

[28] F James Rohlf and James W Archie. A comparison of fourier methods for the description of wing shape in mosquitoes (diptera: Culicidae). *Systematic Zoology*, 33(3):302–317, 1984.

[29] Brenda Sánchez-Velázquez, Hortensia Carrillo-Ruiz, Miguel Ángel Morón, and Sombra Patricia Rivas-Arancibia. Especies de scarabaeidae e hybosoridae (coleoptera: Scarabaeoidea) que habitan en la comunidad del rancho el salado, jolalpan, puebla, méxico. *Dugesiana*, 18(2):207–215.

[30] See also the NetBeans IDE 8.2 New and Noteworthy page. https://netbeans.org/community/releases/82/, 17 de Marzo de 2017.

[31] Krishna Singh, Indra Gupta, and Sangeeta Gupta. Classification of bamboo species by fourier and legendre moment. *International Journal of Advanced Science and Technology*, 50:61–70, 2013.

[32] Ian Sommerville. *Ingeniería del software.* Pearson, 9 edition, 2004.

[33] Alma L Trujillo-Miranda, Hortensia Carrillo-Ruiz, Sombra P Rivas-Arancibia, and A Rosa Andrés-Hernández. Estructura y composición de la comunidad de escarabajos (coleoptera: Scarabaeoidea) en el cerro chacateca, zapotitlán, puebla, méxico. *Revista Mexicana de Biodiversidad*, 87(1):109–122, 2016.

[34] José Ramón Mejía Vilet. Procesamiento digital de imágenes. *Facultad de Ingeniería UASLP, Documento PDF, disponible en la página¡ http://read. pudn. com/downloads159/ebook/711796/Procesamien to_Digital_de_Imagenes. pdf¿[Citado 22 de septiembre de 2012]*, 2005.

[35] Ronald E Walpole, Raymond H Myers, and Sharon L Myers. *Probabilidad y estadística para ingenieros.* Pearson Educación, 1999.

[36] He-Ping Yang, Chun-Sen Ma, Hui Wen, Qing-Bin Zhan, and Xin-Li Wang. A tool for developing an automatic insect identification system based on wing outlines. *Scientific reports*, 5, 2015.

[37] Qing-Bin Zhan and Xin-Li Wang. Elliptic fourier analysis of the wing outline shape of five species of antlion (neuroptera: Myrmeleontidae: Myrmeleontini). *Zoological Studies*, 51(3):399–405, 2012.