INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
MAESTRÍA EN SISTEMAS COMPUTACIONALES

MODELO DE MINERÍA DE DATOS SOBRE LA COLECCIÓN DE ICTIOPLANCTON DE CICIMAR-IPN

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
ISC. ERICK ENRICO JIMÉNEZ VALTIERRA

DIRECTOR DE TESIS:
MC. JESÚS ANTONIO CASTRO, UNAM

MIEMBROS DEL JURADO:
MC. JESÚS ANTONIO CASTRO, UNAM
DRA. SYLVIA PATRICIA A. JIMÉNEZ ROSENBERG, CICIMAR-IPN
MATI. LUIS ARMANDO CÁRDENAS FLORIDO, ITESM
MSC. ILIANA CASTRO LIERA, ITLP

MAESTRÍA EN SISTEMAS COMPUTACIONALES, ITLP

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, AGOSTO 2014.

CONTENIDO

Resumen

La península de Baja California contiene uno de los sistemas costeros más productivos del Pacífico Mexicano, por lo que el estudio de su ecosistema marino se vuelve una actividad fundamental en cuanto a la investigación de su dinámica ambiental y comunidades biológicas y en cuanto al desarrollo de actividades económicas y de turismo.

En el Centro Interdisciplinario de Ciencias Marinas del Instituto Politécnico Nacional (CICIMAR), en la ciudad de La Paz, existen cuatro departamentos en los que se desarrolla investigación acerca de diferentes aspectos del ecosistema marino.

Uno de estos departamentos, el de Plancton y Ecología Marina, alberga la Colección Científica de Huevos y Larvas de Peces del Pacífico Mexicano (ICTIOPLANCTON). De esta colección se derivan varias bases de datos geo-referenciadas, incluyendo aquellas que provienen de los cruceros cuatrimestrales que se realizan en la costa oeste de la Península de Baja California, a cargo del Programa de Investigaciones Mexicanas de la Corriente de California (IMECOCAL; http://imecocal.cicese.mx/).

La información que se obtiene de esta colección es vital para poder observar de manera objetiva la distribución y abundancia de las especies de peces representadas por sus larvas en sus distintos estadios de desarrollo. Esta información es obtenida y registrada para su posterior análisis, de manera que pueda ser usada para procesos de toma de decisiones en distintos ámbitos productivos y de investigación.

La construcción y análisis de estas bases de datos son desarrollados con el apoyo de la tecnología computacional que se tiene dentro de la institución, la extracción y almacenamiento se realiza sobre información que está en forma de hojas de cálculo de Microsoft Excel.

Es importante notar que no es el sistema más adecuado para la cantidad de información que se obtiene y se utiliza y los procesos suelen ralentizarse debido al tiempo que lleva extraer la información desde ese tipo de formato.

La clasificación y predicción de datos sobre información de este tipo podría llevar al análisis más fino de patrones de distribución de las especies y de su potencial uso como especies indicadoras de la dinámica ambiental.

Dada la naturaleza de la información esto podría agilizar de gran manera la integración y análisis de datos, y a la par los procesos de investigación que se realizan en este ecosistema.

El presente proyecto propone ofrecer una perspectiva diferente de análisis automático sobre la información que ya se tiene en las bases de datos en la institución, tomando el enfoque de los procesos de descubrimiento de información en bases de datos, utilizando técnicas de modelos multidimensionales y de minería de datos.

Abstract

The Peninsula of Baja California is one of the most productive in the Mexican Pacific coastal systems, so the study of its marine ecosystem becomes a core activity in terms of its environmental research dynamics and biological communities in the development of economic and tourism activities.

In the Interdisciplinary Center of Marine Sciences of the National Polytechnic Institute (CICIMAR), in the city of La Paz, there are four departments in which research is conducted on different aspects of the marine ecosystem.

One of these departments, the Plankton and Marine Ecology, houses the Scientific Collection of Fish Eggs and Larvae of the Mexican Pacific (ICTIOPLANCTON). From this collection several geo-referenced data are derived, including those from the quarterly cruises that take place on the west coast of the Peninsula of Baja California, by the Mexican Research Program of the California Current (IMECOCAL; http://imecocal.cicese.mx/).

The information obtained in this collection is vital to observe objectively the distribution and abundance of fish species represented by their larvae in their different stages of development. This information is collected and recorded for later analysis, so it can be used for decision-making processes in various production and research areas.

The construction and analysis of these databases are developed with the support of computer technology that is within the institution, extraction and storage is performed on information that is in the form of spreadsheets in Microsoft Excel.

It is important to note that it is not the most appropriate way for the amount of information obtained and used, processes are often slow due to the time it takes to extract the information from that format system.

Classification and prediction of data on such information may lead to finer patterns of species distribution and their potential use as indicators of environmental dynamic analysis.

Given the nature of this information, this could greatly speed up the integration and analysis of data, and at the same time research processes that take place in this ecosystem.

This project aims to provide a different perspective on the automatic analysis of information that is already taken into databases in the institution, taking the approach of the processes of information discovery in databases, using techniques of multidimensional models and data mining.

Capítulo 1 Introducción.

1.1 Introducción

El desarrollo de bases de datos multidimensionales nos permite tener una nueva perspectiva de la información que tenemos en nuestras bases de datos, transformando nuestros sistemas gestores de información de ser solo catálogos a sistemas más completos de análisis profundo y eficiente, agilizando un gran número de procesos y ayudándonos a descubrir patrones y tendencias nunca vistas dentro de nuestras propias bases de datos.

Esto permite que los procesos de toma de decisiones que se desarrollan dentro de una organización tengan una mayor cantidad de información soporte, habiéndole dado a la información un sentido, convirtiéndola de solo datos a conocimiento.

La minería de datos se encarga del análisis de esta información, iniciando con un proceso de extracción y transformación de la información, creación de la estructura multidimensional apropiada para el modelo, y la aplicación de algoritmos de minería de datos.

Un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

Estas técnicas se incorporan para realizar un proceso conjunto al que se le llama proceso de descubrimiento de información en bases de datos (KDD) por sus siglas en inglés. Este tipo de estudios permiten realizar proyecciones sobre el estado actual de la información que se tiene, y desarrollar patrones que nos permitan predecir como cambiará la información.

El conocimiento generado se puede utilizar para generar nuevas estrategias para poder estudiar la información desde una perspectiva más completa y objetiva.

Todo esto, aplicado a las bases de datos de larvas de peces que se tienen en CICIMAR, podría llevar a:

- Realizar un mapa de distribución de especies más completo que incluya posibles escenarios futuros.
- Tomar decisiones pesqueras relevantes, analizando el estado actual de las especies y su estado de desarrollo, de manera que modifiquen los lugares designados para pesca en un momento determinado.
- Desarrollar nuevos modelos de investigación que permitan examinar un mayor número de especies de manera más eficiente, sabiendo con anterioridad donde hacer muestreos.
- Modelar técnicas de protección para especies que se vean amenazadas en alguno de sus estados de desarrollo.
- Facilitar el análisis de eventos oceanográficos tales como el Niño y la Niña, haciéndolo más eficiente en términos de tiempo y calidad.

Todos estos escenarios son posibles gracias a la aplicación de los modelos de bodega de datos sobre la información que ya se tiene almacenada en las bases de datos de la institución, las

cuales contemplan información biológica sobre la especie de larva del pez, así como información acerca de la estación donde se recolectó.

1.2 Investigación previa relevante

Se han realizado trabajos previos que reúnen y organizan información biológica relevante acerca de los ecosistemas tanto marinos como terrestres. Algunos ejemplos son los siguientes.

Fishbase

FishBase (Algunas pantallas importantes se muestran en las Figuras 1.1 y 1.2) es una base de datos con información en línea sobre especies de peces a nivel mundial. En octubre del 2006 incluía la descripción de más de 29,400 especies, 222,300 nombres vulgares en cientos de idiomas, 42,600 fotografías, y referencias a 38,600 trabajos en la literatura científica.

En 1987, Daniel Pauly, inspirado por las Hojas de Identificación de Especies (Species Identification Sheets) y otros productos de Walter Fischer producidos para la FAO en la década de 1970, propuso crear una base de datos estandarizada para especies de peces, como parte del "ICLARM Software Project". El año siguiente, Daniel comenzó a trabajar junto con Rainer Froese, quien había estado trabajando en un sistema experto para identificar larvas de peces. Luego de un intento frustrado de construir un sistema utilizando Prolog, Froese cambió a DataEase, una base de datos relacional para DOS. Para 1989 el proyecto recibió su primer apoyo económico institucional.

En 1993 el proyecto cambió a Microsoft Access y en 1995 se produjo el primer CD-ROM llamado "FishBase 100". Los comentarios en las revistas científicas si bien alabaron el alcance y la idea, destacaron que había numerosos huecos en la información provista. Las ediciones de CD posteriores se han venido editando con una frecuencia anual. La versión de FishBase 2004 por su volumen requería de cinco CD, o un DVD. Para poder visualizarse el sistema requiere Windows 98 o posterior, y no está disponibles en otras plataformas como Mac OS X o Linux.

FishBase estuvo disponible por primera vez en la Web en agosto de 1996, al año siguiente se contrató un webmaster. En ese momento, toda la información de la base de datos estuvo disponible para consulta a través de la red.

A partir del año 2000, FishBase ha sido administrada por el FishBase Consortium. El consorcio está formado por (Instituto, Ciudad de Origen):

- Africamuseum, Tervuren
- Aristotle University of Thessaloniki, Thessaloniki
- Fisheries Centre University of British Columbia, Vancouver
- Food and Agriculture Organization of the United Nations, Rome
- IFM-GEOMAR, Kiel
- Muséum National d'Histoire Naturelle, Paris
- Swedish Museum of Natural History, Stockholm
- WorldFish Center, Penang

En la medida que los especialistas en peces se han enterado de la existencia de FishBase, más de 1370 colaboradores han enviado contribuciones. Para mantener su valor como una base de datos de carácter científica, no se permite el agregado de información original en FishBase; todo

su contenido debe estar basado en material que ha sido publicado con anterioridad.

Figura 1.1, Fishbase.org

Note: Lists may be incomplete. Some lists may be very long and will take time to load
Note: A new dropdown list will appear if a country has a sub-country (ex. Canada, USA, etc.)

Information by Ecosystem

- All fishes
- Point data
- Ecosystem info
- Resilience of fishes
- Trophic pyramids
- Species Ecology Matrix
- Deep-water
- Ecopath parameters
- Identification by pictures
- Identification keys

Note: Lists may be incomplete. Some lists may be very long and will take time to load

Information by Topic

Trophic ecology
- Diet
- Food items
- Food consumption
- Ration
- Predators

Physiology/Behavior
- Metabolism
- Gill area
- Brains
- Vision
- Fish sounds
- Swim. speed

Life history
- Growth
- L-W relationship
- Length frequencies
- Recruitment
- Reproduction
- Maturity
- Spawning
- Fecundity
- Eggs
- Egg dev.
- Larvae
- Larval dynamics
- Abundance

Uses
- Aquaculture
- Aquaculture profiles
- Introductions
- Diseases
- Ciguatera
- Processing
- Ecotoxicology
- Genetics
- Allele frequencies
- Heritability
- Otoliths
- Mass conversion

Miscellaneous
- Treaties & Conv.
- CITES
- CMS
- National databases
- Names by Language
- Collaborators
- Public aquariums
- Expeditions
- Video
- Fish stamps and coins
- Uploaded photos online

Note: Lists may be incomplete. Some lists may be very long and will take time to load

Tools

- Quick Identification
- Identification keys
- Identification by morphometrics
- Adverse introductions
- Global introductions
- Invasiveness
- Species by ecosystem
- Graphs
- SeaFood Advisory
- Shifting Baselines WP2 - Online Toolset
- Preferred algae/plants of herbivorous fishes
- Match names
- Disease diagnosis
- My Fish Page
- Life-history tool
- L-F Analysis
- Information gaps
- Random Species
- Sea Around Us
- FishBase for Americas
- FishBase for Africa
- FishBase for the Red Sea
- ISSCAAP Troph
- FAO aquaculture
- FAO catches
- Catch analysis
- ICES catch
- Catch-MSY
- Classification List
- Classification Tree
- Fish statistics
- World records
- Country codes
- Catalogue of Life
- Fish collections
- Collection History
- Trophic pyramids
- Ecopath parameters
- AquaMaps
- New species in FishBase
- New species in Welt der Fische
- New photos
- Web Stats
- Top 100
- Coastal Transects Analysis Model (CTAM)

Note: Tools without radio button are available from the Species Summary page.

Figura 1.2, Fishbase.org

Biótica

El Sistema de Información Biótica (Portada mostrada en Figura 1.3) ha sido diseñado especialmente para el manejo de datos curatoriales, nomenclaturales, geográficos, bibliográficos y de parámetros ecológicos. Tiene el propósito de ayudar en la captura y actualización de la información.

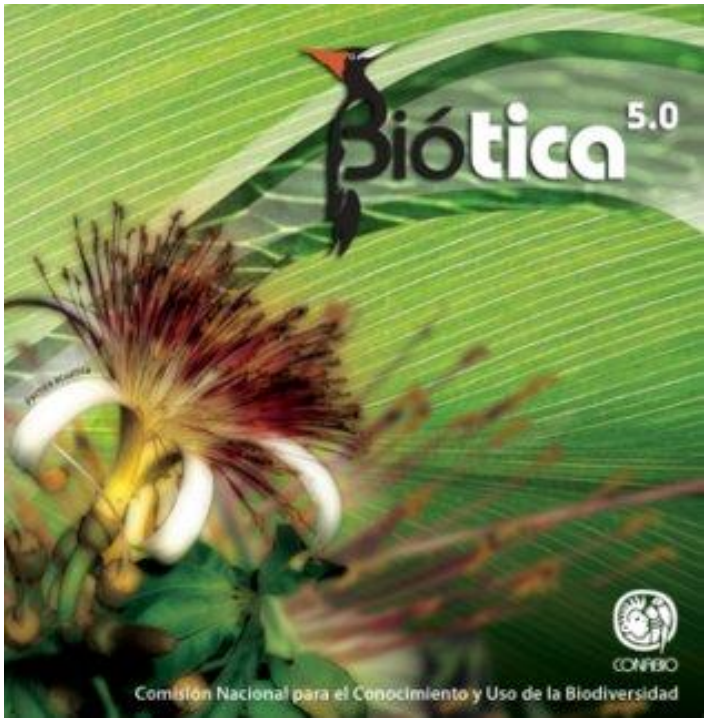


Figura 1.3, Biótica Portada

Biótica fue desarrollado en forma modular tanto en la estructura de la base de datos como en su sistema tomando en cuenta la gran variedad de necesidades de la comunidad biológica (taxónomos, curadores, biogeógrafos, ecólogos, etnobiólogos, etc.); además de otras características como son códigos de barras para la etiquetación de ejemplares; es posible ligar la información de la base de datos con información manejada por otras aplicaciones como imágenes, sonidos, páginas WWW, hojas de cálculo, otras bases de datos, etc.

Es posible utilizar Biótica tanto en un ambiente monousuario como en un ambiente multiusuario (red) con bases de datos en Microsoft Access y Microsoft SQL Server.

Biótica 5.0 se divide en diez módulos: Base de datos, Directorio, Nomenclatural, Ejemplar, Ecología, Geográfico, Bibliografía, Colecciones, Herramientas y Ayuda.

Biótica incluye catálogos de:

- Nombres científicos de algunos grupos biológicos.
- Autoridades.
- Instituciones y colecciones.
- Estados y municipios para México (INEGI 2005).

- Regiones hidrológicas, marinas y terrestres prioritarias de México, así como las ecorregiones marinas de Norteamérica.
- Características que pueden asociarse al taxón.
- Características que pueden asociarse al ejemplar.
- Características que pueden asociarse al sitio.
- Parámetros poblacionales.
- Tipos nomenclaturales.
- Relaciones entre taxones.

Como se puede ver, los anteriores sistemas de organización han sido solo para la integración de la información con fines de catalogación y clasificación. Sin embargo no hay herramientas que tengan también la capacidad de análisis que brinda la implementación de los procesos KDD, por lo cual este proyecto se plantea como una innovación en el campo de los sistemas computacionales para la investigación biológica.

El caso de estudio de este proyecto es la base de datos ICTIOPLANCTON.

Las colecciones biológicas son fundamentales para conocer y promover el conocimiento sobre la biodiversidad. Su objetivo es contener una representatividad de las especies que se pueden presentar en un espacio y tiempo y bajo ciertas condiciones ambientales. Con ello nos será más fácil reconocer a aquellas comunidades que nos indiquen cambios en el ambiente, permitiéndonos interpretar esos cambios en posibles afectaciones en el ámbito ecológico y/o económico.

En el estudio de las comunidades marinas se ha observado que la distribución de las especies sigue gradientes reconocibles asociados a variables ambientales de diferente naturaleza, con lo que diversos organismos han sido utilizados como indicadores de características particulares de este ambiente (1).

El hecho de que los huevos y larvas de peces tengan requerimientos ambientales diferentes a los de los adultos, propicia que estas comunidades se destaquen como indicadoras. Las asociaciones de larvas de peces se caracterizan por ser muy diversas y por estar completamente condicionadas por los factores ambientales. Como resultado, presentan una distribución mucho más limitada en espacio y tiempo (2).

Generalmente, el análisis de asociaciones de larvas de peces se hace sin distinción de los diferentes estadios que las larvas atraviesan durante su desarrollo. Al integrar a todos estos estadios en una sola entidad, se acepta que todas las variables afectan de la misma forma y con la misma intensidad a cada uno de ellos. Los supuestos bajo este concepto, minimizan el hecho de que el comportamiento, las necesidades y las capacidades de las larvas cambian durante su desarrollo. Conforme avanza en el desarrollo, la larva va adquiriendo capacidades que le permiten ser más selectiva en cuanto a sus presas y más eficaz en cuanto a la evasión de depredadores, siendo afectada su distribución por distintas variables ambientales. Por lo tanto, no solo es importante conocer a que factores está asociada la distribución de las larvas de peces, sino que factores influyen más en cada uno de sus estadios larvarios. Esto nos permitirá conocer de una manera más fina las condiciones que prevalecen en el ambiente en determinado espacio y tiempo, usando esta información para la planeación estratégica de protección de especies, comunidades biológicas y/o Centros de Actividad Biológica, para el manejo sustentable de

pesquerías y para modelación de posibles impactos causados por cambios climáticos a corto y largo plazo.

1.3 Objetivo de la investigación

El objetivo principal de la investigación es agilizar los procesos de análisis de información que aportan las bases de datos de larvas de peces de la colección ICTIOPLANCTON en términos de tiempo y calidad. Para tal efecto se propone diseñar e implementar un modelo de bases de datos y su posterior proceso de minería, mediante un algoritmo de árboles de decisión sobre la información histórica contenida en dichas bases de datos.

1.3.1 Objetivos específicos:

- Seleccionar datos iniciales a partir la información histórica aportada por las bases de datos de larvas de peces de la colección ICTIOPLANCTON.
- Realizar un proceso ETL (Extracción, Transformación y Carga) que convierta esta base de datos de formato Excel a bases de datos relacionales SQL y lo alimente al modelo de datos.
- Identificar los patrones de distribución de las larvas de peces por estadio de desarrollo en diferentes escenarios ambientales mediante un modelo de minería de datos.

1.4 Importancia de la investigación

Uno de los conceptos más importantes dentro del ámbito de la investigación biológica reside en el estudio de los llamados indicadores ambientales y de su empleo para la medición de cambios naturales en la biodiversidad, con enfoque en el manejo de los recursos marinos (3).

Los indicadores nos permiten sacar conclusiones acerca del estado del ambiente y de su dinámica ante la posible intrusión de fenómenos o elementos extraños al mismo, al examinar detenidamente cambios en su estado y en la forma en cómo interactúan entre ellos, como la aparición de nuevas especies en lugares en los que normalmente no se les encuentra.

Los indicadores biológicos nos permiten analizar de manera más objetiva y eficiente los factores que influyen en el cambio de la dinámica de un ecosistema en particular, están diseñados para proveer señales sobre eventos de amplio significado y hacer perceptibles tendencias o fenómenos no detectables inmediatamente (3).

Dentro de la biología marina las larvas de peces son considerados indicadores ambientales de gran utilidad, ya que las mismas necesitan de condiciones muy específicas para eclosionar y sobrevivir a través de sus estadios de desarrollo tempranos, los cuales requieren de condiciones ambientales muy específicas (4).

Mediante una comprensión más profunda y eficiente acerca de las larvas de peces, de la distribución de las especies, de su abundancia e incluso estado fisiológico, es posible detectar alteraciones en las condiciones normales del ambiente en que se desarrollan, entender las causas de estas alteraciones y realizar modelaciones y predicciones para extrapolarlas al ecosistema que habitan.

Esto apoya en gran manera a toda una serie de procesos de toma de decisiones, involucrando procesos productivos y del ámbito de investigación y desarrollo tanto regional como global.

Además del conocimiento de la dinámica ambiental, a partir del análisis de la información que aportan estos indicadores biológicos, se establecen normativas para el establecimiento de pesquerías y la protección de recursos marinos.

Por lo tanto se vuelve importante el desarrollo de sistemas que permitan agilizar este proceso de análisis, de manera que ya no sea tratada solo con fines de catalogación y clasificación, sino que tengan también la posibilidad de generar resultados y conclusiones analizando de manera más eficiente los datos que ya se han reunido.

Esto tendrá un impacto importante en los procesos de toma de decisiones en los ámbitos anteriormente mencionados, lo cual nos dará una ventaja inicial y permitirá modificar regulaciones y procesos, de acuerdo a cambios en el ecosistema, en tiempo prácticamente real.

1.5 Hipótesis:

El proyecto agilizará los procesos de análisis e investigación sobre información de larvas de peces en el ecosistema de la península de Baja California con el fin de apoyar procesos de toma de decisiones de distintos ámbitos productivos y de investigación. Lo anterior mediante la implementación de un modelo de bases datos y su posterior proceso de minería, identificando los patrones de distribución de las larvas de peces por estadio de desarrollo en diferentes escenarios ambientales en base a los datos y pudiendo detectar anomalías, anticipando posibles cambios en la dinámica ambiental que puedan afectar al ecosistema.

1.6 Limitaciones:

- Se trabajó únicamente con información biológica de larvas de peces obtenida mediante muestreos cuatrimestrales realizados en la costa oeste de la península de Baja California.
- Solamente se utilizó el algoritmo de Arboles de Decisión de Microsoft dentro del proceso de minería de datos.
- Solamente se consideraron algunas características ambientales acerca de las estaciones georeferenciadas y sus características hidrográficas.
- Solamente se trabajó con bases de datos proporcionadas por el Departamento de Plancton y Ecología Marina del CICIMAR.

1.7 Contribución al conocimiento

El desarrollo de este proyecto pretende extender el uso de herramientas informáticas como apoyo a los procesos de investigación biológica de trabajar con sus bases de datos con fines de catalogación y clasificación a una nueva etapa en la que la misma herramienta proporcione medios para agilizar los procesos de análisis que se realizan en los centros de investigación de la región, dando la oportunidad de actuar en tiempo real a todos los posibles escenarios que se deriven de los diferentes cambios que pudiese sufrir el ecosistema marino.

Aunque el sistema será elaborado específicamente para trabajar con datos de biología marina, se desarrollará de manera que se pueda migrar a otros tipos de datos con el fin de apoyar el análisis de la información de investigación ambiental que se realizan en la región.

Como contribución computacional de este proyecto se desea dar inicio a una nueva etapa en el desarrollo de sistemas gestores de bases de datos que ya tengan inmersa la capacidad de análisis, ya que actualmente el ámbito nacional de desarrollo de aplicaciones para la investigación ambiental no se ha tomado en cuenta la agregación de este tipo de capacidades.

1.8 Metodología de investigación

El proceso de descubrimiento de conocimiento en bases de datos (KDD), es el modelado y análisis automático y exploratorio de grandes repositorios de datos. KDD es el proceso organizado para identificar patrones válidos, novedosos, útiles y comprensibles a partir de conjuntos de datos grandes y complejos (5). La minería de datos (DM) es el núcleo del proceso KDD, y abarca la inferencia de algoritmos que exploran los datos, desarrollan el modelo y descubren patrones previamente desconocidos. El modelo se usa para entender fenómenos dentro de los datos, para el análisis y para la predicción.

La actual accesibilidad y abundancia de los datos hace que el proceso de minería de datos y de descubrimiento del conocimiento sea de gran importancia.

1.8.1 Proceso KDD

El proceso KDD es iterativo en cada paso, lo cual quiere decir que puede ser necesario moverse hacia pasos anteriores a fin de realizar el proceso correctamente, como se muestra en la Figura 1.4. En cada paso no hay una fórmula definida para realizarlo, sino que se deben aplicar diferentes técnicas dependiendo del problema que se tiene en mente. También es necesario comprender el proceso y las diferentes necesidades y posibilidades en cada paso.

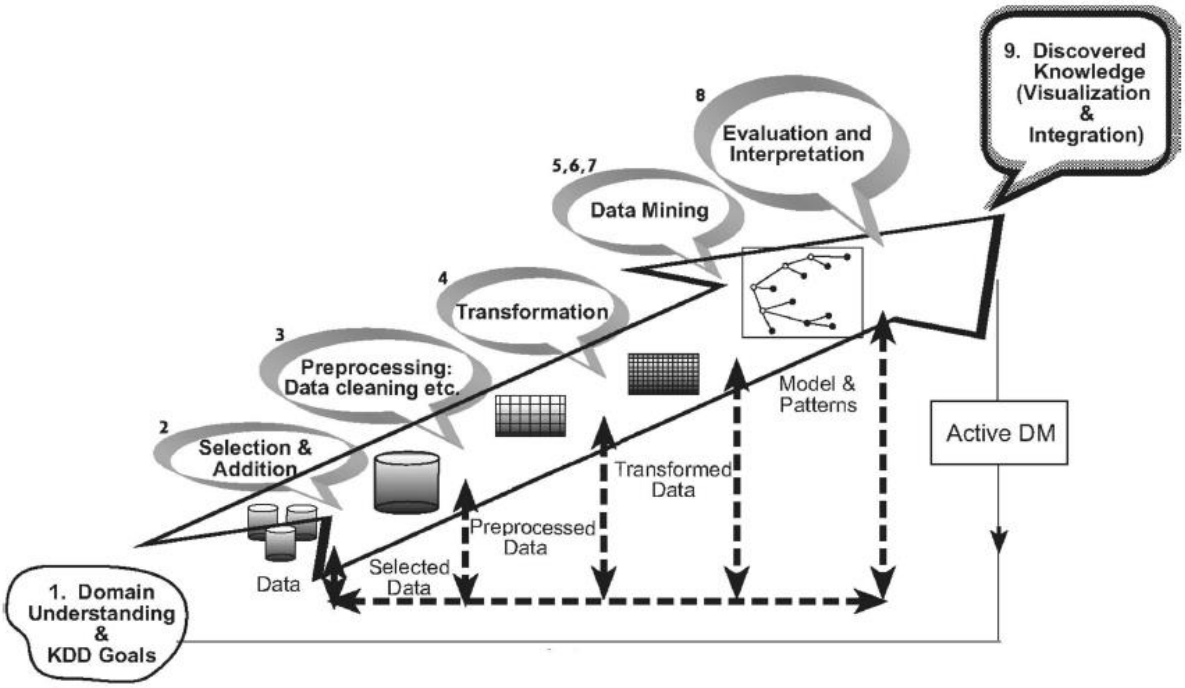


Figura 1.4, Proceso KDD

El proceso comienza determinando las metas KDD y termina con la implementación del conocimiento descubierto. Entonces se cierra el ciclo y comienza la parte activa de la minería de datos.

Esta es una breve descripción de los 9 pasos KDD (6)

1. **Desarrollar una buena comprensión del dominio de la aplicación a usarse**. Este es el paso preparatorio inicial. Prepara la escena para comprender lo que debería hacerse con las decisiones (acerca de la transformación, algoritmos, representación, etc.). La gente que se encuentra a cargo de un proceso KDD debe entender y definir las metas del usuario final y del ambiente en el que tomará lugar el proceso de descubrimiento de conocimiento. Mientras KDD proceda, pueden haber revisiones sobre este paso. Una vez comprendidas las metas KDD, comienza el pre procesamiento de los datos, lo cual se define en los siguientes 3 pasos.
2. **Seleccionar y crear un conjunto de datos en el que tomará lugar el descubrimiento**. Habiendo definido las metas, deben determinarse los datos que se usarán para el proceso. Esto incluye encontrar cuales datos están disponibles, obtener datos adicionales necesarios y la integración de todos ellos en un solo conjunto, incluyendo los atributos que serán considerados para el proceso. Este paso es muy importante, porque la minería de datos aprende y descubre a partir de los datos disponibles. Esta es la evidencia base para construir los modelos. Si hacen falta atributos importantes, entonces el proceso entero podría fallar. Desde este punto de vista, mientras más atributos sean considerados, mejor. Por otro lado, para recolectar, analizar y procesar conjuntos de datos demasiado largos o complejos se tiene un aumento en el costo y tiempo. Esto se debe tener en cuenta al elegir los datos y sus atributos.
3. **Pre procesamiento y limpieza**. En este paso se procede a limpiar los datos del conjunto seleccionado, haciéndolos más fiables. Esto incluye manejar datos faltantes y la remoción de ruido. Hay muchos métodos para realizar este paso, desde no hacer nada hasta transformar el conjunto de datos completo. Esto puede comprender métodos estadísticos complejos o el uso de minería de datos.
4. **Transformación de los datos**. En este paso se prepara y desarrolla la generación de datos mejorados para la minería de datos. Aquí se incluyen la reducción de dimensiones y la transformación de atributos. Este paso puede ser crucial para el éxito del proceso KDD y es usualmente muy específico al tipo de proyecto que se realiza. Una vez completados estos 4 pasos, los siguientes 4 se refieren al proceso de minería de datos, en donde el enfoque se da en los aspectos algorítmicos empleados para cada proyecto.
5. **Elegir la tarea de minería de datos apropiada**. En este punto se debe decidir qué tipo de minería de datos se usará. Esto depende más que nada de las metas KDD y también de los pasos previos. Hay 2 metas principales en la minería de datos: la predicción y la descripción. La predicción se conoce usualmente como minería de datos supervisada, mientras que la descripción incluye los aspectos no supervisados y de visualización de la minería de datos. La mayor parte de las técnicas de minería de datos están basadas en aprendizaje inductivo, en el que se construye un modelo explícitamente o implícitamente al generalizar a partir de un número suficiente de ejemplos de entrenamiento. La idea principal del aproximamiento inductivo es que el modelo entrenado se puede aplicar para casos futuros. La estrategia también toma en cuenta el nivel de meta-aprendizaje para el conjunto particular de datos disponibles. Esto es, qué tanto sabe el algoritmo acerca de los datos que se tienen disponibles.
6. **Elegir el algoritmo de minería de datos**. Una vez que se tiene la estrategia, ahora debemos definir nuestras tácticas. Este paso incluye la selección del método específico que se usará para la búsqueda de patrones. Por ejemplo, si consideramos la precisión contra la comprensión, la

primera es mejor con redes neuronales, mientras que la segunda se da mejor con árboles de decisión. Para cada estrategia de meta aprendizaje hay varias posibilidades para que se logre el proceso. El meta-aprendizaje se centra en explicar qué causa que un algoritmo de minería de datos sea exitoso o no en un problema particular. Además, este acercamiento intenta comprender las condiciones bajo las que un algoritmo de minería de datos es más apropiado. Cada algoritmo tiene parámetros y tácticas de aprendizaje.

7. **Emplear el algoritmo de minería de datos**. Finalmente, la implementación del algoritmo. En este paso puede que sea necesario emplear el algoritmo varias veces hasta que se llegue a un resultado satisfactorio. Por ejemplo al reconfigurar los parámetros del algoritmo, como el número mínimo de instancias en una sola hoja de un árbol de decisión.
8. **Evaluar**. En este paso se evalúan e interpretan los patrones minados (reglas, fiabilidad, etc.), con respecto a las metas definidas en el primer paso. Aquí consideramos los pasos de pre-procesamiento con respecto a su efecto sobre los resultados del algoritmo de minería de datos. Este paso se centra en la comprensibilidad y utilidad del modelo inducido. En este paso también se documenta el conocimiento descubierto, para su uso posterior.
9. **Usar el conocimiento descubierto**. En este paso se incorpora el conocimiento en un sistema diferente, para acciones posteriores. El conocimiento se vuelve activo en el sentido de que podemos hacer cambios en el sistema y medir los efectos que esto causa.

Capítulo 2
Marco Teórico.

2.1 Base de datos (BD)

Una base de datos o banco de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. Actualmente, debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos están en formato digital (electrónico), por lo que se ha desarrollado y se ofrece un amplio rango de soluciones al problema del almacenamiento de datos (7).

Existen programas denominados sistemas manejadores de bases de datos, abreviado DBMS, que permiten almacenar y posteriormente acceder a los datos de forma rápida y estructurada. Las propiedades de estos SGBD, así como su utilización y administración, se estudian dentro del ámbito de la Informática.

2.2 Sistema manejador de bases de datos (DBMS)

Un sistema manejador de bases de datos es una colección de datos interrelacionados y un conjunto de programas que acceden a los mismos. La colección de datos, usualmente referida como base de datos, contiene información relevante para la empresa o campo a explotar. La meta primaria de un DBMS es proveer una manera de almacenar y recuperar información de la base de datos, de manera que sea tanto conveniente como eficiente (8).

Los sistemas de bases de datos están diseñados para manejar grandes cantidades de información, y la administración de los datos incluye tanto estructuras de definición para el almacenamiento de la información como mecanismos usados para la manipulación de la información. Además, los sistemas de bases de datos deben asegurar la información, aun si el sistema llega a dejar de funcionar o en casos de intentos de acceso no autorizados. Si los datos deben ser compartidos entre varios usuarios el sistema debe evitar posibles resultados anómalos.

Las bases de datos son ampliamente usadas, en los bancos, aerolíneas, universidades, transacciones crediticias, telecomunicaciones, finanzas, ventas, manufactura y recursos humanos.

2.3 Instancias y esquemas

Las bases de datos cambian a través del tiempo mientras la información es agregada y eliminada. El conjunto de información almacenado en la base de datos en un momento particular se le llama instancia. El diseño en general de la base de datos se llama el esquema de la base de datos (6). Los esquemas no son cambiados muy frecuentemente, si es que se hace alguna vez.

El concepto de esquemas de bases de datos y de instancias puede ser comprendido con la analogía sobre un programa escrito en un lenguaje de programación. Un esquema de bases de datos corresponde las declaraciones de variables, junto con las definiciones de tipos de datos, en un programa.

Cada variable tiene un valor particular en un instante dado. Los valores de las variables en el programa en un punto particular corresponden a la instancia de un esquema de bases de datos.

Los sistemas de bases de datos tienen varios esquemas, particionados de acuerdo a los niveles de abstracción.

El esquema físico describe el diseño de la base de datos al nivel físico, mientras que el esquema lógico lo define a su nivel lógico. Una base de datos puede también tener varios esquemas en sus diferentes vistas, a veces llamadas sub esquemas, que describen diferentes vistas de las bases de datos.

2.4 Modelos de datos

Subyacente a la estructura de la base de datos se encuentra el modelo de datos, el cual es una colección de herramientas conceptuales para describir datos, sus relaciones, semánticas y restricciones de consistencia (8).

2.4.1 Modelo Entidad-Relación (ER)

El modelo de datos ER se basa en la percepción de un mundo real que consiste en colecciones de objetos básicos llamados entidades, y relaciones entre estos objetos. Una entidad es una cosa u objeto en el mundo real que es distinguible de otros objetos. Por ejemplo cada persona es una entidad y las cuentas de banco se pueden considerar también como tales (7).

Las entidades están descritas en una base de datos por un conjunto de atributos, por ejemplo los atributos número de cuenta y balance podrían describir una cuenta particular en un banco y forman atributos en el conjunto de entidades llamado cuenta. De manera similar los atributos nombre_cliente, calle_cliente y ciudad_cliente pueden describir una entidad llamada cliente.

Una relación es una asociación entre varias entidades. Por ejemplo una relación depósito asociará a un cliente con cada una de las cuentas que tenga. El conjunto de todas las entidades del mismo tipo y el conjunto de todas las relaciones del mismo tipo son llamados un conjunto de entidades o conjunto de relaciones, respectivamente.

El esquema de una base de datos puede ser expresado gráficamente por un diagrama E-R, el cual está construido por los siguientes componentes:

Rectángulos, que representan conjuntos de entidades.

Elipses, que representan atributos.

Rombos, que representan relaciones entre conjuntos de entidades.

Líneas, los cuales enlazan atributos a conjuntos de entidades y conjuntos de entidades a relaciones.

Cada componente es etiquetado con la entidad o relación que representa.

Como ejemplo considere parte de un sistema bancario de bases de datos que consiste en los clientes y las cuentas que estos clientes tienen. El diagrama E-R correspondiente será mostrado en la Figura 2.1. El diagrama indica que hay 2 conjuntos de entidades, clientes y cuentas, con

atributos como en la anterior consideración. El diagrama también muestra una relación depósito entre el cliente y su cuenta.

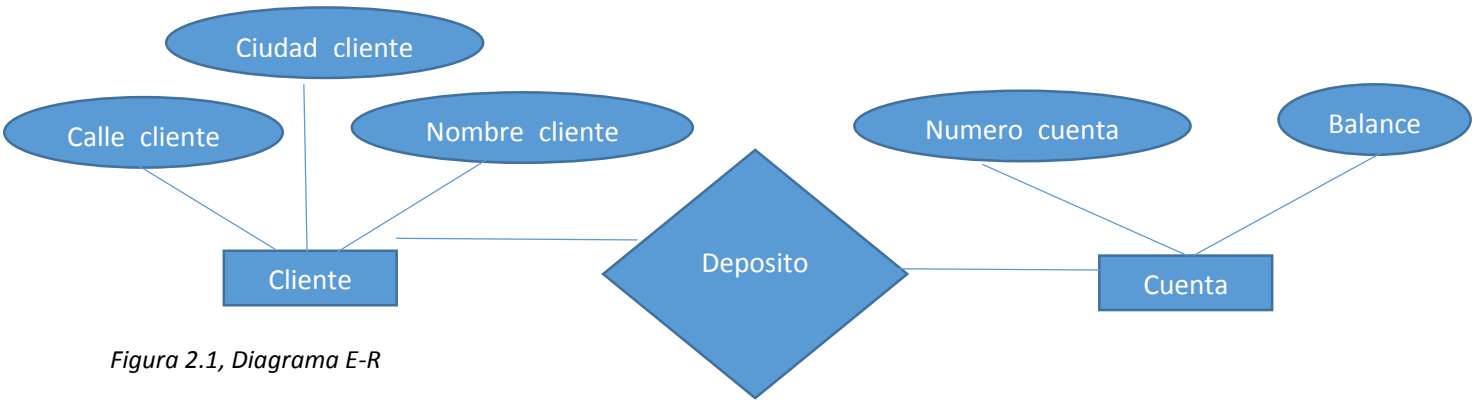


Figura 2.1, Diagrama E-R

En adición a las entidades y relaciones el modelo E-R representa ciertas restricciones a las cuales se debe apegar el contenido de la base de datos. Una restricción importante es la cardinalidad, que expresa el número de entidades a las que otra entidad puede ser asociada mediante un conjunto de relaciones, por ejemplo si cada cuenta debe pertenecer solo a un cliente el modelo E-R puede expresar esa restricción.

2.4.2 Modelo Relacional

El modelo relacional usa un conjunto de tablas para representar tanto los datos como las relaciones entre ellos. Cada tabla tiene múltiples columnas y cada columna tiene un nombre único (8). La siguiente figura muestra una base de datos relacional compuesta por 3 tablas: una muestra los detalles de los clientes del banco, la segunda muestra cuentas y la tercera muestra a que clientes pertenecen estas cuentas. El modelo relacional es un ejemplo de un modelo basado en registros. Los modelos basados en registros son llamados así porque la base de datos está estructurada en registros de formato fijo de varios tipos. Cada tabla contiene registros de un tipo particular. Cada tipo de registro define un número fijo de campos o atributos. Las columnas de la tabla corresponden a los atributos del tipo de registro.

No es difícil ver como las tablas pueden ser almacenadas en archivos. Por ejemplo, un carácter especial, (como la coma), podría ser usado para delimitar los diferentes atributos de un registro, y otro más podría ser usado para delimitar registros. Los modelos relacionales ocultan estos detalles de bajo nivel de los desarrolladores de bases de datos y usuarios en la implementación.

El modelo relacional es el más ampliamente usado en la actualidad y una vasta mayoría de los sistemas de bases de datos están basados en este modelo. El modelo relacional está a un nivel más bajo de abstracción que el modelo E-R. Los diseños de bases de datos generalmente son realizados en modelos E-R y, posteriormente, son traducidos a modelos relacionales. Es fácil ver que las tablas cliente y cuenta corresponden a los conjuntos de entidades del mismo nombre, mientras que la tabla deposito corresponde al conjunto de relaciones deposito.

Nombre_Cliente	Calle_Cliente	Ciudad_Cliente
Jhonson	12 Av Hayes	San Diego
Smith	32 Road Victoria	Palo Alto

Numero_Cuenta	Balance
A-101	500
A-215	700

Nombre_Cliente	Numero_Cuenta
Jhonson	A-101
Smith	A-215

2.5 Lenguajes de bases de datos

Un sistema de bases de datos provee un lenguaje de definición de datos, para especificar el esquema de la base de datos, y un lenguaje de manipulación de datos para expresar consultas y actualizaciones sobre la base de datos. En la práctica los lenguajes de manipulación y definición de datos no son dos lenguajes separados. En vez de eso, simplemente forman parte de un solo lenguaje de bases de datos; tal es el caso del lenguaje SQL (8).

2.5.1 Lenguaje de definición de datos (DDL)

Se especifica un esquema de bases de datos a través de un conjunto de definiciones expresadas por un lenguaje especial llamado lenguaje de definición de datos (DDL) (8).

Por ejemplo la siguiente sentencia en el lenguaje SQL define la tabla cuenta:

Create table cuenta

(numero_cuenta char(10),

Balance integer)

La ejecución de esta sentencia DDL crea la tabla cuenta. Aparte actualiza un conjunto especial de tablas llamadas el diccionario de datos o directorio de datos. Un diccionario de datos contiene metadatos, esto es, datos acerca de los datos. El esquema de una tabla es un ejemplo de metadatos. Un sistema de base de datos consulta el diccionario de datos antes de leer o modificar los datos existentes.

Se especifica la estructura de almacenamiento y métodos de acceso usados por el sistema de bases de datos con un conjunto de sentencias en un tipo especial de DDL llamado lenguaje de almacenamiento y definición de datos. Estas sentencias definen los detalles de implementación de los esquemas de bases de datos, que están usualmente ocultos a los usuarios.

Los valores de los datos almacenados en la base de datos deben satisfacer ciertas restricciones de consistencia. Por ejemplo suponga que el balance de una cuenta no debe bajar de $100. El DDL provee facilidades para especificar estas restricciones. El sistema de bases de datos analiza estas restricciones cada vez que la base de datos es actualizada.

2.5.2 Lenguaje de manipulación de datos (DML)

La manipulación de datos consiste en:

- La recuperación de información almacenada en la base de datos.
- La inserción de nueva información en la base de datos.
- La eliminación de información de la base de datos.
- La modificación de la información en la base de datos.

Un lenguaje de manipulación de datos es un lenguaje que permite a los usuarios acceder o manipular datos según están organizados por el modelo de datos apropiado. Hay básicamente 2 tipos:

DML procedural requiere que un usuario especifique cuales datos se necesitan y como tomar esos datos.

DML declarativo (también referido como DML no procedimental) requiere que el usuario especifique cuales datos se necesitan sin especificar como obtenerlos.

Los DML declarativos son generalmente más fáciles de comprender y usar que los DML procedimentales. Sin embargo, ya que el usuario no tiene que especificar como obtener los datos, el sistema de base de datos tiene que descifrar formas eficientes de acceder a los datos. El componente DML del lenguaje SQL es no procedimental.

Una consulta es una sentencia que solicita la recuperación de información. La porción de DML que se encarga de la recuperación de la información se le llama lenguaje de consulta. Aunque es técnicamente incorrecto, es práctica común el uso de los términos lenguaje de consulta y lenguaje de manipulación de datos como sinónimos.

2.6 Sistemas para el apoyo de decisiones (DSS):

Un DSS es un sistema dinámico que requiere conocimiento más allá de lo que brinda un sistema de información tradicional. Demanda el uso de una base de conocimiento para que el tomador de decisiones esté mejor informado y se le apoye en el procesamiento de información y el análisis de alternativas de solución (9).

Los sistemas para apoyar la toma de decisiones (también conocidos, por sus siglas en inglés, como DSS) son en esencia muy diversos, nacen del interés de aplicar métodos cuantitativos al proceso de toma de decisiones, pero comparten algunas características en común que son listadas a continuación:

- Se emplean en contextos de decisión no estructurados o semiestructurados. Un contexto estructurado es aquel que tiene objetivo no conflictivos, claramente definido, con pocas alternativas de decisión y normalmente son conocidos los efectos de las decisiones.
- Tratan de apoyar al tomador de decisiones más que reemplazarlo.

- Apoyan todas las fases del proceso de toma de decisiones.
- Se enfocan en la efectividad del proceso de toma de decisiones, más que en su eficiencia.
- Usan datos y modelos predefinidos.
- Son interactivos y, en general, amigables.
- Se desarrollan a través de un proceso iterativo y evolutivo.
- Apoyan los niveles estratégicos y tácticos de las empresas.
- Apoyan la toma de decisiones independientes o interdependientes.
- Apoyan contextos de toma de decisiones individuales, grupales y en equipo.

Los componentes de un DSS son:

- El Sistema de Administración de Datos. Compuesto en esencia por la base de datos y una facilidad para realizar consultas.
- El Sistema de Administración de Modelos. Responsable de controlar el almacenamiento y recuperación de los datos relacionados con los modelos cuantitativos, en base a los cuales se estructura el proceso de análisis de decisiones.
- La Máquina de Conocimiento. Permite derivar nuevo conocimiento en base a los datos o al conocimiento previamente derivado. La máquina de conocimiento maneja estrategias de solución específicas de acuerdo al contexto, reglas derivadas, restricciones inherentes a una situación particular, probabilidades a priori. Básicamente podemos ubicar que el conocimiento contenido en un DSS pueden ser HECHOS e HIPOTESIS (reglas o relaciones que se cree existen entre los hechos). La máquina de conocimiento tiene dos funciones básicas:
  a) La Adquisición del Conocimiento; y
  b) La Recuperación del Conocimiento (Máquina de Inferencias).
- La Interfaz con el Usuario. Esta interfaz debería adecuarse al usuario. Se pretende que tenga capacidades de generación de voz, reconocimiento de voz, manejo de lenguaje natural, presentación de los datos, conocimiento, opciones y en general de toda la información de una manera adecuada al usuario (ejemplo de KEIM), se busca agregar afectividad a la interfaz y el uso de agentes inteligentes de manera adicional.
- El Usuario del DSS. Existen diversos tipos de usuarios de un DSS, pero en esencia podemos distinguir: a) El Administrador del DSS; b) El Mantenedor de los Datos, Base de Conocimiento, Reglas, etc.; y c) El Tomador de Decisiones.

2.7 Bodega de Datos (Data Warehouse):

Es un conjunto de datos integrados u orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales apoyan el proceso de toma de decisiones de la administración y está orientada al manejo de grandes volúmenes de datos provenientes de diversas fuentes o diversos tipos (10).

Estos datos cubren largos períodos de tiempo, lo que trae consigo que se tengan diferentes esquemas de los datos fuente, La concentración de esta información está orientada a su análisis para apoyar la toma de decisiones oportunas y fundamentadas. Previo a su utilización se deben aplicar procesos de análisis, selección y transferencia de datos seleccionados desde las fuentes.

El ciclo del desarrollo de la bodega de datos no difiere en mucho de las fases de perfeccionamiento de todos los desarrollos de software. Las fases y las secuencias son las mismas, pero existen variantes únicas asociadas a la bodega de datos y son las siguientes:

Planeación

En esta fase se determinan:

- El enfoque que se optará para la implementación: Top-Down (De Arriba abajo), Bottom-up (De abajo a arriba) o una combinación de estas dos.
- La metodología de desarrollo: Las más usuales son el método de análisis y diseño estructurado y el método del desarrollo en espiral.

Requerimientos

Especificación clara y precisa de las funciones que se esperan obtener de la bodega de datos. Estas deben definirse desde varias perspectivas: propietario, arquitecto o desarrollador de la bodega de datos y desde la visión del usuario. Se definen las áreas tema que apoyará la bodega de datos, las dimensiones de categorización (tiempo, geografía, industria, grupo de clientes, línea de producto, etc.).

Análisis

Consiste en convertir todos los requerimientos conseguidos en la fase anterior en especificaciones concretas que sirvan de base para el diseño. Se definen los modelos lógicos de los datos para la bodega de datos, los mercados de datos, definir los procedimientos de conexión con las fuentes de datos y la bodega de datos y las herramientas de acceso del usuario final.

Diseño

Los modelos lógicos conseguidos en la anterior fase se convierten en modelos físicos. Se generan los diseños para programas y procesos que se requieren según la arquitectura, tanto a nivel de los datos como de aplicación.

Construcción

Se conoce también como diseño físico y consiste en plasmar en la práctica, los diseños lógicos de la fase anterior. Incluye la construcción de programas que creen y modifiquen las bases de datos, que extraigan datos de las fuentes, programas para transformación de datos tales como integración, resumen y adición, programas para la actualización de los datos, programas para búsquedas en bases de datos muy grandes.

Montaje

Relacionados con la instalación, puesta en marcha y uso de la bodega de datos. Un elemento importante consiste en concientizar a los usuarios sobre la disponibilidad, beneficios y presentación de la bodega de datos, esto se conoce como comercialización de la información.

2.8 Modelos de bases de datos multidimensionales

En un modelo de datos multidimensional (esquema mostrado en Figura 2.2) los datos se organizan alrededor de los temas de la organización, formando así la llamada tabla de hechos. La estructura de datos manejada en este modelo son matrices multidimensionales o hipercubos que pueden ser estructurados en diferentes arquitecturas (dependiendo del uso que se le vaya a dar a los datos) y del tipo de los mismos (11).

2.9 Hipercubo

Un hipercubo consiste en un conjunto de celdas, cada una se identifica por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones. Un hipercubo, por tanto, deberá ser reestructurado cada vez que se le agreguen datos o se modifiquen los ya existentes, ya que la información no está en tablas sino organizada de manera dimensional.

2.10 Hecho

Es el objeto a analizar. Posee atributos de tipo cuantitativo llamados de hechos o de síntesis. Sus valores (medidas) se obtienen generalmente por la aplicación de una función estadística que resume un conjunto de valores en un único valor. Por ejemplo: ventas en dólares, cantidad de unidades en inventario, cantidad de unidades de producto vendidas, horas trabajadas, promedio de piezas producidas, consumo de combustible de un vehículo, etcétera.

2.11 Dimensiones

Representan cada uno de los ejes en un espacio multidimensional. Suministran el contexto en el que se obtienen las medidas de un hecho. Algunos ejemplos son: tiempo, producto, cliente, departamento, entre otras. Las dimensiones se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado. Los componentes de una dimensión se denominan niveles y se organizan en jerarquías, verbigracia, la dimensión tiempo puede tener niveles día, mes y año.

Los hechos se guardan en tablas de hechos y las dimensiones en tablas de dimensiones, sin embargo hay diferentes diseños que podemos usar dependiendo de cómo queramos acceder a la información y del tipo de aplicación que vayamos a desarrollar.

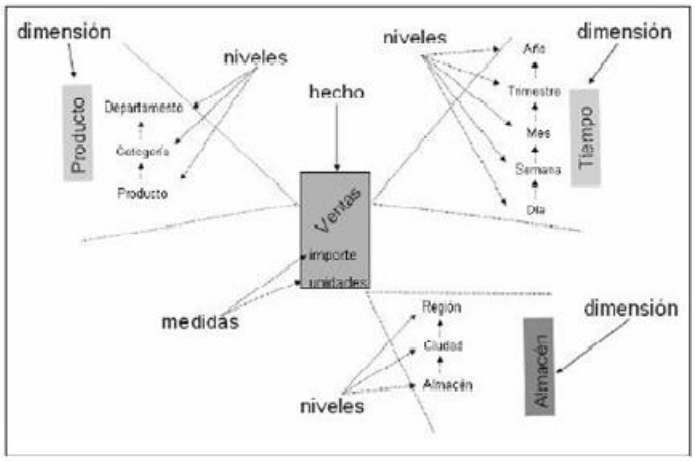


Figura 2.2, Modelo Multidimensional

2.12 OLAP y sus modos de almacenamiento:

OLAP (OnLine Analitical Processing) o procesamiento analítico en línea, es una técnica que nos permite analizar datos en línea de la información contenida en un hipercubo o a un conjunto de tales, dándonos respuestas rápidas a consultas complejas. Debido a las propiedades mencionadas, es típicamente usado para la toma de decisiones, presentando la información a través de sentencias naturales, de las cuales podemos obtener fácilmente conocimiento nuevo viendo patrones o eventos que no podríamos ver a simple vista (12).

OLAP ofrece un conjunto de operadores que facilitan la concepción de consultas, algunos de ellos son Slice & Dice, Swap, Drill Down, Drill Up, Roll-Up, Drill-Across, Drill-Through, etc…

OLAP puede trabajar con los siguientes modos de almacenamiento;

2.12.1 Almacenamiento ROLAP (Relational OLAP)

En este modo se emplea una arquitectura con tres niveles: la base de datos relacional contiene a los datos, el motor OLAP nos da el análisis y alguna de muchas herramientas tanto propietarias como libres. Se usa para presentar la información conformada por los resultados de las consultas que se le hagan al sistema. Así, el motor de análisis OLAP integrado con la herramienta de presentación es la que permite al usuario realizar sus análisis OLAP, convierte sus análisis sobre datos multidimensionales a consultas SQL ejecutadas sobre las BD relacionales, y después se encarga de devolver los resultados (13).

Esta arquitectura es capaz de usar datos que ya se hayan calculado anteriormente en alguna otra consulta o de generar los mismos dinámicamente desde la información en las tablas relacionales. Debido a esto puede tener ciertos retardos en el procesamiento de la consulta. Por ello es que soporta técnicas de optimización para acelerar las mismas, como tablas particionadas, entre otras.

2.12.2 Almacenamiento MOLAP (multidimensional OLAP)

Un sistema MOLAP usa una BD multidimensional (BDMD), en la que la información se almacena multidimensionalmente. El sistema MOLAP utiliza una arquitectura de dos niveles: la BDMD y el motor analítico. La BDMD es la encargada del manejo, acceso y obtención de los datos y el nivel de aplicación es el responsable de la ejecución de las consultas OLAP.

El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP (13).

La información procedente de los sistemas transaccionales se carga en el sistema MOLAP. Una vez cargados los datos en la BDMD, se realiza una serie de cálculos para obtener datos agregados a través de las dimensiones del DW, agregando así nueva información a la BDMD.

Luego de llenar esta estructura, se generan índices y se emplean algoritmos de tablas hash para mejorar los tiempos de accesos de las consultas. Una vez que el proceso de agregado ha

finalizado, la BDMD está lista para su uso. Los usuarios solicitan informes a través de la interfaz y la lógica de aplicación de la BDMD obtiene los datos.

2.12.3 Almacenamiento HOLAP (Hybrid OLAP)

Estas arquitecturas usan aspectos de ambas técnicas de almacenamiento y análisis, ROLAP y MOLAP. En una solución con HOLAP, los registros detallados (los volúmenes más grandes) se mantienen en la BD relacional, mientras que los agregados lo hacen en un almacén MOLAP independiente.

2.13 Minería de Datos:

La minería de datos es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y gestión de datos, procesamiento de datos, el modelo y las consideraciones de inferencia, métricas de Intereses, consideraciones de la Teoría de la Complejidad Computacional, post-procesamiento de las estructuras descubiertas, la visualización y actualización en línea (6).

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis cluster), registros poco usuales (la detección de anomalías) y dependencias (Asociación Minera regla). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada y puede ser utilizado en el análisis adicional o, por ejemplo, en la máquina de aprendizaje y análisis predictivo. Por ejemplo, el paso de minería de datos podría identificar varios grupos en los datos que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones. Ni la recolección de datos, preparación de datos, ni la interpretación de los resultados y la información son parte de la etapa de minería de datos, pero pertenecen a todo el proceso KDD (Knowledge Discovery in Databases) como pasos adicionales.

Los términos relacionados con el dragado de datos, la pesca de datos y espionaje de los datos se refieren a la utilización de métodos de minería de datos a las partes de la muestra que son (o pueden ser) demasiado pequeños para las inferencias estadísticas fiables que se hicieron acerca de la validez de cualquiera de los patrones descubiertos. Estos métodos pueden, sin embargo, ser utilizados en la creación de nuevas hipótesis que se prueban contra las poblaciones de datos más grandes.

Un proceso típico de minería de datos consta de los siguientes pasos generales:

- **Selección del conjunto de datos**, tanto en lo que se refiere a las variables objetivo (aquellas que se quieren predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.

- **Análisis de las propiedades de los datos**, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
- **Transformación del conjunto de datos de entrada**. Se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema. A este paso también se le conoce como preprocesamiento de los datos.
- **Seleccionar y aplicar la técnica de minería de datos**. Se construye el modelo predictivo, de clasificación o segmentación.
- **Extracción de conocimiento**. Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesamiento diferente de los datos.
- **Interpretación y evaluación de datos**, una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

  Si el modelo final no superara esta evaluación el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.

  Una vez validado el modelo, éste ya está listo para su explotación. Los modelos obtenidos por técnicas de minería de datos se aplican incorporándolos en los sistemas de análisis de información de las organizaciones e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje **PMML** (Predictive Model Markup Language), de manera que los modelos de minería de datos sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar.

  Tradicionalmente, las técnicas de minería de datos se aplicaban sobre información contenida en almacenes o bodegas de datos. De hecho, muchas grandes empresas e instituciones han creado y alimentan bases de datos especialmente diseñadas para proyectos de minería de datos en las que centralizan información potencialmente útil de todas sus áreas de negocio. No obstante, actualmente está cobrando una importancia cada vez mayor la minería de datos no estructurados como información contenida en archivos de texto, en Internet, etc.

  Como ya se ha comentado, las técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística. Dichas técnicas no son más que algoritmos más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

  Las técnicas más representativas son:

- **Redes neuronales**.- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de

interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

  - El Perceptrón.
  - El Perceptrón Multicapa.
  - Los Mapas Autoorganizados, también conocidos como redes de Kohonen.

- **Regresión lineal**.- Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.
- **Árboles de decisión**.- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.
- **Modelos estadísticos**.- Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.
- **Agrupamiento o Clustering**.- Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos:
  - Algoritmo K-medias.
  - Algoritmo K-medianas.
- **Reglas de asociación**.- Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados:

- **Algoritmos supervisados** (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.
- **Algoritmos no supervisados** (o del descubrimiento del conocimiento). Se descubren patrones y tendencias en los datos.

La Minería de Datos ha sufrido transformaciones en los últimos años de acuerdo con cambios tecnológicos, de estrategias de marketing, la extensión de los modelos de compra en línea, etc. Los más importantes de ellos son:

- La importancia que han cobrado los datos no estructurados (texto, páginas de Internet, etc.).
- La necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.
- La exigencia de que los procesos funcionen prácticamente en línea (por ejemplo, en casos de fraude con una tarjeta de crédito).
- Los tiempos de respuesta. El gran volumen de datos que hay que procesar en muchos casos para obtener un modelo válido es un inconveniente pues esto implica grandes cantidades de tiempo de proceso y hay problemas que requieren una respuesta en tiempo real.

Capítulo 3 Herramientas.

3.1 Microsoft Excel 2013

Microsoft Excel es una aplicación de hojas de cálculo desarrollada por Windows que incluye herramientas de cálculo, graficación de datos, tablas, y un lenguaje de programación de macros llamado VBA Visual Basic for Applications), Excel forma parte de Microsoft Office.

CICIMAR decidió utilizar esta herramienta debido a la simplicidad de uso del mismo y por claridad a la hora de introducir y analizar los datos.

3.2 Microsoft Excel 2013 VBA

Visual Basic for Applications, (Visual Basic para Aplicaciones), es una implementación del lenguaje de programación orientado a eventos Visual Basic 6 y su ambiente de desarrollo integrado.

VBA nos permite crear funciones definidas por usuario, automatizar procesos y acceder a las APIs de Windows, y otras funcionalidades de bajo nivel, a través de las librerías dinámicas o DLL.

Excel 2013 contiene una implementación de VBA que permite programar macros, las cuales son procedimientos automatizados para facilitar la tarea de la construcción y modificación de los datos y tablas dentro de la hoja de cálculo, esta fue la razón por la cual se decidió utilizar esta herramienta, ya que la información original no tenía el esquema adecuado para el proceso.

3.3 Microsoft SQL Server 2012

Microsoft SQL Server es un Sistema de manejo de bases de datos relacionales desarrollado por Microsoft. Como un software de bases de datos, su función principal es la de almacenar y recuperar información a través de consultas realizadas por otras aplicaciones de software, ya sea en la misma computadora o en otras conectadas por red. Hay muchas implementaciones de SQL server específicas para cargas de trabajo de distintos tamaños y para distintos tipos de aplicaciones, incluyendo distintos números de usuarios concurrentes. Su lenguaje primario de consultas es T-SQL y ANSI SQL (14).

En este proyecto se eligió usar la implementación 2012 debido a que contiene una implementación para inteligencia de negocios, llamada Businnes Intelligence Development Studio que posee una serie de herramientas y procesos para el trabajo con estructuras de datos multidimensionales y modelos de análisis para implementación de algoritmos de minería de datos.

3.4 Microsoft SQL Server Business Intelligence Development Studio

Businnes Intelligence Development Studio (Estudio de desarrollo de inteligencia de negocios) es un ambiente de desarrollo integrado de Windows y se usa para desarrollar análisis de datos e inteligencia de negocios utilizando los servicios de análisis de Microsoft SQL Server, servicios de reportes y servicios integrados.

Está basado en el ambiente de desarrollo de Microsoft Visual Studio, pero se le agregaron servicios específicos de SQL server, así como tipos de proyectos, incluyendo herramientas,

controles y proyectos para reportes, flujos de datos ETL, cubos OLAP, y estructuras de minería de datos (15).

Esta herramienta se utilizó para la construcción de la estructura de datos multidimensional en el proceso de minería de datos, su modificación y extracción de información, así como su posterior análisis.

3.5 Microsoft SQL Server Analysis Services

Microsoft SQL Server Analysis Services (Servicios de Análisis de SQL Server) es una herramienta OLAP, de minería de datos y de reportes de Microsoft SQL Server. Es usada como una herramienta para analizar y dar sentido a información que podría estar repartida en múltiples bases de datos, o en distintas tablas (16).

Este recurso viene integrado en SQL Server como una herramienta de inteligencia de negocios y de bodegas de datos.

Analysis Services se utilizó para realizar el proceso de minería de datos y para el posterior análisis de los resultados obtenidos.

3.6 Microsoft Visual Studio 2012

MS Visual Studio es un ambiente de desarrollo integrado de Microsoft. Se usa para desarrollar programas de computadora para la familia de sistemas operativos Microsoft Windows, así como sitios web, aplicaciones web y servicios web. Visual Studio usa plataformas de desarrollos de software de Windows tales como las Windows API, Formas de Windows, Microsoft Silverlight etc.

MS Visual Studio se utilizó para el desarrollo de una aplicación para la transformación de formato de los datos originales, debido a que su estructura original no era la apropiada.

Capítulo 4
Desarrollo.

4.1 Datos iniciales

Los datos iniciales se obtuvieron en formato de hojas de cálculo en Excel y su estructura se muestra en la Figura 4.1.

IMECOCAL 0001: Preflexion-Transformacion: Datos normalizados

							HABITAT	D	M(MEP)	M	M(MEP)	M(MEP)	D	M(MEP)	D	D
							AFINIDAD	TM-SBTR	SA-TRAN	SA-TRAN	SA-TRAN	TRAN	TA-SBTR	TRAN	TRAN-SBTR	SBTR
							FILOGENIA	187	323	199	201	203	993	397	1331	1333
							FAMILIA	Bathylagidae	Myctophidae	Bathylagidae	Bathylagidae	Bathylagidae	Myctophidae	Myctophidae	Paralepididae	Paralepididae
							REPRODUCCION	TA	Ago-Dic	Feb-Mar-Abr	Ago-Dic	Ago-Dic	Ene-Jun	TA-Ene-Jun	TA-Ene-Jun	Ene-Jun
							CODIGO CALCOFI	56	151	68	69	71	811	226	921	60
							CODIGO CALCOFI	ARGSIA	ARISCI	BATOCH	BATPAC	BATWES	CAUPRI	CERTOW	CITFRA	CITGOR
EST	LAT	LONG	T SUP.	S SUP.	F.E.A.	B.Z.ml/1000m³	B.Z. S/SALPAS ml/1000m³	*Argentina sialis* Gilbert, 1890			*Bathylagus pacificus* Gilbert, 1890	*Bathylagoides wesethi* Bolin, 1938			*Citharichthys fragilis* Gilbert, 1890	
100.30	31.67	-116.78	15.26	32.62	7.7197	1.71	1.71	0	0	0	0	0	0	0	31	0
100.35	31.51	-117.12	15.95	33.12	7.1173	109.24	109.24	0	0	0	0	0	0	0	0	0
100.40	31.35	-117.45	15.91	33.33	5.9190	163.82	163.82	0	0	0	0	0	0	0	18	0
100.45	31.19	-117.78	15.56	33.50	8.2866	156.72	156.72	0	0	0	0	0	0	0	0	0
100.50	30.99	-118.14		33.54	8.1626	124.75	124.75	0	0	0	16	0	0	0	24	0
100.55	30.83	-118.44	14.70	33.48	7.4932	115.54	115.54	0	0	0	15	0	0	0	7	0
100.60	30.67	-118.78	16.45	33.50	7.6598	165.77	157.80	0	0	0	8	0	0	0	0	0
103.30	31.10	-116.41	15.35	33.28	4.8249	126.54	126.54	0	0	0	0	0	0	0	0	0
103.35	30.94	-116.75	15.96	33.28	7.6350	84.92	84.92	0	0	0	0	0	0	8	0	0
103.40	30.77	-117.08	16.07	33.27	7.6959	113.61	113.61	0	0	0	0	0	0	0	0	0
103.45	30.60	-117.41	15.88	33.33	6.5719	123.30	123.30	0	0	0	0	0	0	0	0	0
103.50	30.42	-117.74	15.74	33.35	7.4010	334.34	334.34	0	0	0	0	0	0	0	30	0
103.55	30.27	-118.07	15.30	33.25	5.9317	161.74	161.74	0	0	12	0	0	0	12	24	0
103.60	30.10	-118.41	15.34	33.23	6.4624	213.82	213.82	0	0	6	0	0	0	0	52	0

Figura 4.1, Datos Iniciales

Estos datos nos muestran información acerca de muestreos de larvas de peces realizados en distintos puntos (estaciones) de la Península de Baja California. En cada uno de estos registros se recabó información acerca de una serie de atributos, los cuales denotan la siguiente información;

Atributos de Estación

EST

Estación, identificador numérico de la estación en la que se realizó el muestreo.

LAT

Latitud de la estación en la que se realizó el muestreo.

LONG

Longitud de la estación en la que se realizó el muestreo.

T.SUP

Temperatura superficial del agua en la estación, al momento de realizar el muestreo.

S.SUP

Salinidad superficial del agua en la estación, al momento de realizar el muestreo.

B.Z. ml/1000m^3

Cantidad de biomasa de zooplancton medida en mililitros por cada 1000 metros cúbicos de agua en la estación, al momento de realizar el muestreo.

Atributos de Especie

Hábitat

Los huevos y larvas de peces pertenecen al grupo animal denominado zooplancton. Éste a su vez es una división del Pláncton(que significa "Errante"), y contiene a todos los organismos (virus, bacterias, plantas y animales) que viven en la columna de agua y no tienen capacidad de vencer las corrientes, es decir, que su distribución y sobrevivencia depende de las condiciones hidrológicas (corrientes, temperatura, salinidad, profundidad) del cuerpo de agua que habitan y no pueden desplazarse por medios propios hacia cuerpos de agua o regiones con características más favorables.

El plancton se divide en dos grandes grupos, el Holoplancton, que contiene a los organismos que pasan toda su vida perteneciendo al plancton, y el Meroplancton, que contiene organismos que solo durante parte de su ciclo de vida pertenecen al plancton, como los huevos y las larvas de peces que en su vida juvenil y adulta tienen facultades y estructuras para vencer las corrientes y desplazarse a voluntad hacia sitios de crianza y reproducción favorables.

El hábitat establece el ambiente en el que los juveniles y adultos de la especie se van a desarrollar y reproducir. Se relaciona con las condiciones geológicas como distancia costa-océano, tipo de fondo (rocoso, arenoso, arrecifal), y todo esto se relaciona con las condiciones locales del ambiente en donde se distribuyen las especies. Si las larvas en los estadios de desarrollo más avanzados no se encuentran cerca del hábitat donde se distribuyen los juveniles y/o adultos provocaría la muerte de estas larvas, ya que no tendrían los requerimientos necesarios para continuar su desarrollo. A la larga, esto representa una disminución en las poblaciones adultas.

Atributos de Hábitat

–	No aplica
B	Batypelágico
BD	Bento-demersal
BP	Bento-pelágico
D	Demersal
PO	Pelágico oceánico
RA	Arrecifal

Afinidad

La afinidad establece los rangos de distribución latitudinal de las especies (norte-sur; Polo-Ecuador). Un cambio en la distribución de las especies y/o sus abundancias con respecto a su afinidad nos indica eventos de calentamiento y enfriamiento, de cambio en la intensidad de las corrientes. Todos estos cambios están relacionados con condiciones anómalas del ambiente natural donde se distribuyen las especies.

Atributos de Afinidad

−	No aplica
CWM	Masa de Agua central del Pacífico
ECP	Pacífico Central Este
SA-SBTR	Subartico-Subtropical
SA-TM	Subartico-Templado
SA-TR	Subartico-Tropical
SBTR	Subtropical
TM	Templado
TM-SBR	Templado-Subtropical
TM-TR	Templado-Tropical
TR	Tropical
TR-SBTR	Tropical-Subtropical

Reproducción

Los peces adultos tienen una distribución muy restringida en espacio y tiempo durante la reproducción. Cada especie tiene características particulares con el fin de asegurar las mejores condiciones para el desarrollo de los huevos y las larvas ya que, al ser planctónicos, necesitan contar desde el principio de su desarrollo con un ambiente en condiciones favorables.

Esta cualidad de ser depositadas para su desarrollo en sitios con características ambientales muy estrictas y definidas hace que las larvas de peces sean indicadores biológicos de gran utilidad de condiciones y cambios en el ambiente costero y oceánico, ya que su presencia, ausencia y abundancia nos indicará si estas condiciones se han modificado a escala local, latitudinal y temporal.

Atributos de Reproducción

V	Verano
P	Primavera
O	Otoño
I	Invierno
pico	Periodo en el que se van a encontrar las mayores abundancias de huevos y larvas en el ambiente
?	Se desconoce para la especie, no hay estudios al respecto
−	No se cuenta con el dato por falta de identificación a nivel de especie

Familia

La familia es el grupo filogenético (evolutivo) al que pertenecen las especies.

Filogenia

Orden evolutivo de las especies o taxa.

Atributo de Filogenia

323 *Aristostomias scintillans* (Gilbert, 1915)
617 *Aulopus bajacali* Parin & Kotlyar, 1984
- No se cuenta con el dato por falta de identificación a nivel de especie

Código Calcolfi

Es un código numérico empleado para hacer equivalentes las bases de datos de larvas de CalCOFI (California Cooperative Oceanic Fisheries Investigations) con investigaciones bases de datos de larvas de peces de la Corriente de California.

Atributo Calcofi

068 *Lipolagus ochotensis* (Schmidt, 1938)
071 *Brama japonica* Hilgendorf, 1878
552 *Ceratoscopelus townsendi* (Eigenmann & Eigenmann, 1889)
- No se cuenta con el dato por falta de identificación a nivel de especie

4.2 Transformación de los datos.

Para poder iniciar el estudio de los datos iniciales fue necesario realizar una serie de transformaciones de estructura a la base de datos, así como la agregación de algunos atributos.

IMECOCAL 0097: DATOS NORMALIZADOS: TRANSFORMACION

							HABITAT	PC		D	D+	D	M	M(MEP)	M(MEP)	D	EPI-BATI	D	M(E-B)	M(MEP)	M
							AFINIDAD							TRAN			ETP			TRAN	
							FILOGENIA	175	355	187	831	1021	199	203	967	1015	975	993	585	397	1135
							FAMILIA														
							REPRODUCCION	V/O-I	TA	TA					TA	V-O	TA		V		
							CODIGO CALCOFI	26/33	317	056	748	609	068	071	553	607	559	811	979	226	816
							CODIGO CALCOFI	ENGSP1	ARCRIS	ARGSIA	ARTLAT	ATRNOB	BATOCH	BATWES	BRAJAP	CALBRA		CAUPRI	CREHOL	CERTOW	CHINIG
EST	LAT	LONG	FECHA	HORA	T SUP	SAL SUP	BZ	Engraulidae type 1		*Argentina sialis* Gilbert, 1890	*Artedius lateralis* (Girard, 1854)	*Atractoscion nobilis* (Ayres, 1860)		*Bathylagoides wesethi* Bolin, 1938	*Brama japonica* Hilgendorf, 1878				*Ceratias holboelli* Krøyer, 1845		*Chiasmodon niger* Johnson, 1864
100.30	31.69	116.78	06/10/97	1	18.1	33.5822	1460.73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.35	31.49	117.11	06/10/97	0	18.85	33.594	90.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.40	31.56	117.76	06/10/97	1	19.01	33.6351	93.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.45	31.18	117.78	05/10/97	1	19.07	33.6	127.53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.50	31.03	118.10	05/10/97	1	19.37	33.601	288.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.55	30.86	118.45	06/10/97	1	18.43	33.59	165.49	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.60	31.10	119.32	06/10/97	0	18.98	33.61	184.29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.30	30.09	116.39	06/10/97	1	16.22	33.5729	1050.53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.35	30.94	116.74	05/10/97	1	18.3	33.5561	238.46	0	0	0	0	0	0	12	0	0	0	0	0	0	0
103.40	31.27	117.13	05/10/97	1	17.84	33.6	208.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.45	30.73	117.44	06/10/97	0	19.99	33.5368	234.18	0	5	0	0	0	0	0	0	0	0	0	0	0	0
103.50	30.43	117.75	06/10/97	0	18.95	33.62	185.34	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.55	30.28	118.08	06/10/97	0	18.94	33.5329	164.23	0	0	0	0	0	5	0	0	0	0	0	0	0	0
103.60	30.12	118.40	05/10/97	0	18.56	33.507	485.64	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107.32	30.45	116.15	05/10/97	0	17.41	33.54	453.44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107.35	30.36	116.36	06/10/97	0	18.57	33.5904	341.00	0	0	0	0	0	0	11	0	0	0	0	0	0	0
107.40	30.19	116.69	06/10/97	0	19.41	33.69	110.34	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107.45	30.00	116.94	06/10/97	0	18.23	33.5672	256.42	0	0	0	0	0	0	13	0	0	0	0	0	0	0
107.50	29.85	117.35	05/10/97	1	18.76	33.45	172.88	0	0	0	0	0	0	7	0	0	0	0	0	0	0

Figura 4.2, Inclusión de atributos fecha y horario a la base de datos.

El primer cambio que se realizó fue la inclusión de los atributos de fecha y horario de cada muestreo como se muestra en la Figura 4.2.

La fecha incluye el día, mes y año del muestreo. El horario se refiere a en que momento del día se realizó el muestro (0 = noche, 1 = día). El horario se agregó debido a que las condiciones ambientales e hidrológicas en las estaciones pueden variar en gran manera de día a noche.

Una vez incluidos estos datos se realizó una simplificación de los nombres de los atributos para facilitar su manejo en la base de datos, según se muestra en la Figura 4.3.

IMECOCAL 0001: DATOS NORMALIZADOS: TRANSFORMACION																				
							NUM	1	2	3	4	5	6	7	8	9	10	11	12	13	14
							HABITAT	PC		D	D+	D	M					D			M
							AFINIDAD							TRAN			ETP			TRAN	
							FILOGENIA	175	355	187	831	1021	199	203	967	1015	975	993	585	397	1135
							FAMILIA														
							REPRODUCCION	V/O-I	TA	TA					TA	V-O	TA		V		
							CODIGO CALCOFI	26/33	317	056	748	609	068	071	553	607	559	811	979	226	816
							CODIGO CALCOFI														
EST	LAT	LONG	FECHA	HORA	TS	SS	BZ	Engraulidae type 1		*Argentina sialis* Gilbert, 1890	*Artedius lateralis* (Girard, 1854)	*Atractoscion nobilis* (Ayres, 1860)		*Bathylagoides wesethi* Bolin, 1938	*Bvarna japonica* Hilgendorf, 1878				*Ceratias holboelli* Krøyer, 1845		*Chiasmodon niger* Johnson, 1864
100.30	31.69	116.78	16/01/01	1	18.1	33.5822	1460.73	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.35	31.49	117.11	16/01/01	1	18.85	33.594	90.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.40	31.56	117.76	17/01/01	1	19.01	33.6351	93.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.45	31.18	117.78	17/01/01	1	19.07	33.6	127.53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.50	31.03	118.10	17/01/01	0	19.37	33.601	288.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.55	30.86	118.45	17/01/01	0	18.43	33.59	165.49	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100.60	31.10	119.32	19/01/01	0	18.98	33.61	184.29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.30	30.09	116.39	19/01/01	1	16.22	33.5729	1050.53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.35	30.94	116.74	19/01/01	0	18.3	33.5561	238.46	0	0	0	0	0	0	12	0	0	0	0	0	0	0
103.40	31.27	117.13	19/01/01	1	17.84	33.6	208.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.45	30.73	117.44	20/01/01	0	19.99	33.5368	234.18	0	5	0	0	0	0	0	0	0	0	0	0	0	0
103.50	30.43	117.75	20/01/01	0	18.95	33.62	185.34	0	0	0	0	0	0	0	0	0	0	0	0	0	0
103.55	30.28	118.08	21/01/01	0	18.94	33.5329	164.23	0	0	0	0	0	5	0	0	0	0	0	0	0	0
103.60	30.12	118.40	21/01/01	0	18.56	33.507	485.64	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107.32	30.45	116.15	21/01/01	1	17.41	33.54	453.44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107.35	30.36	116.36	22/01/01	1	18.57	33.5904	341.00	0	0	0	0	0	0	11	0	0	0	0	0	0	0
107.40	30.19	116.69	22/01/01	1	19.41	33.69	110.34	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107.45	30.00	116.94	23/01/01	1	18.23	33.5672	256.42	0	0	0	0	0	0	13	0	0	0	0	0	0	0
107.50	29.85	117.35	23/01/01	1	18.76	33.45	172.88	0	0	0	0	0	0	7	0	0	0	0	0	0	0

Figura 4.3, Simplificación de los nombres de atributos

Simplificación de los nombres de atributos

Como vemos en la figura 12.3 se cambiaron los nombres de **T.SUP** a **TS**, de **S.SUP** a **SS**, y de **B.Z. ml/1000m^3**a **BZ.**

Una vez que se incluyeron los atributos necesarios y se simplificó su identificador se procedió a la transformación de la base de datos a una base de datos relacional en SQL Server.

Para esto se programó una aplicación en Visual Studio 2012 que permite tomar estos datos a partir de su texto plano y modificar su estructura, de manera que fuera posible exportarla directamente a SQL Server 2012.

La parte relevante del código se muestra en la Figura 4.4:

```
char hora[100];
char ts[100];
char ss[100];
char bz[100];
cadena[0]='\0';


while(!feof(archivo)) {
    car=fgetc(archivo);

    if (car=='\t' || car=='\n') {
        if (strlen(cadena)!=0) {
            if (fila==1 && columna>=10) {
                strcpy(especies[columna-10],cadena);
                if (columna%150==0)
                    int a=0;
            }
            if (columna==1) strcpy(estadio,cadena);
            if (columna==2) strcpy(estacion,cadena);
            if (columna==3) strcpy(latitud,cadena);
            if (columna==4) strcpy(longitud,cadena);
            if (columna==5) strcpy(fecha,cadena);
            if (columna==6) strcpy(hora,cadena);
            if (columna==7) strcpy(ts,cadena);
            if (columna==8) strcpy(ss,cadena);
            if (columna==9) strcpy(bz,cadena);
            if (fila>=2 && columna>=10)
                //if (strcmp(cadena,"0")!=0)
fprintf(nuevo,"%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\n",estadio,estacion,latitud,longitud,fecha,hora,ts,ss,bz,especies[columna-10],cadena);

            cadena[0]='\0';
            pos=0;
            columna++;
            if (car=='\n') {
                fila++;
                columna=1;
            }
```

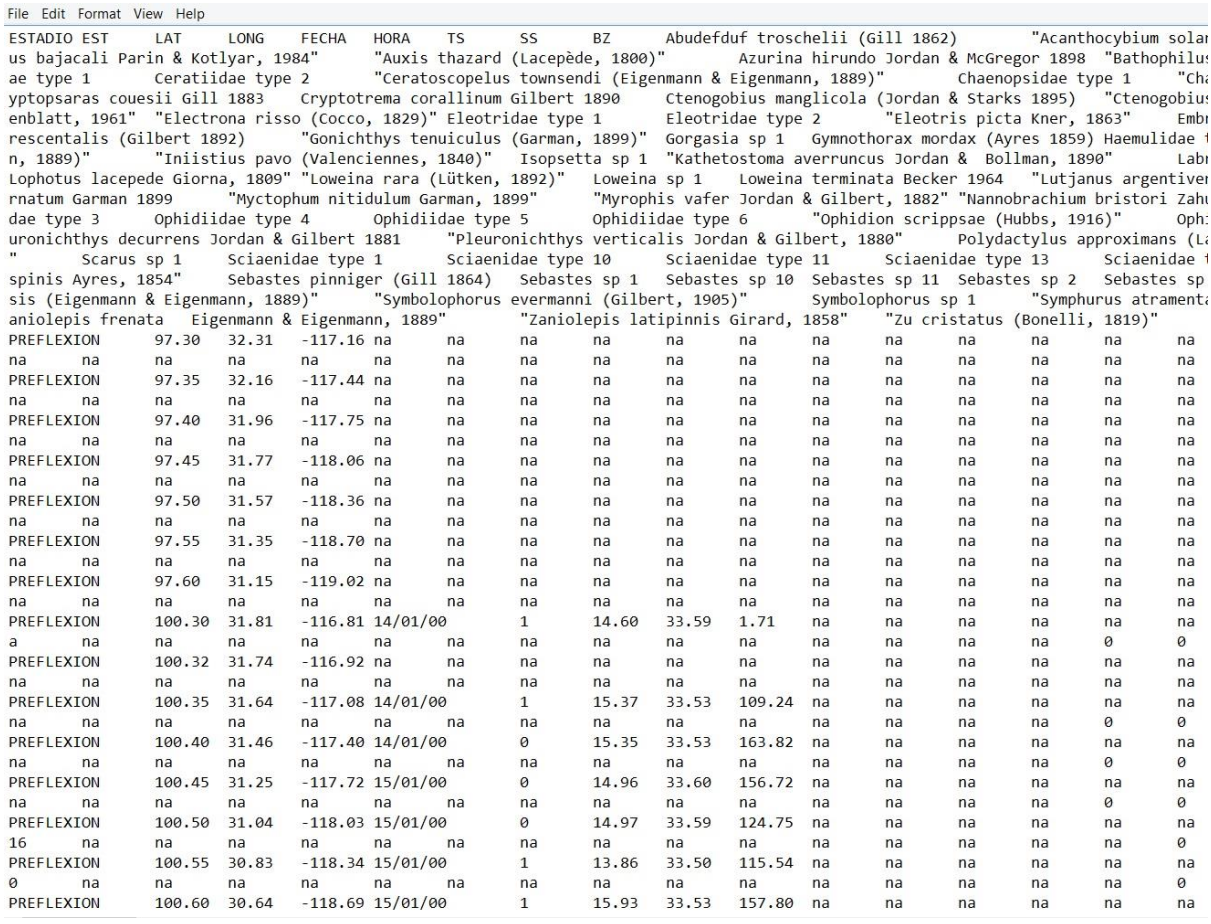
Figura 4.4, Código para transformación de estructura

Transformación de la estructura de la información

Una vez que se creó este código fue necesario hacer una transformación en la estructura de los registros de la base de datos, de manera que fuera posible tomar un registro de muestra de estación por cada especie encontrada. Después esta estructura fue almacenada en texto plano para su posterior transformación con el programa. Los pasos intermedios y estructura final se muestran en las Figuras 4.5, 4.6 y 4.7.

ESTADIO	EST	LAT	LONG	FECHA	HORA	TS	SS	BZ	[illegible]	[illegible]	*Albula* sp 1
PREFLEXION	97.30	32.31	-117.16	na	na	na	na	na	na	na	na
PREFLEXION	97.35	32.16	-117.44	na	na	na	na	na	na	na	na
PREFLEXION	97.40	31.96	-117.75	na	na	na	na	na	na	na	na
PREFLEXION	97.45	31.77	-118.06	na	na	na	na	na	na	na	na
PREFLEXION	97.50	31.57	-118.36	na	na	na	na	na	na	na	na
PREFLEXION	97.55	31.35	-118.70	na	na	na	na	na	na	na	na
PREFLEXION	97.60	31.15	-119.02	na	na	na	na	na	na	na	na
PREFLEXION	100.30	31.81	-116.81	14/01/00	1	14.60	33.59	1.71	na	na	na
PREFLEXION	100.32	31.74	-116.92	na	na	na	na	na	na	na	na
PREFLEXION	100.35	31.64	-117.08	14/01/00	1	15.37	33.53	109.24	na	na	na
PREFLEXION	100.40	31.46	-117.40	14/01/00	0	15.35	33.53	163.82	na	na	na
PREFLEXION	100.45	31.25	-117.72	15/01/00	0	14.96	33.60	156.72	na	na	na
PREFLEXION	100.50	31.04	-118.03	15/01/00	0	14.97	33.59	124.75	na	na	na
PREFLEXION	100.55	30.83	-118.34	15/01/00	1	13.86	33.50	115.54	na	na	na
PREFLEXION	100.60	30.64	-118.69	15/01/00	1	15.93	33.53	157.80	na	na	na
PREFLEXION	103.30	31.18	-116.40	16/01/00	0	14.84	33.51	126.54	na	na	na
PREFLEXION	103.33	31.10	-116.52	na	na	na	na	na	na	na	na
PREFLEXION	103.35	31.02	-116.65	16/01/00	0	15.49	33.52	84.92	na	na	na
PREFLEXION	103.40	30.82	-116.97	16/01/00	1	15.47	33.62	113.61	na	na	na
PREFLEXION	103.45	30.61	-117.30	16/01/00	1	15.22	33.56	123.30	na	na	na
PREFLEXION	103.50	30.42	-117.61	16/01/00	0	15.27	33.57	334.34	na	na	na
PREFLEXION	103.55	30.22	-117.93	16/01/00	0	14.43	33.47	161.74	na	na	na
PREFLEXION	103.60	30.03	-118.28	15/01/00	0	14.78	33.45	213.82	na	na	na
PREFLEXION	107.32	30.56	-116.12	17/01/00	0	16.15	33.57	98.23	na	na	na
PREFLEXION	107.33	30.52	-116.20	na	na	na	na	na	na	na	na
PREFLEXION	107.35	30.46	-116.30	17/01/00	1	15.29	33.52	110.10	na	na	na
PREFLEXION	107.40	30.26	-116.60	17/01/00	1	14.92	33.48	85.76	na	na	na
PREFLEXION	107.45	30.04	-116.92	17/01/00	1	15.27	33.61	441.58	na	na	na
PREFLEXION	107.50	29.85	-117.23	17/01/00	0	16.42	33.49	87.66	na	na	na
PREFLEXION	107.55	29.66	-117.56	18/01/00	0	16.51	33.58	73.23	na	na	na
PREFLEXION	107.60	29.45	-117.90	18/01/00	1	16.74	33.60	43.18	na	na	na
PREFLEXION	110.34	29.92	-115.78	na	na	na	na	na	na	na	na
PREFLEXION	110.35	29.87	-115.87	19/01/00	1	14.98	33.51	55.28	na	na	na

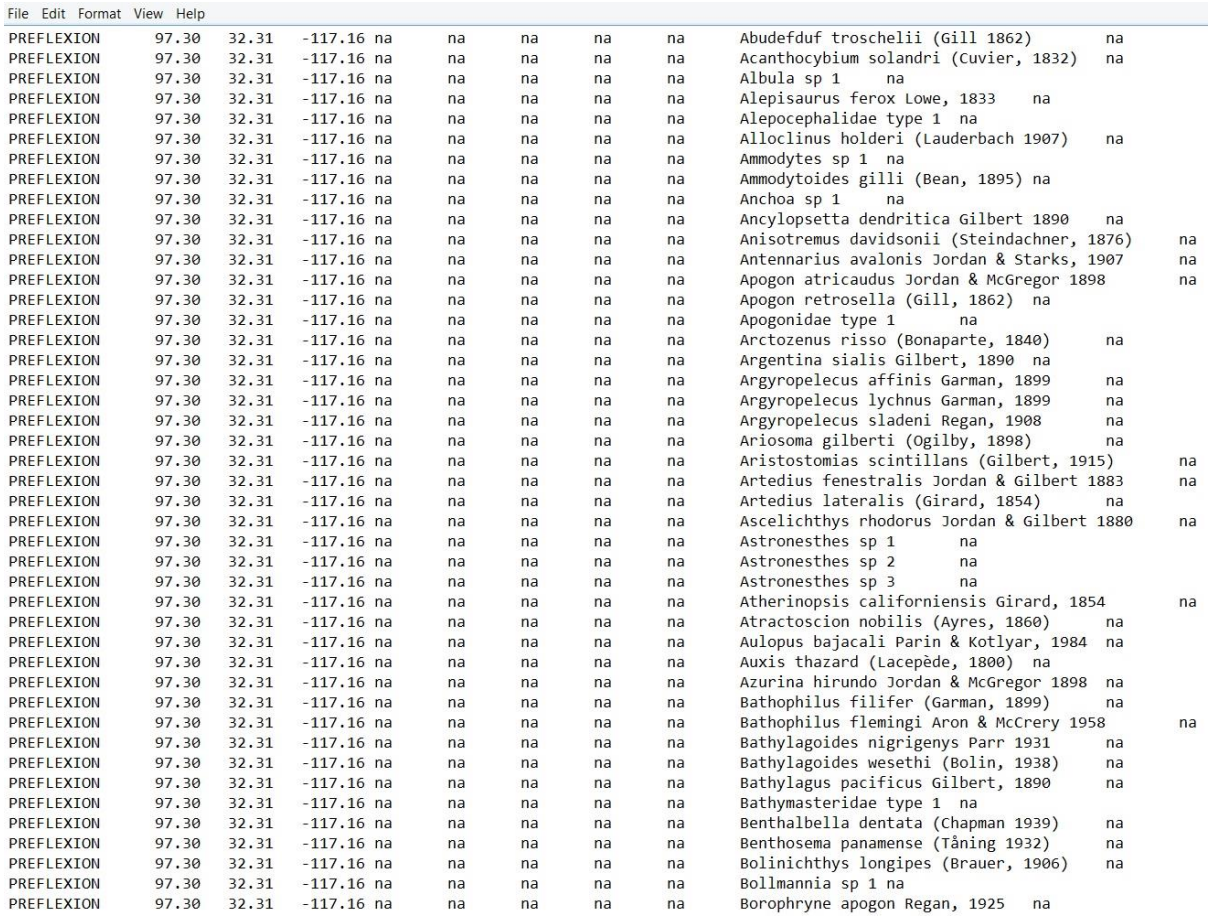
Figura 4.5, Estructura intermedia



File Edit Format View Help

```
ESTADIO EST     LAT     LONG    FECHA   HORA    TS      SS      BZ      Abudefduf troschelii (Gill 1862)        "Acanthocybium solar
us bajacali Parin & Kotlyar, 1984"      "Auxis thazard (Lacepède, 1800)"        Azurina hirundo Jordan & McGregor 1898  "Bathophilus
ae type 1       Ceratiidae type 2       "Ceratoscopelus townsendi (Eigenmann & Eigenmann, 1889)"        Chaenopsidae type 1     "Cha
yptopsaras couesii Gill 1883    Cryptotrema corallinum Gilbert 1890     Ctenogobius manglicola (Jordan & Starks 1895)   "Ctenogobius
enblatt, 1961"  "Electrona risso (Cocco, 1829)" Eleotridae type 1        Eleotridae type 2       "Eleotris picta Kner, 1863"     Embr
rescentalis (Gilbert 1892)      "Gonichthys tenuiculus (Garman, 1899)"  Gorgasia sp 1   Gymnothorax mordax (Ayres 1859) Haemulidae 1
n, 1889)"       "Iniistius pavo (Valenciennes, 1840)"   Isopsetta sp 1  "Kathetostoma averruncus Jordan &  Bollman, 1890"       Labr
Lophotus lacepede Giorna, 1809" "Loweina rara (Lütken, 1892)"   Loweina sp 1    Loweina terminata Becker 1964   "Lutjanus argentiver
rnatum Garman 1899      "Myctophum nitidulum Garman, 1899"      "Myrophis vafer Jordan & Gilbert, 1882" "Nannobrachium bristori Zahu
dae type 3      Ophidiidae type 4       Ophidiidae type 5       Ophidiidae type 6       "Ophidion scrippsae (Hubbs, 1916)"      Ophi
uronichthys decurrens Jordan & Gilbert 1881     "Pleuronichthys verticalis Jordan & Gilbert, 1880"      Polydactylus approximans (La
"       Scarus sp 1     Sciaenidae type 1       Sciaenidae type 10      Sciaenidae type 11      Sciaenidae type 13      Sciaenidae 1
spinis Ayres, 1854"     Sebastes pinniger (Gill 1864)   Sebastes sp 1   Sebastes sp 10  Sebastes sp 11  Sebastes sp 2   Sebastes sp
sis (Eigenmann & Eigenmann, 1889)"      "Symbolophorus evermanni (Gilbert, 1905)"       Symbolophorus sp 1      "Symphurus atramenta
aniolepis frenata   Eigenmann & Eigenmann, 1889"        "Zaniolepis latipinnis Girard, 1858"    "Zu cristatus (Bonelli, 1819)"
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      97.35   32.16   -117.44 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      97.40   31.96   -117.75 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      97.45   31.77   -118.06 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      97.50   31.57   -118.36 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      97.55   31.35   -118.70 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      97.60   31.15   -119.02 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      100.30  31.81   -116.81 14/01/00        1       14.60   33.59   1.71    na      na      na      na      na      na
a       na      na      na      na      na      na      na      na      na      na      na      na      0       0
PREFLEXION      100.32  31.74   -116.92 na      na      na      na      na      na      na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      na      na      na
PREFLEXION      100.35  31.64   -117.08 14/01/00        1       15.37   33.53   109.24  na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      0       0
PREFLEXION      100.40  31.46   -117.40 14/01/00        0       15.35   33.53   163.82  na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      0       0
PREFLEXION      100.45  31.25   -117.72 15/01/00        0       14.96   33.60   156.72  na      na      na      na      na      na
na      na      na      na      na      na      na      na      na      na      na      na      na      0       0
PREFLEXION      100.50  31.04   -118.03 15/01/00        0       14.97   33.59   124.75  na      na      na      na      na      na
16      na      na      na      na      na      na      na      na      na      na      na      na      na      na      0
PREFLEXION      100.55  30.83   -118.34 15/01/00        1       13.86   33.50   115.54  na      na      na      na      na      na
0       na      na      na      na      na      na      na      na      na      na      na      na      na      na      0
PREFLEXION      100.60  30.64   -118.69 15/01/00        1       15.93   33.53   157.80  na      na      na      na      na      na
```

Figura 4.6, Texto plano de la estructura intermedia



File Edit Format View Help

```
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Abudefduf troschelii (Gill 1862)        na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Acanthocybium solandri (Cuvier, 1832)   na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Albula sp 1     na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Alepisaurus ferox Lowe, 1833    na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Alepocephalidae type 1  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Alloclinus holderi (Lauderbach 1907)    na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Ammodytes sp 1  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Ammodytoides gilli (Bean, 1895) na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Anchoa sp 1     na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Ancylopsetta dendritica Gilbert 1890    na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Anisotremus davidsonii (Steindachner, 1876)     na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Antennarius avalonis Jordan & Starks, 1907      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Apogon atricaudus Jordan & McGregor 1898        na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Apogon retrosella (Gill, 1862)  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Apogonidae type 1       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Arctozenus risso (Bonaparte, 1840)      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Argentina sialis Gilbert, 1890  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Argyropelecus affinis Garman, 1899      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Argyropelecus lychnus Garman, 1899      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Argyropelecus sladeni Regan, 1908       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Ariosoma gilberti (Ogilby, 1898)        na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Aristostomias scintillans (Gilbert, 1915)       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Artedius fenestralis Jordan & Gilbert 1883      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Artedius lateralis (Girard, 1854)       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Ascelichthys rhodorus Jordan & Gilbert 1880     na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Astronesthes sp 1       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Astronesthes sp 2       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Astronesthes sp 3       na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Atherinopsis californiensis Girard, 1854        na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Atractoscion nobilis (Ayres, 1860)      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Aulopus bajacali Parin & Kotlyar, 1984  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Auxis thazard (Lacepède, 1800)  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Azurina hirundo Jordan & McGregor 1898  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bathophilus filifer (Garman, 1899)      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bathophilus flemingi Aron & McCrery 1958        na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bathylagoides nigrigenys Parr 1931      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bathylagoides wesethi (Bolin, 1938)     na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bathylagus pacificus Gilbert, 1890      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bathymasteridae type 1  na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Benthalbella dentata (Chapman 1939)     na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Benthosema panamense (Tåning 1932)      na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bolinichthys longipes (Brauer, 1906)    na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Bollmannia sp 1 na
PREFLEXION      97.30   32.31   -117.16 na      na      na      na      na      Borophryne apogon Regan, 1925   na
```

Figura 4.7, Estructura final después de la utilización del programa de transformación

Una vez obtenido un registro por cada muestreo de estación y por cada especie, se agregó el campo **CANTIDAD**, que indica el número de larvas encontradas en ese registro.

La estructura final de la base de datos, vista en Excel, se muestra en la Figura 4.8.

Con esta estructura final se procedió a exportar la base de datos a SQL Server para la creación del modelo de minería y para su posterior análisis a través del módulo Analiysis Services de SQL Server 2012.

A	B	C	D	E	F	G	H	I	J	K
ESTADIO	EST	LAT	LONG	FECHA	HORA	TS	SS	BZ	ESPECIE	CANTIDAD
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Syacium sp 2	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symbolophorus californiensis (Eigenmann & Eigenmann, 1889)	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symbolophorus evermanni (Gilbert, 1905)	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symbolophorus sp 1	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symphurus atramentatus Jordan & Bollman, 1890	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symphurus atricaudus (Jordan & Gilbert, 1880)	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symphurus sp 11	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Symphurus williamsi Jordan & Culver, 1895	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Synchiropus atrilabiatus (Garman, 1899)	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Syngnathidae type 1	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Syngnathus californiensis Storer, 1845	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Synodus lucioceps (Ayres, 1855)	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Synodus sp 1	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Taractichthys steindachneri (Döderlein, 1883)	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Tarletonbeania crenularis (Jordan & Gilbert, 1880)	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Tarletonbeania sp 1	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	TAXA 1	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	TAXA 2	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	TAXA 3	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	TAXA 4	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	TAXA 5	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	TAXA 6	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Tetragonurus atlanticus Lowe, 1839	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Tetragonurus cuvieri Risso, 1810	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Trachipterus altivelis Kner, 1859	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Trachurus symmetricus (Ayres, 1855)	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Trichiurus nitens Garman 1899	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Triphoturus mexicanus (Gilbert, 1890)	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Triphoturus nigrescens (Brauer, 1904)	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Typhlogobius californiensis Steindachner, 1879	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Umbrina roncador Jordan & Gilbert, 1882	0
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Umbrina sp 1	na
PREFLEXION	103.5	30.42	-117.61	26/01/98	1	16.9	33.58	148.54	Unidentified fish larvae	0

Figura 4.8, Estructura final en Excel

4.3 Creación de la base de datos

Una vez obtenida la estructura final de información, se procedió a la transformación de la misma a bases de datos relacionales en Microsoft SQL Server 2012 a través de su herramienta de exportación.

Primeramente se creó la base de datos en la que se almacenarán las tablas de la estructura, como se muestra en la Figura 4.9.

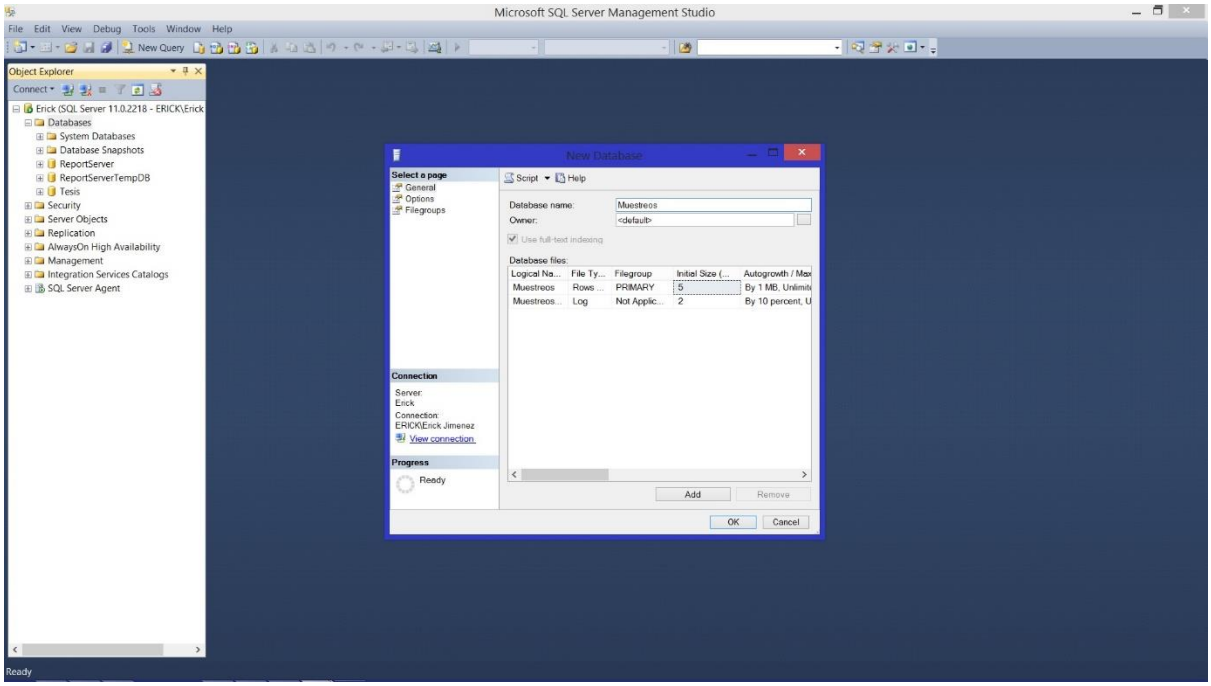


Figura 4.9, Creación de la nueva base de datos

Hecho esto, se eligió en el menú **TASKS** la opción **IMPORT DATA** y se realizó la importación de los datos a Microsoft SQL Server 2012(Figura 12.10).

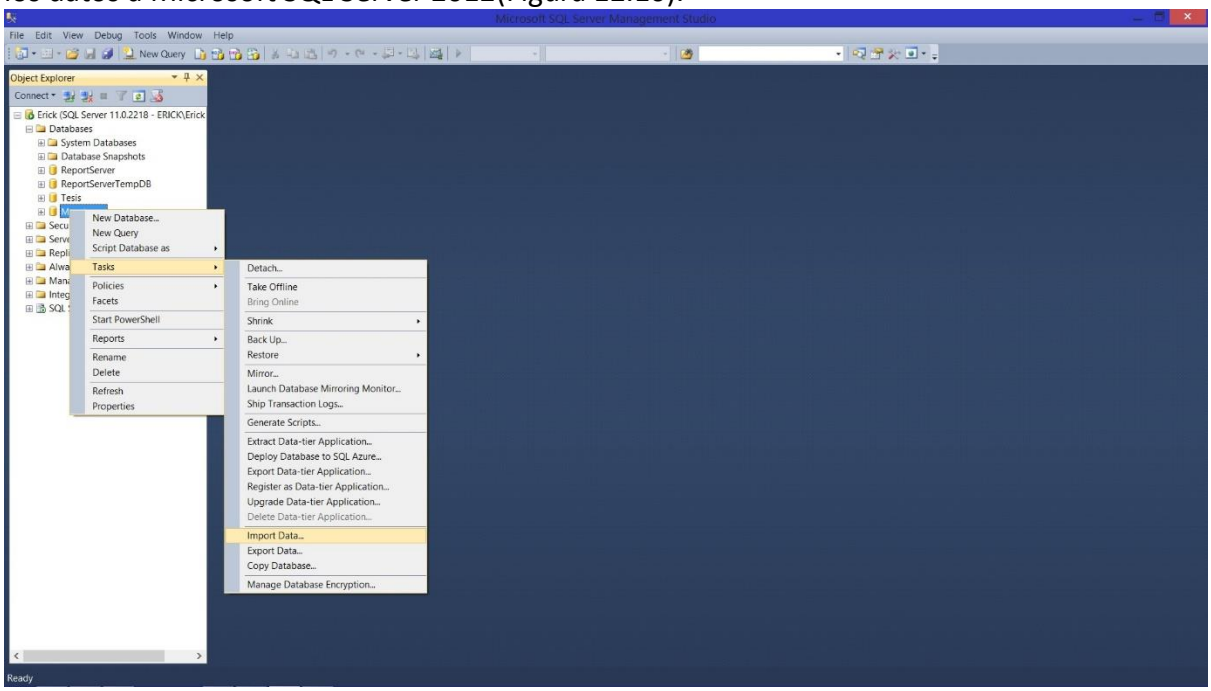


Figura 4.10, Acceso a herramienta de importación de datos

Dentro de la herramienta de importación se eligió nuestra fuente de datos. En este caso son archivos de Microsoft Excel, y se seleccionó la ruta al archivo (Figura 4.11).

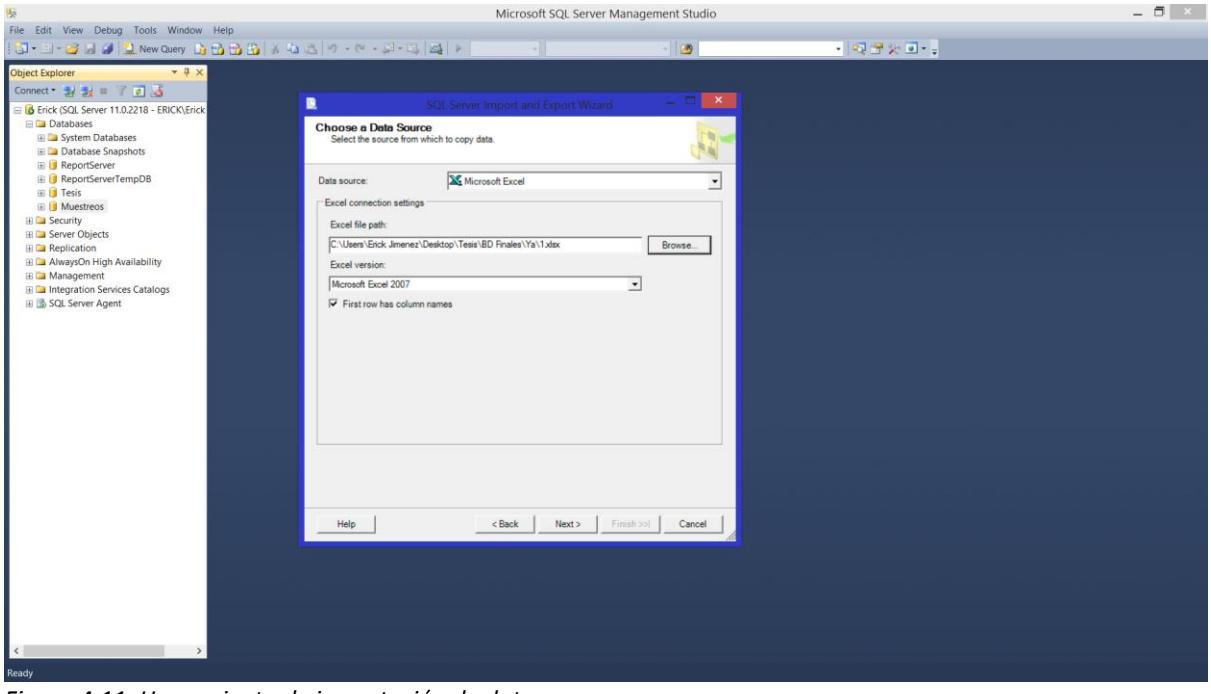


Figura 4.11, Herramienta de importación de datos

Después se eligió la cuenta de usuario, el servidor y la base de datos (a la cual se importó la estructura) (Figura 4.12).

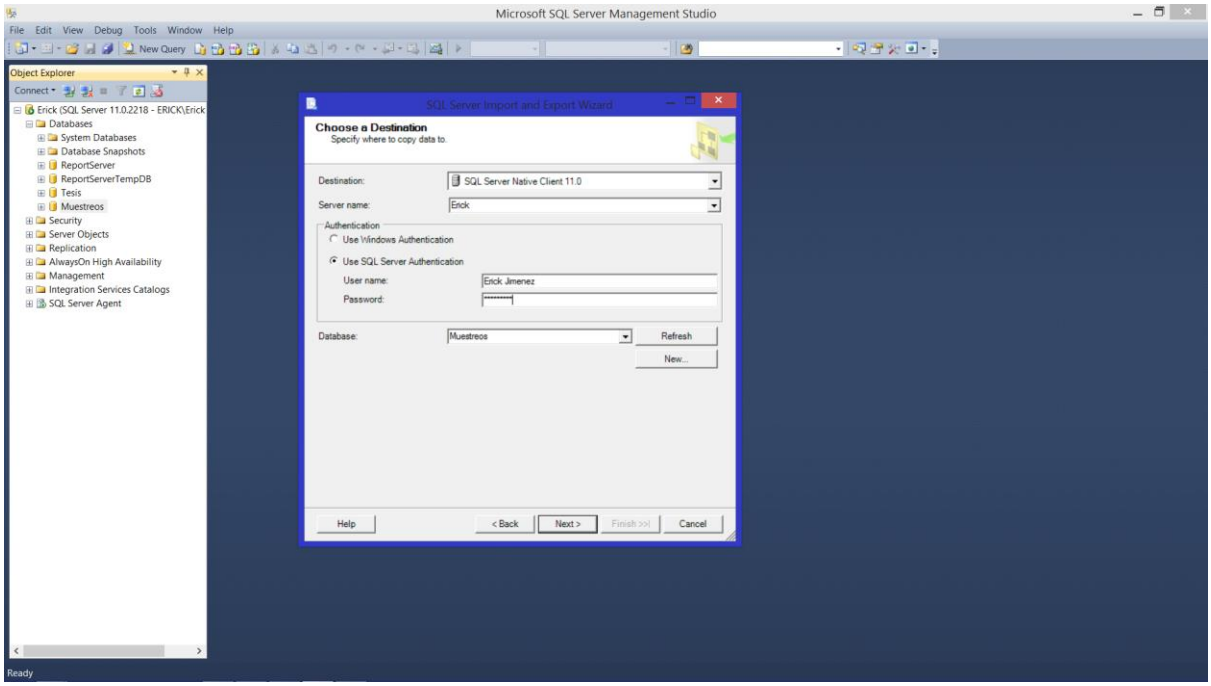


Figura 4.12, Selección de cuenta, BD y servidor

Se especifica que queremos información de una tabla (Figura 4.13).

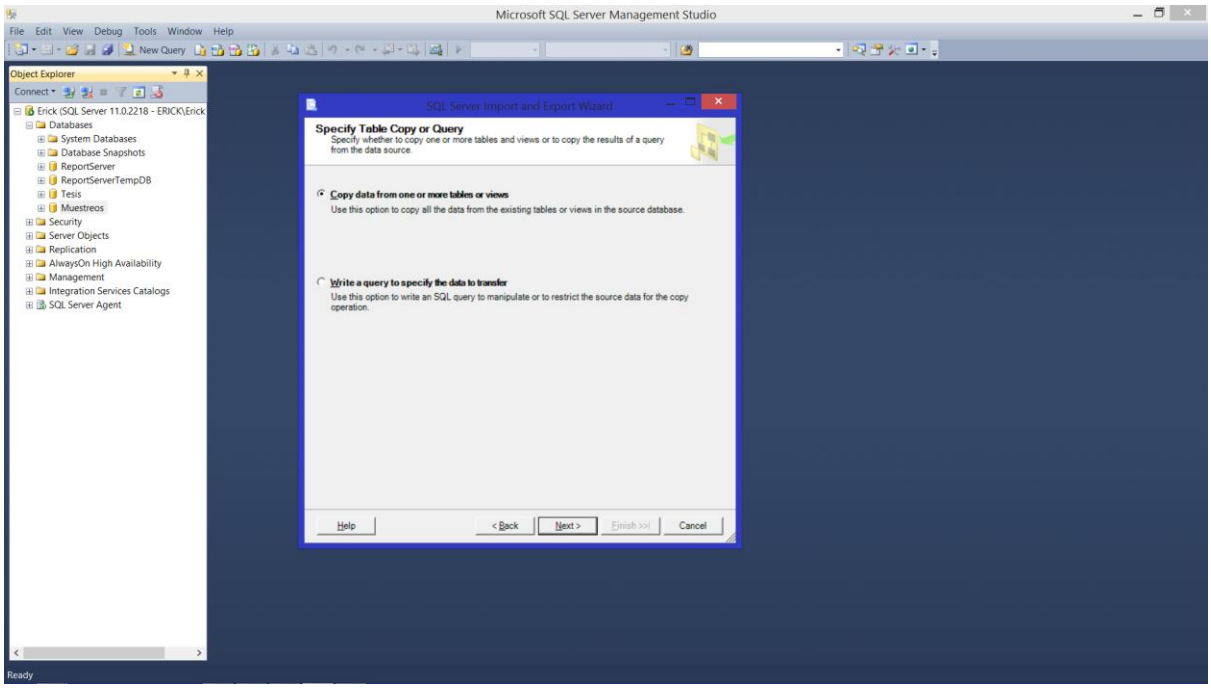


Figura 4.13, Especificación de la fuente de datos

Se elige la hoja de la cual se obtendrán los datos. En este caso se eligió a la única existente, ya que se almacenaron todos los datos en una sola hoja. Al dar **PREVIEW** nos permite realizar una vista previa de la información (Figura 4.14).

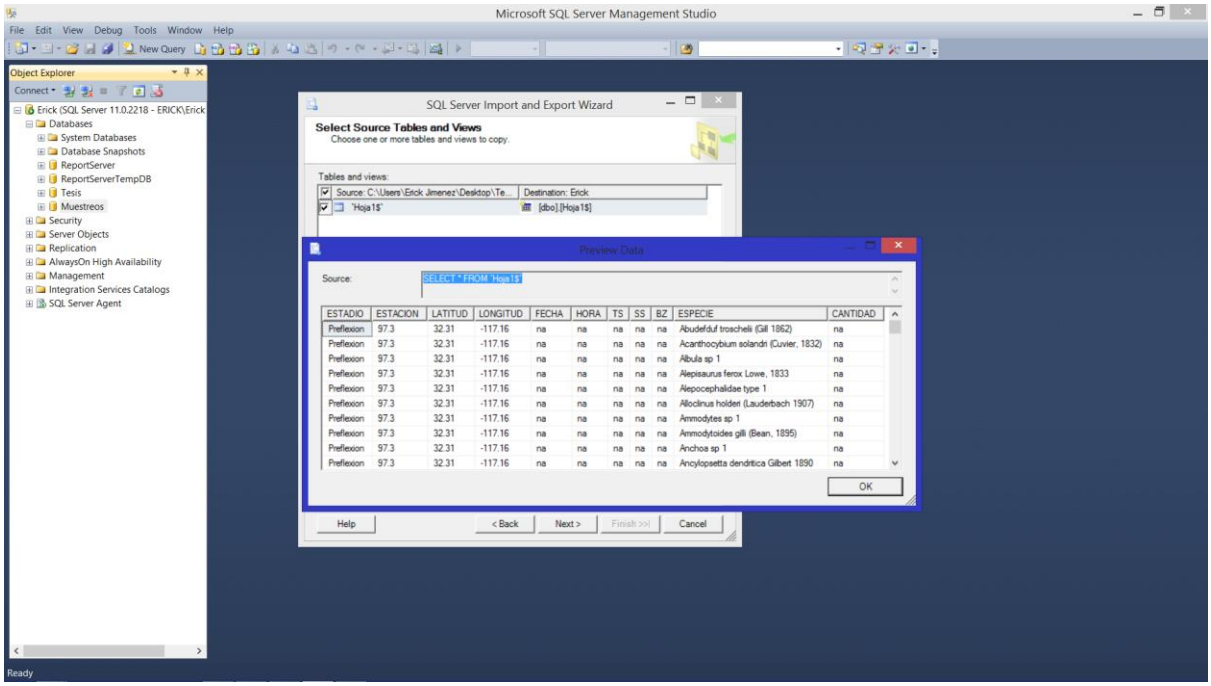


Figura 4.14, Especificación de la fuente de datos

Después se elige la opción de Run (Figura 4.15).

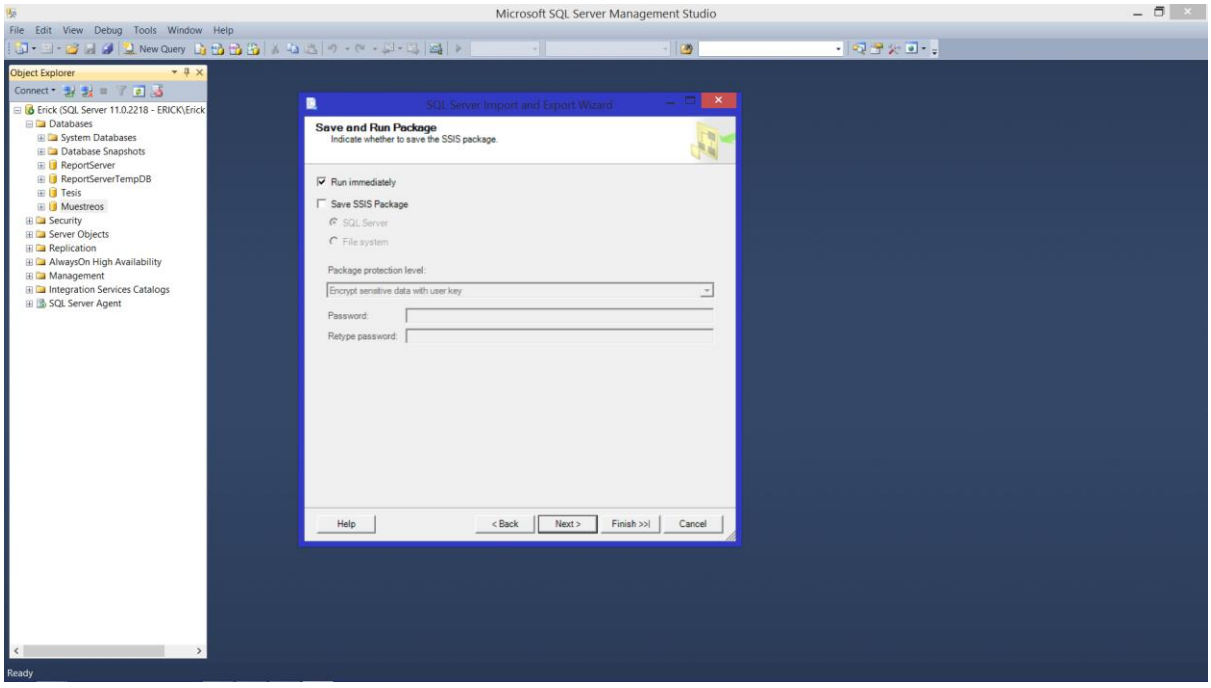


Figura 4.15, Opción de ejecución

Y finalmente se ejecuta la herramienta para que se realice la importación (Figura 4.16).

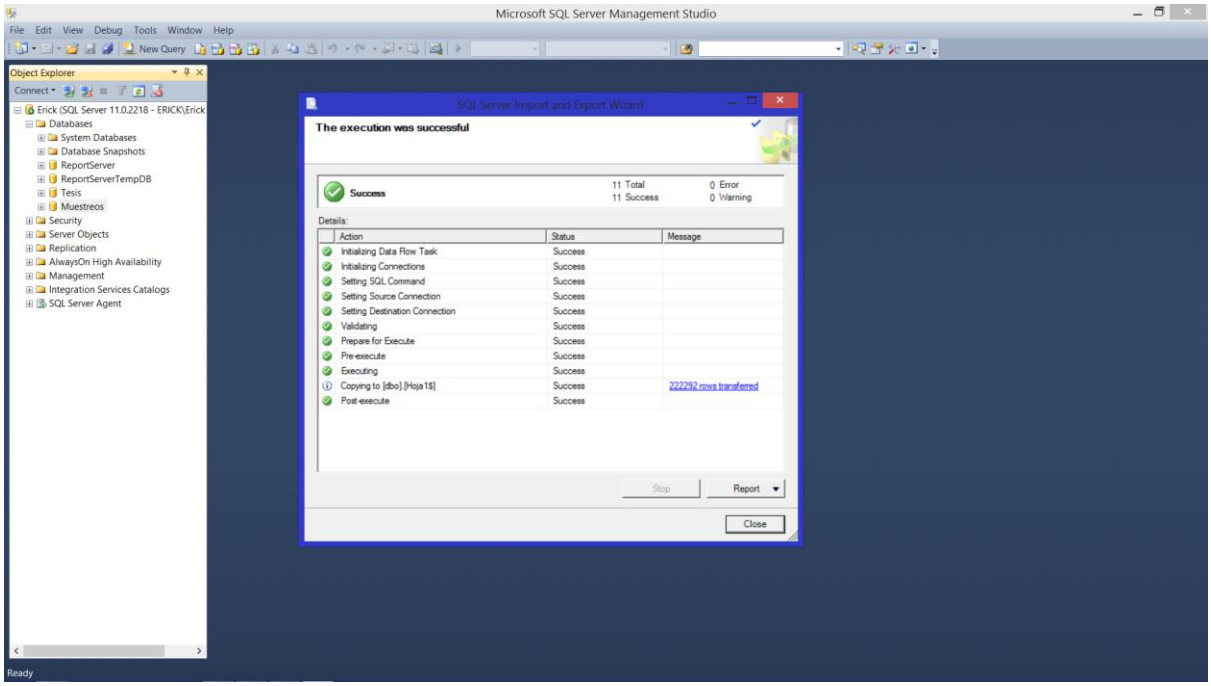


Figura 4.16, Importación exitosa

De esta manera se creó la base de datos que se usó directamente para los modelos de minería de datos (Figura 4.17).

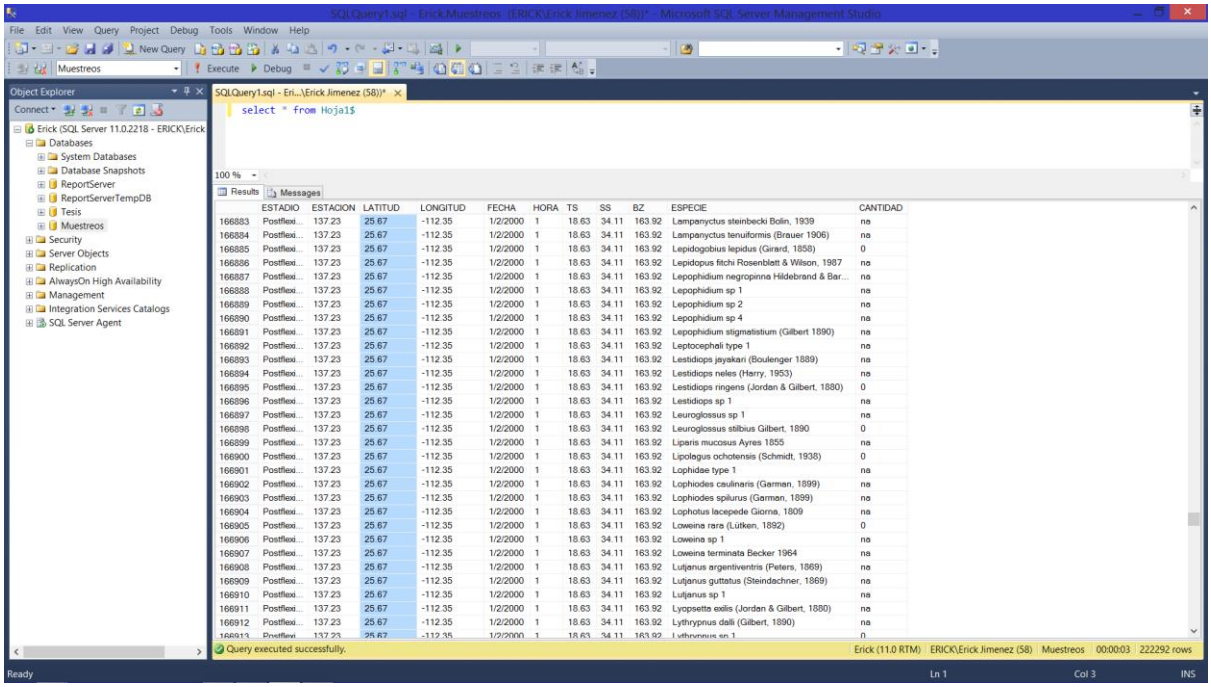


Figura 4.17, Base de datos final

4.4 Creación de un modelo de árboles de decisión.

Una vez obtenida la base de datos necesaria para el modelo de minería, se procedió a la creación del mismo.

Creación de un proyecto de minería de datos

Para la creación de un proyecto de minería de datos, ir a la opción de New Project y se elige Analysis Services Multidimensional and Data Mining Project, se dá clic en OK (Figura 4.18).

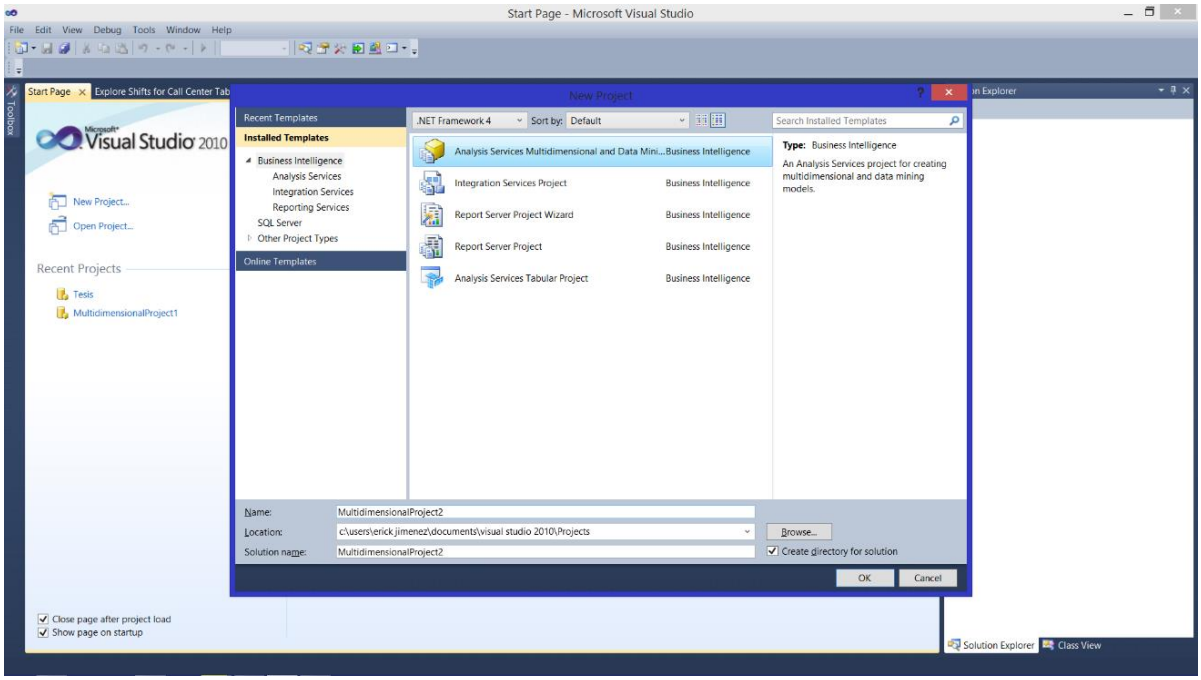


Figura 4.18, Creación de un proyecto de minería de datos.

Selección de la fuente de datos

Es necesario seleccionar nuestra base de datos como fuente de datos para el proyecto de minería. Para realizar esto, ir a la opción Data Sources, se dá clic derecho y se elige la opción de New Data Source. A continuación se dá clic en Next (Figura 4.19).

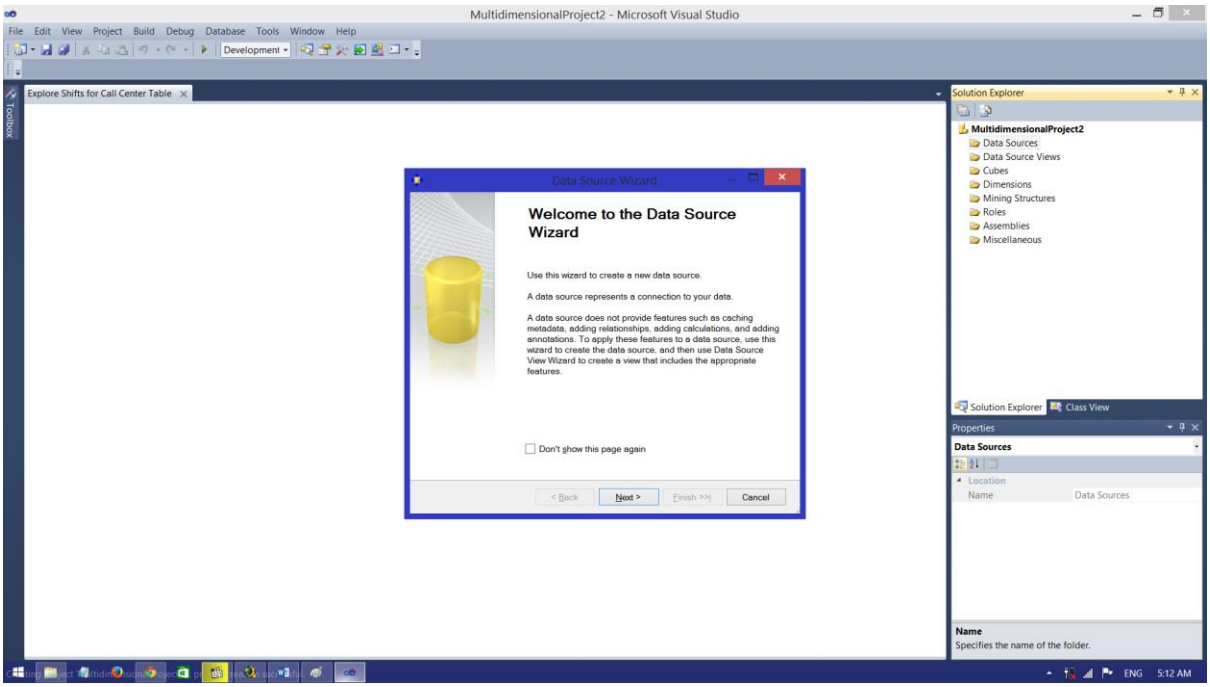


Figura 4.19, Selección de la fuente de datos.

A continuación elegir la conexión que usaremos para acceder a la base de datos. Si ya tenemos una conexión, solo damos clic en ella, de lo contrario, daremos clic en New (Figura 4.20).

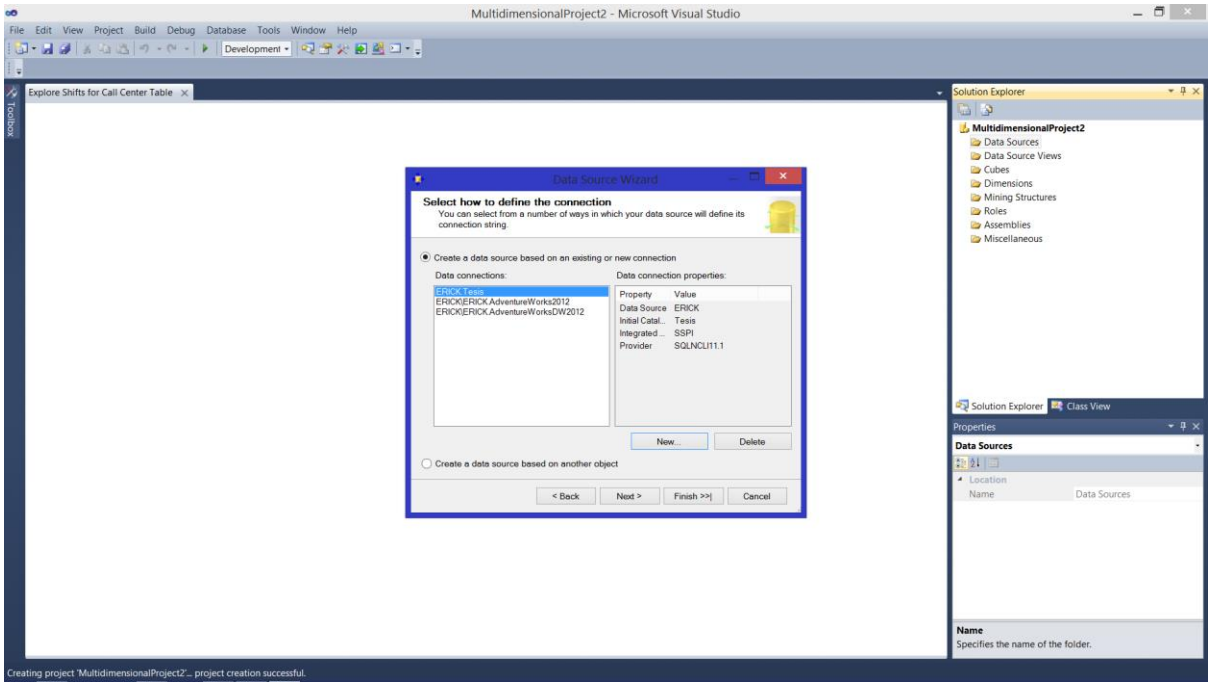


Figura 4.20, Selección de la conexión a la base de datos.

En este menú se selecciona el proveedor de servicio, nuestro servidor, el tipo de autenticación y la base de datos a la queremos conectar y se da clic en OK (Figura 4.21).

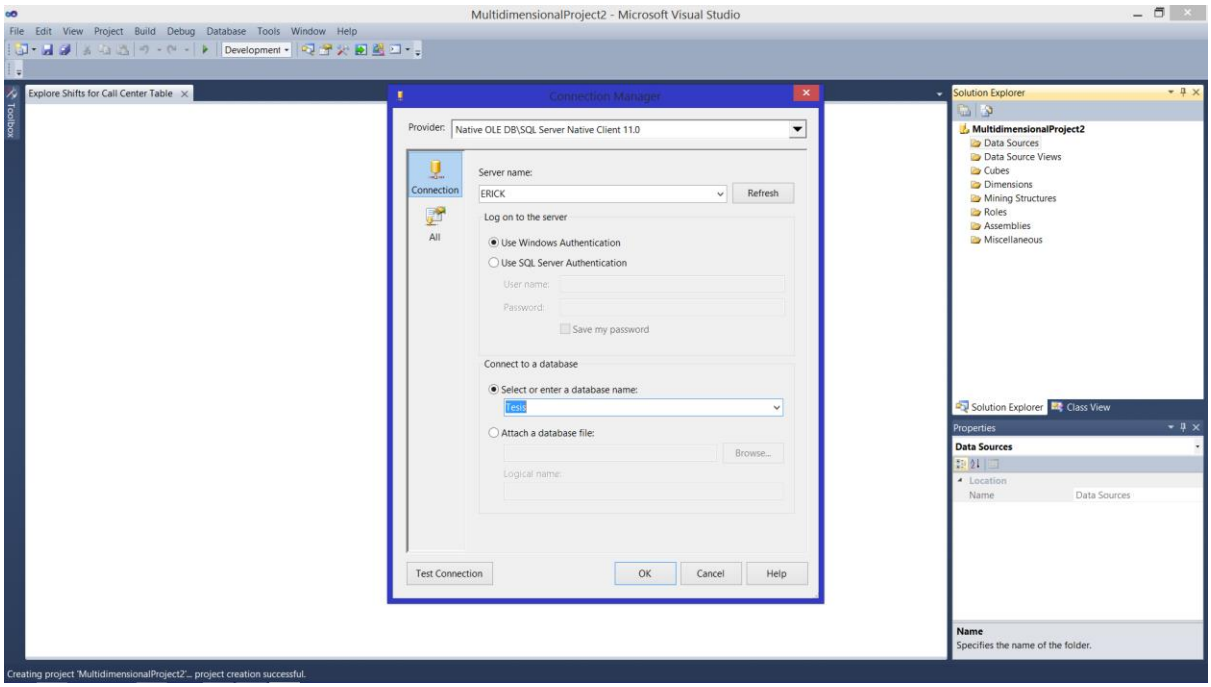


Figura 4.21, Creación de la conexión a la base de datos.

Dar clic en nuestra nueva conexión y después en Finish (Figura 4.22).

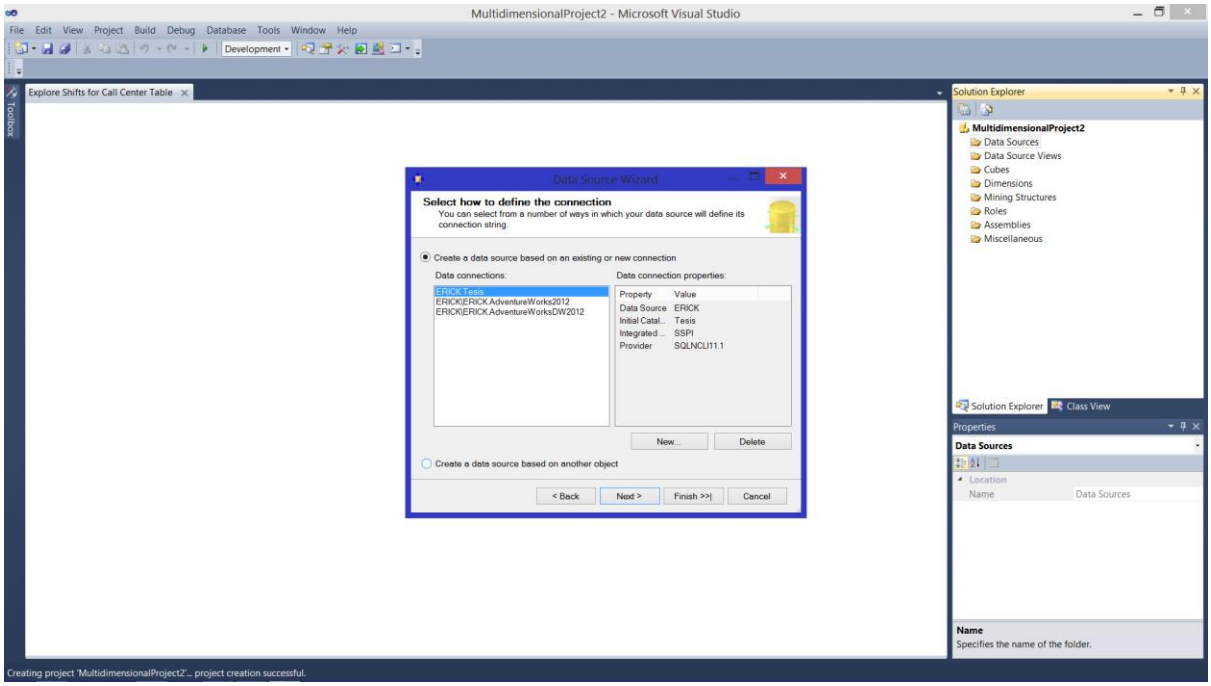


Figura 4.22, Creación de la conexión a la base de datos.

Poner el nombre de nuestra nueva fuente de datos y dar clic en Finish, para tener lista nuestra nueva fuente de datos (Figura 4.23).

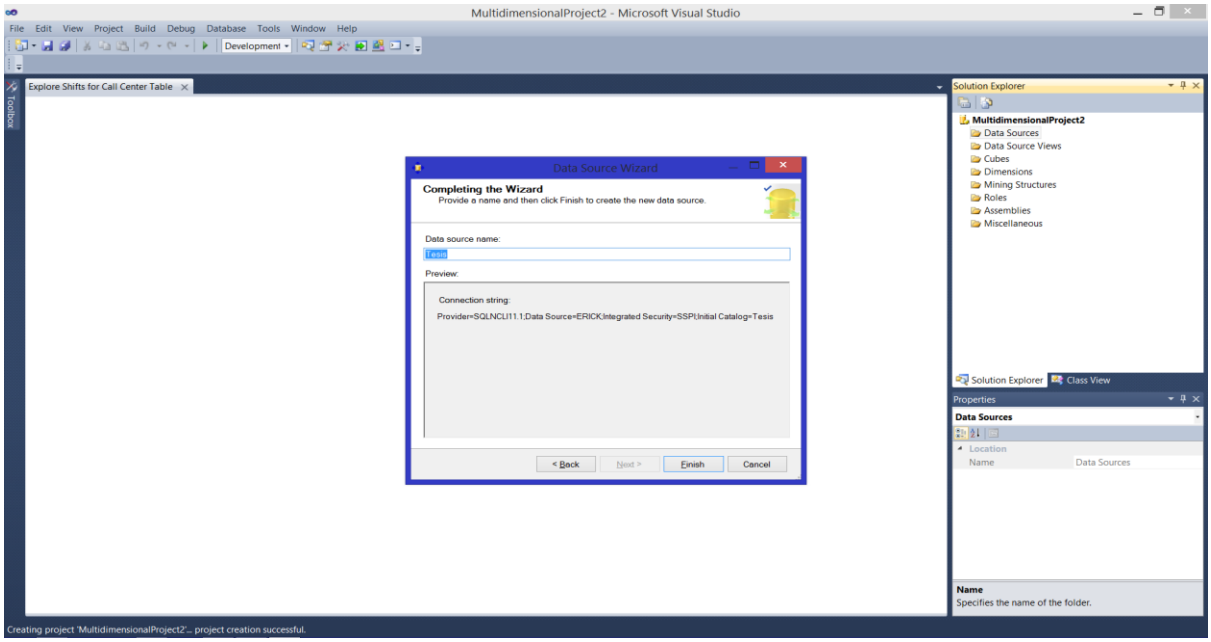


Figura 4.23, Creación de la conexión a la base de datos.

Creación de las vistas de fuentes de datos.

Para la creación de un modelo de minería de datos es necesario crear estructuras llamadas vistas de fuentes de datos que nos permiten observar de manera gráfica la estructura de la base de datos con la que trabajaremos. En este caso se crearon vistas para cada una de las tablas de las especies que se examinaron.

Para crear una nueva vista de fuente de datos, dar clic derecho en Data Source Views y seleccionar la opción New Data Source View, dar clic en Next. (Figura 4.24)

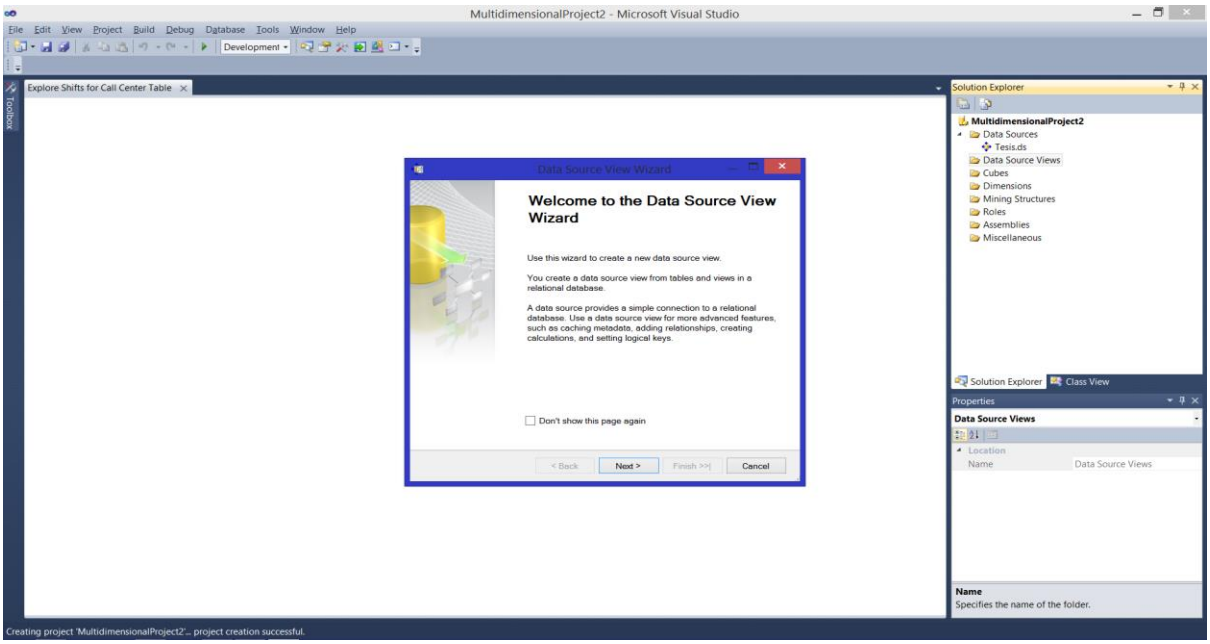


Figura 4.24, Creación de la vista de fuentes de datos.

Elegir la fuente de datos creada previamente, dar clic en next. Se dejan las opciones por default en el menú de Name Matching, ya que en este caso no es de interés especificar llaves externas para la coincidencia de nombres dentro de la base de datos. Dar clic en Next (Figura 4.25).

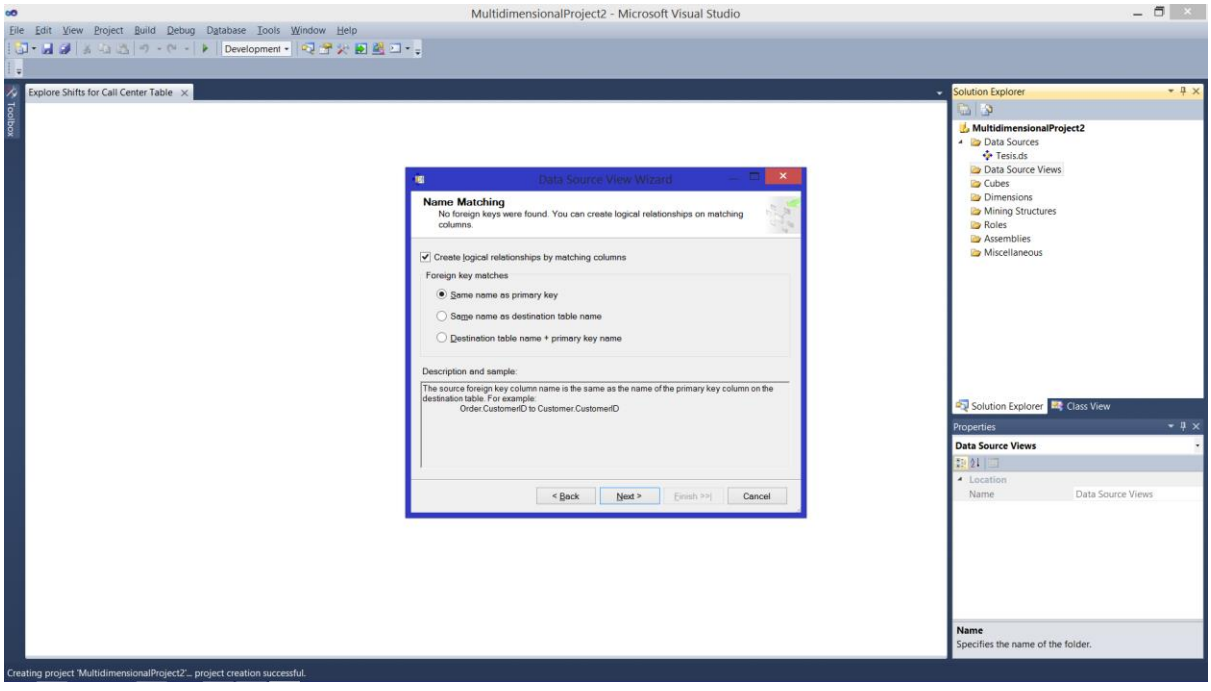


Figura 4.25, Creación de la vista de fuentes de datos.

En el siguiente menú elegir la tabla específica de nuestra base de datos de la cual queremos crear nuestra vista y dar clic en Next (Figura 4.26).

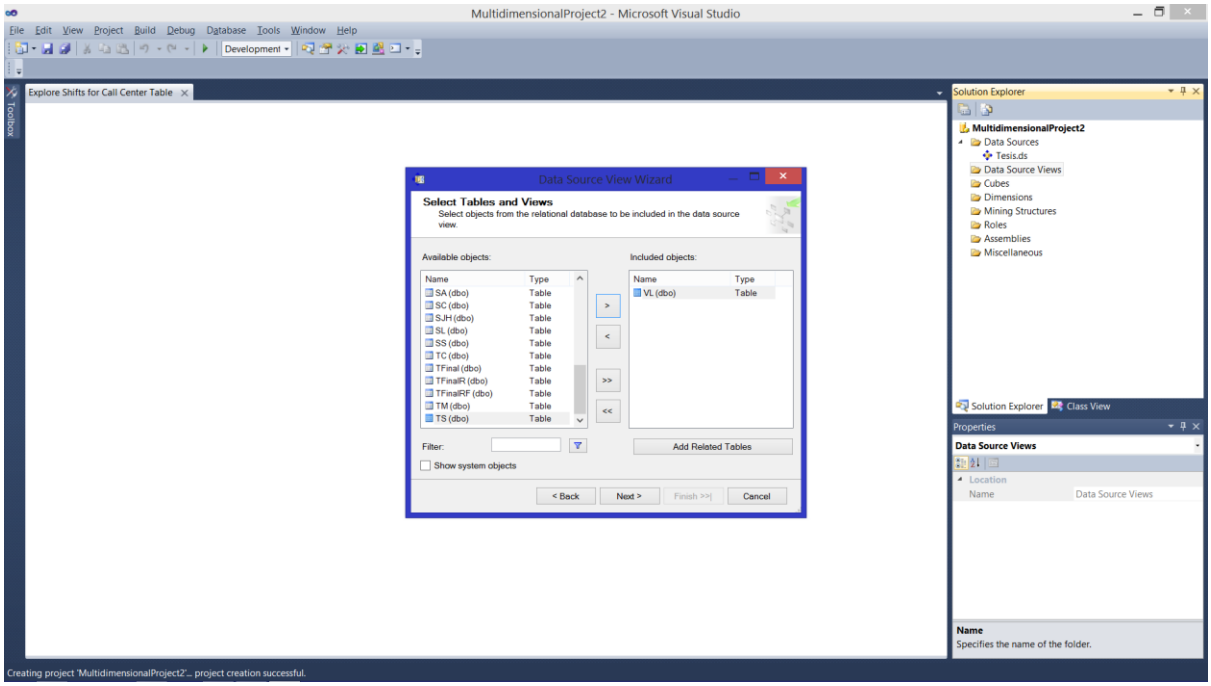


Figura 4.26, Creación de la vista de fuentes de datos.

Una vez elegida, darle un nombre a nuestra vista y dar clic en Finish (Figura 4.27).

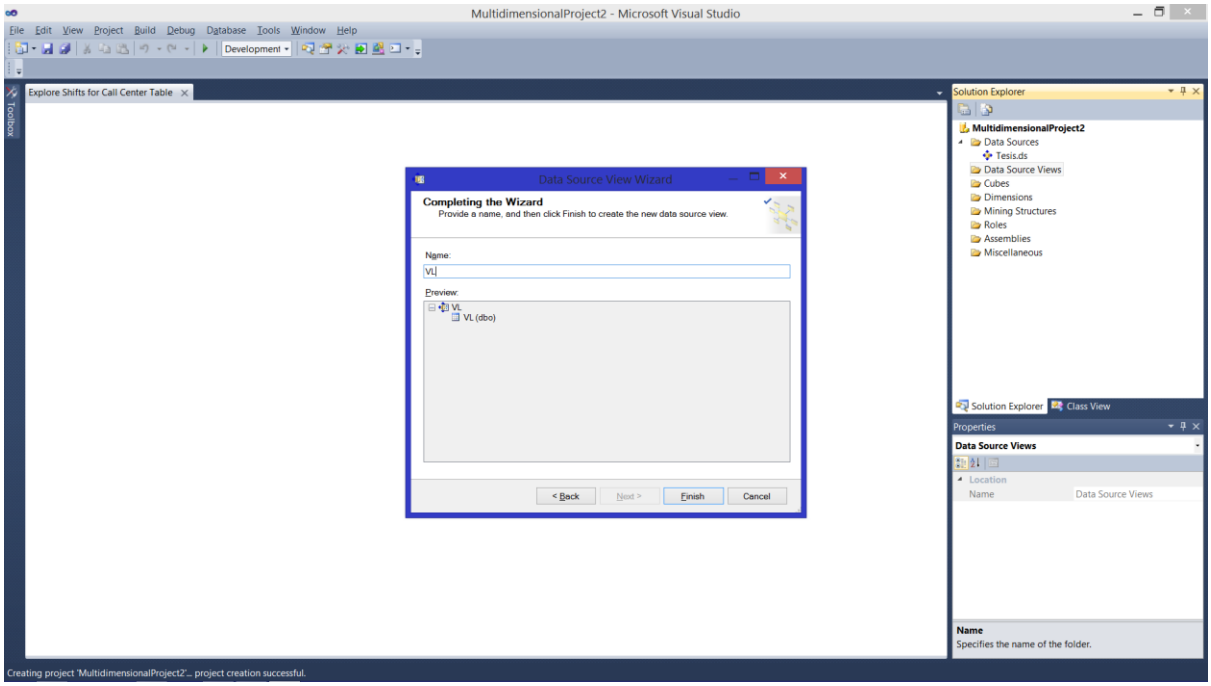


Figura 4.27, Creación de la vista de fuentes de datos.

Una vez hecho, esto podemos observar nuestra tabla de manera gráfica y la podemos usar para crear nuestro modelo (Figura 4.28).

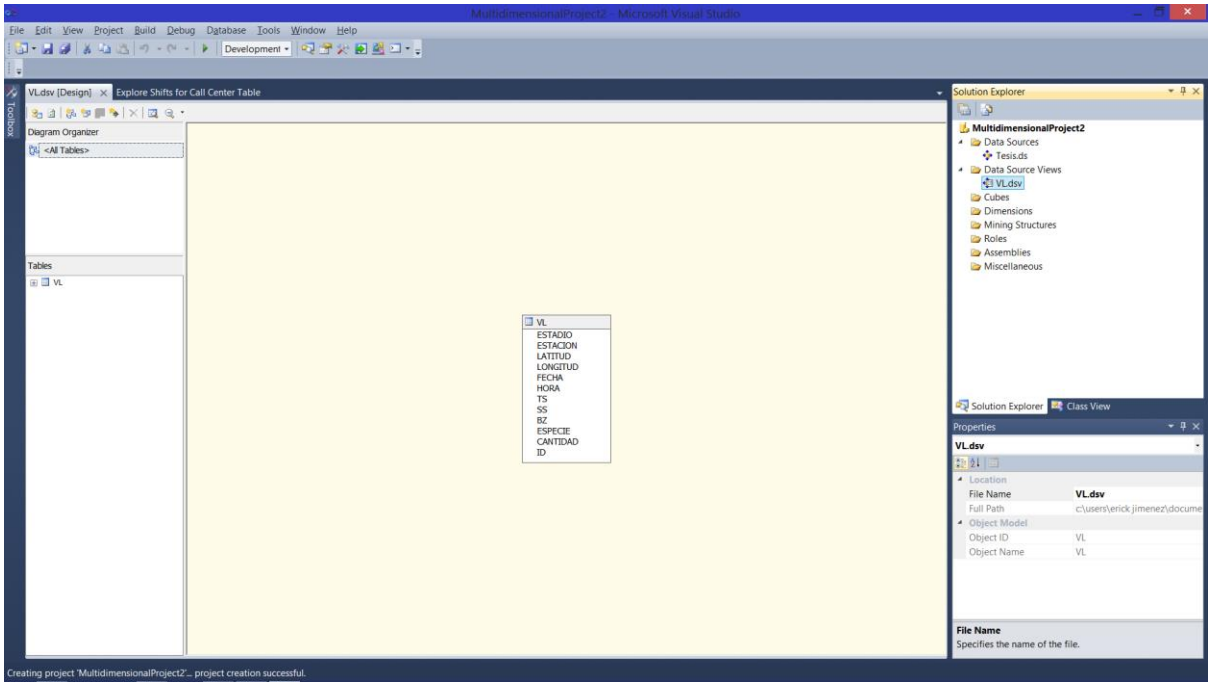


Figura 4.28, Vista de fuente de datos.

Creación de la estructura de minería de datos

Una vez creadas las estructuras necesarias para nuestro modelo, elegir la opción New Mining Structure, dar clic en siguiente y elegir la opción “From a existing relational database or data warehouse”, ya que hemos preparado previamente la base de datos para este proceso, y dar clic en Next (Figura 4.29).

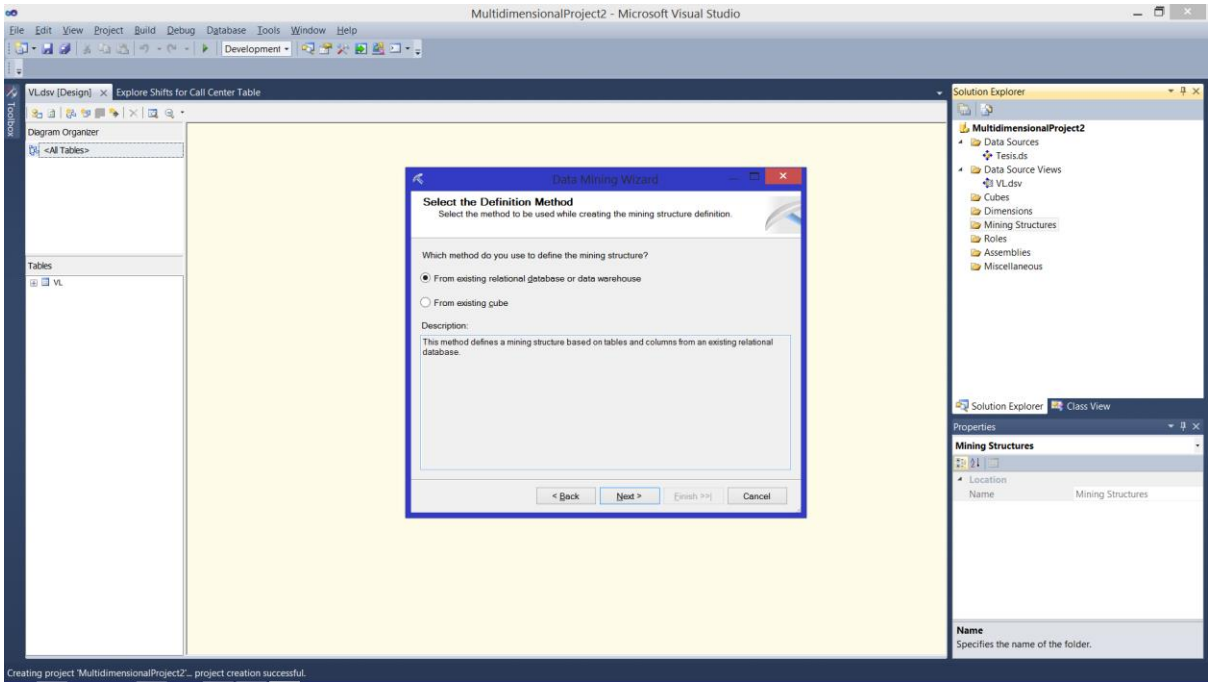


Figura 4.29, Creación de la estructura de minería de datos.

A continuación elegir la técnica de minería que utilizaremos, en este caso Microsoft Decision Trees, y dar clic en Next (Figura 4.30).

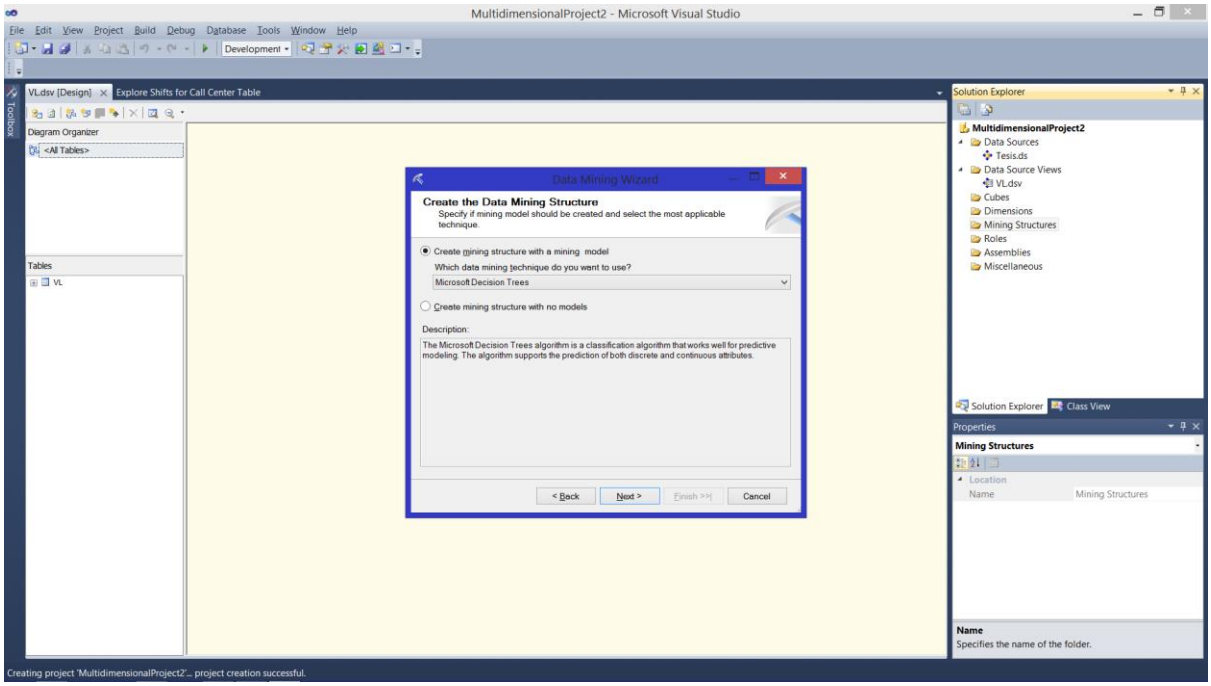


Figura 4.30, Elección de la técnica de minería.

En los siguientes menús elegir la vista que creamos anteriormente, dar clic en Next y después elegir el tipo “Case” para nuestra vista (ya que será nuestra única vista), dar clic en Next (Figura 4.31).

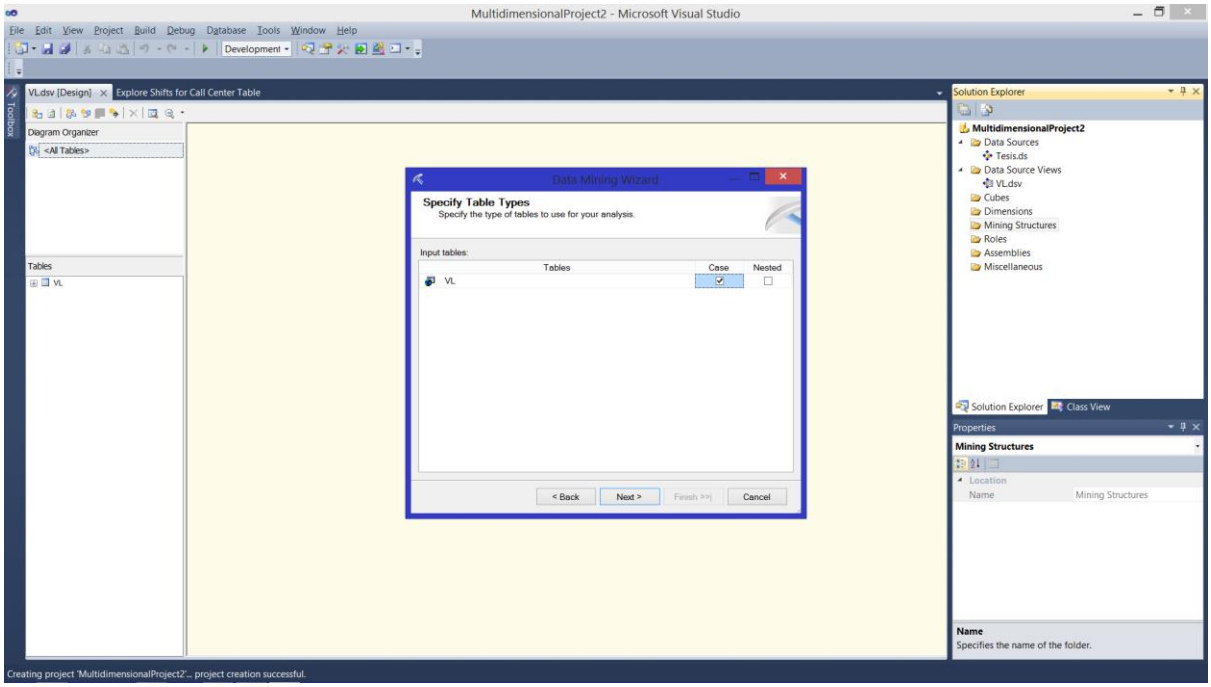


Figura 4.31, Tipo “Case”.

En el siguiente menú se selecciona la estructura de los atributos para el modelo que crearemos, esto es, especificaremos cual atributo será el campo llave, cuales atributos se usarán como entrada de datos y cuales atributos se quieren predecir. Para este modelo se utilizó el atributo ID como llave principal, ya que contiene un identificador único para cada registro. Como entrada se especificaron todos los atributos, ya que todos tienen información importante para realizar

nuestro proceso de predicción y, puesto que nuestro objetivo es conocer las condiciones biológicas necesarias para el desarrollo de las larvas, se especificaron como predictibles los atributos BZ, Cantidad, Especie, Estadio, Hora, SS y TS.

Sin embargo, el modelo nos permite cambiar posteriormente la especificación de los atributos en caso de que deseemos hacerlo. Asimismo, nos permite especificar atributos para que sean tanto de entrada como predictibles.

La especificación de atributos queda como se muestra en la Figura 4.32.

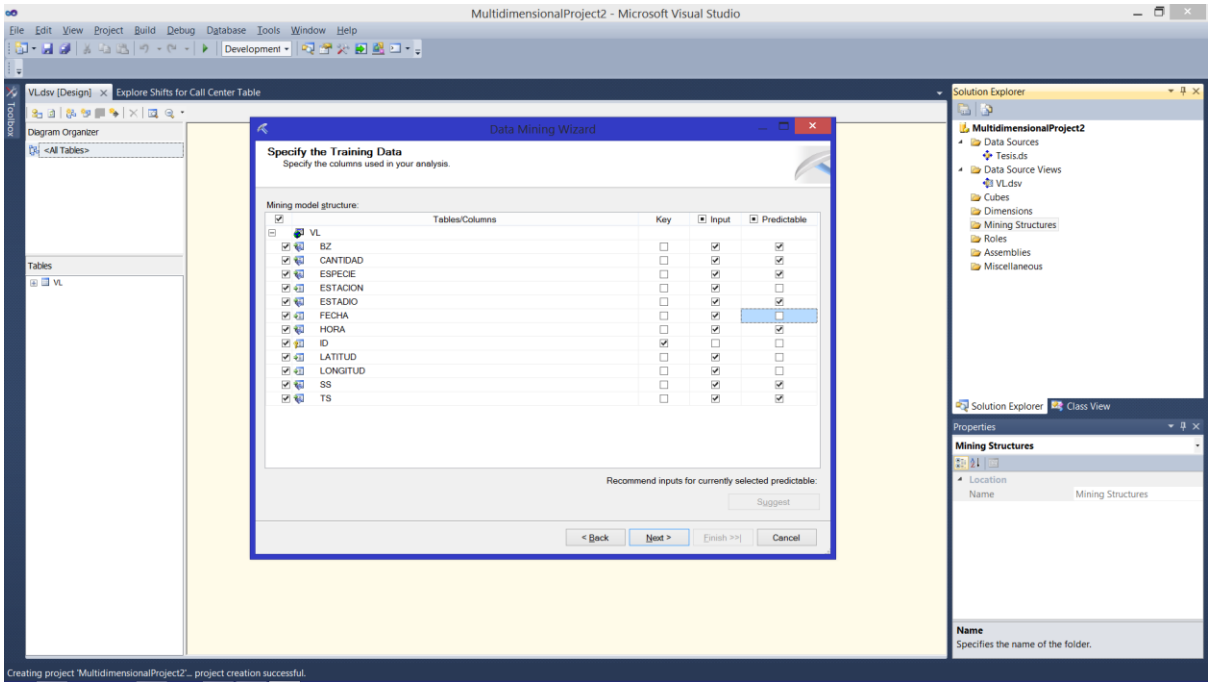


Figura 4.32, Especificación de atributos.

Una vez hecho esto, se seleccionan los tipos de contenido y de datos de cada uno de nuestros atributos. Esto es importante porque aquí definimos si queremos que se trate como un atributo discreto o continuo, y el formato que deseamos que use el modelo. La configuración final se muestra en la Figura 4.33.

Como se ve en la Figura 4.33 los atributos SS, TS, BZ y CANTIDAD, fueron tratados como discretos. Esto es debido a que se realizó un proceso de discretizacion en ellos.

Ya que la finalidad de nuestro modelo es la de obtener información sobre las condiciones biológicas que rodean a la incidencia o ausencia de una especie, no fue de interés la especificación de abundancias sobre cada especie, por lo que en este atributo se le dio el valor 0 a todos aquellos registros en los que no se hubiera encontrado la especie, y se le dio el valor 1 a cualquier cantidad mayor que 0, esto con el fin de especificar únicamente ausencia o incidencia.

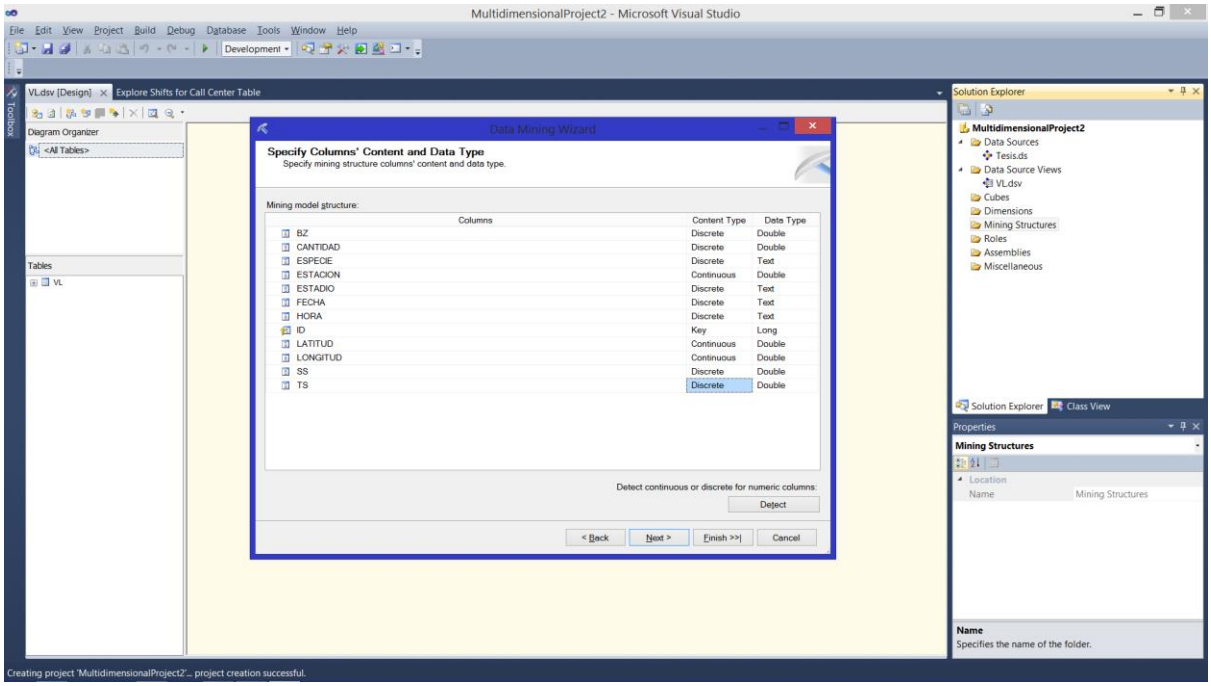


Figura 4.33, Especificación tipos de datos.

En el caso de los atributos TS, SS y BZ, se notó que no sería de mucha utilidad obtener predicciones sobre valores específicos de cada atributo, ya que los comportamientos biológicos de las larvas no son lineales. Un resultado demasiado específico no necesariamente indica un comportamiento. Por lo tanto, se crearon rangos de manera que se tuvieron valores bajos, medios y altos para cada uno de estos atributos. La especificación se muestra en la Figura 12.34

Atributo BZ		
Rango de valores	Cantidad	Valor en la tabla
1-100.4	Bajo	1
100.5-350.4	Medio Bajo	2
350.5-700.4	Medio Alto	3
Mayor que 700.5	Alto	4
Atributo SS		
Rango de valores	Cantidad	Valor en la tabla
32.5-33.5	Bajo	1
33.55-34.5	Medio	2
34.55-35.5	Alto	3
Atributo TS		
Rango de valores	Cantidad	Valor en la tabla
12.1-17.1	Bajo	1
17.15-22.1	Medio	2
22.15-27	Alto	3

Figura 4.34, Discretizacion de atributos BZ, SS y TS.

Una vez que hemos definido los atributos, se pasa al siguiente menú y se selecciona el porcentaje de la información que usaremos para entrenar a nuestro modelo. Al hacerse varias pruebas con varios porcentajes se determinó que los mejores resultados se obtienen al utilizar el 50% de los datos. Esto se debe a que, dentro de los datos el porcentaje de incidencias de cada

especie es menor que el de las ausencias. Por lo tanto, se debió utilizar un porcentaje mayor al 30% por default de los datos para realizar el entrenamiento, con el fin de que se tomen en cuenta la mayor cantidad de casos de incidencia posibles.

El número de casos se deja en blanco, ya que no es de interés especificar una cantidad de casos de prueba (Figura 4.35).

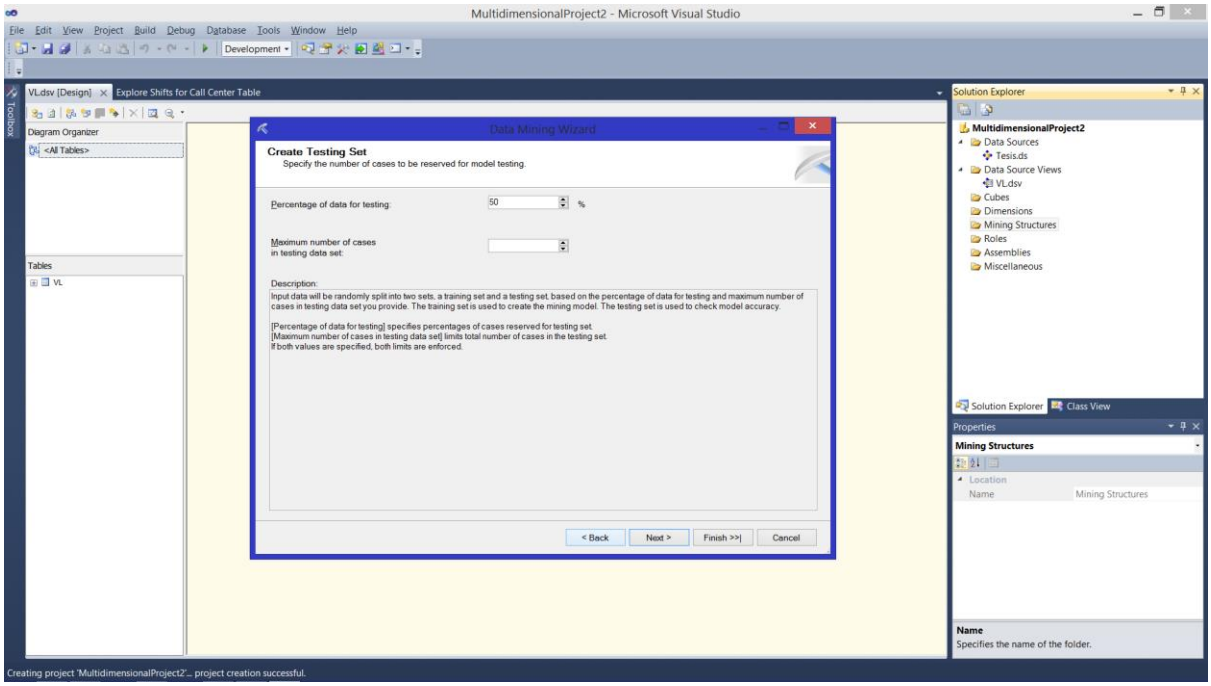


Figura 4.35, Especificación del porcentaje de datos para el entrenamiento.

Hecho esto dar clic en Next y se especifica el nombre de nuestra estructura de minería. Para terminar dar clic en Finish y podremos ver que nuestra estructura ha sido creada (Figura 4.36).

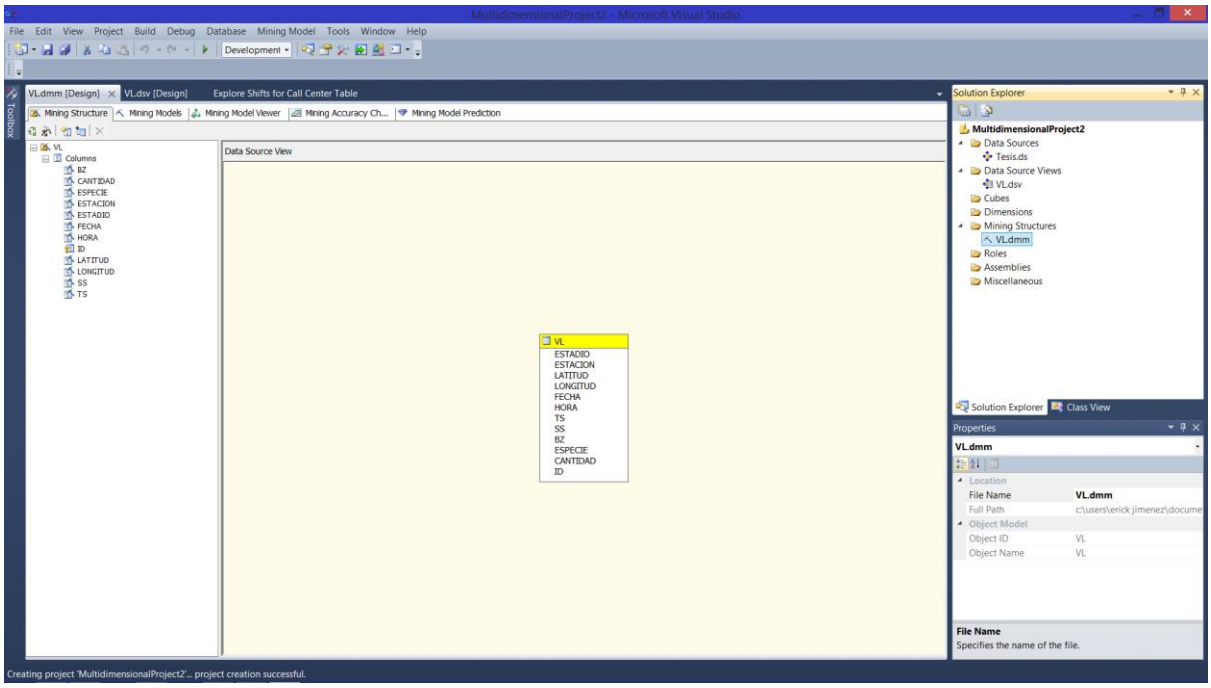


Figura 4.36, Estructura final.

Ya que hemos creado nuestra estructura, podemos empezar el proceso de minado, pero debemos procesarla primero. Para esto dar clic derecho en nuestra estructura y dar clic en Process. Nos saldrá una advertencia de contenido. Dar clic en Yes para llegar al siguiente menú (Figura 4.37).

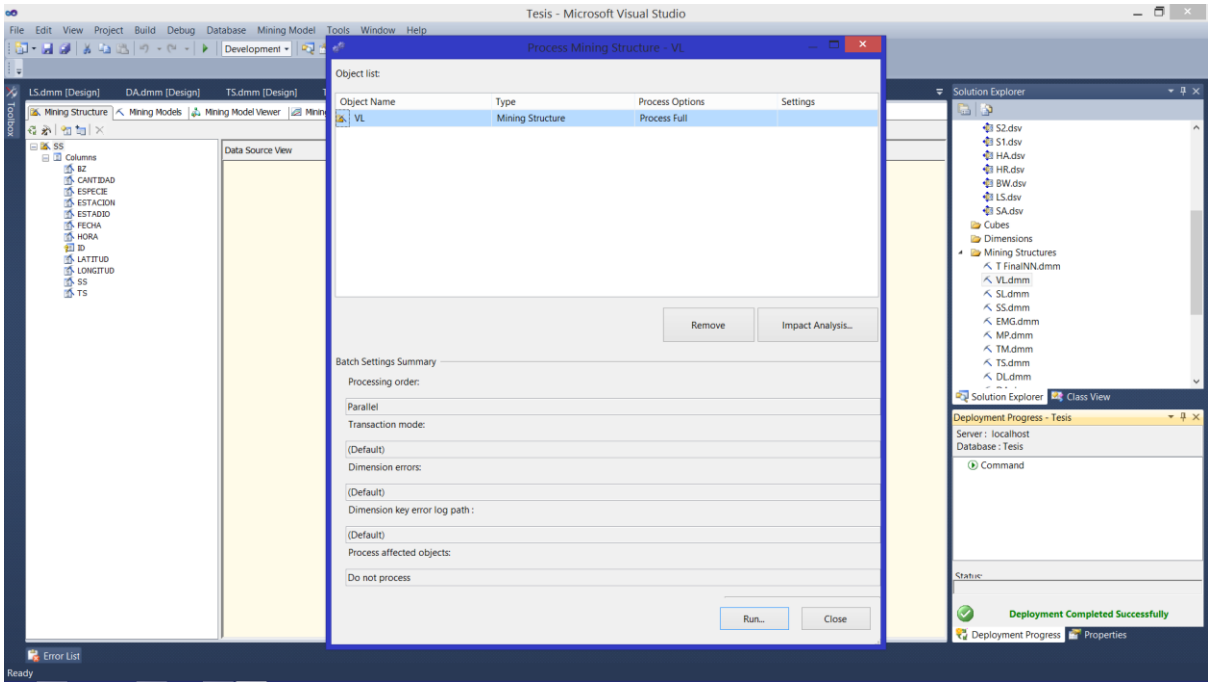


Figura 4.37, Procesamiento de la estructura.

En este menú dar clic en Run para iniciar el proceso de procesamiento y una vez terminado dar clic en Close (Figura 4.38).

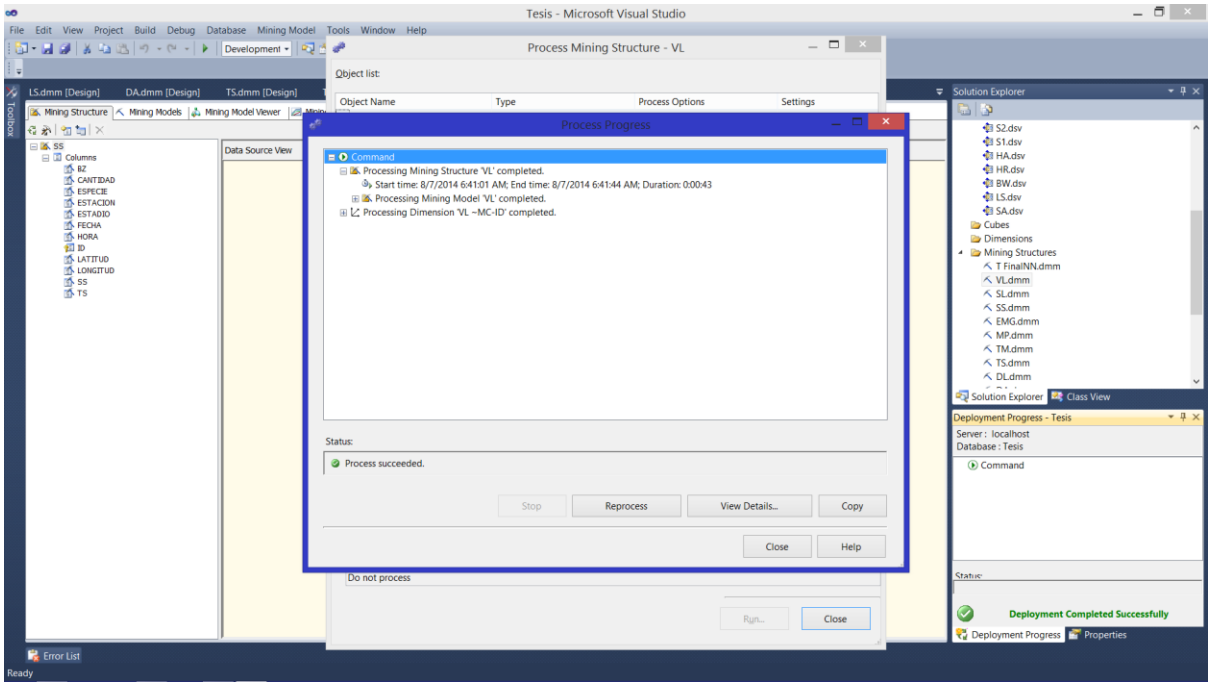


Figura 4.38, Procesamiento finalizado.

Visor del modelo

Una vez que finalizado el procesamiento del modelo, podemos observar los resultados del análisis a través de la herramienta de visión del modelo. Para acceder a esta herramienta debemos ir a la pestaña Mining Model Viewer (Figura 4.39).

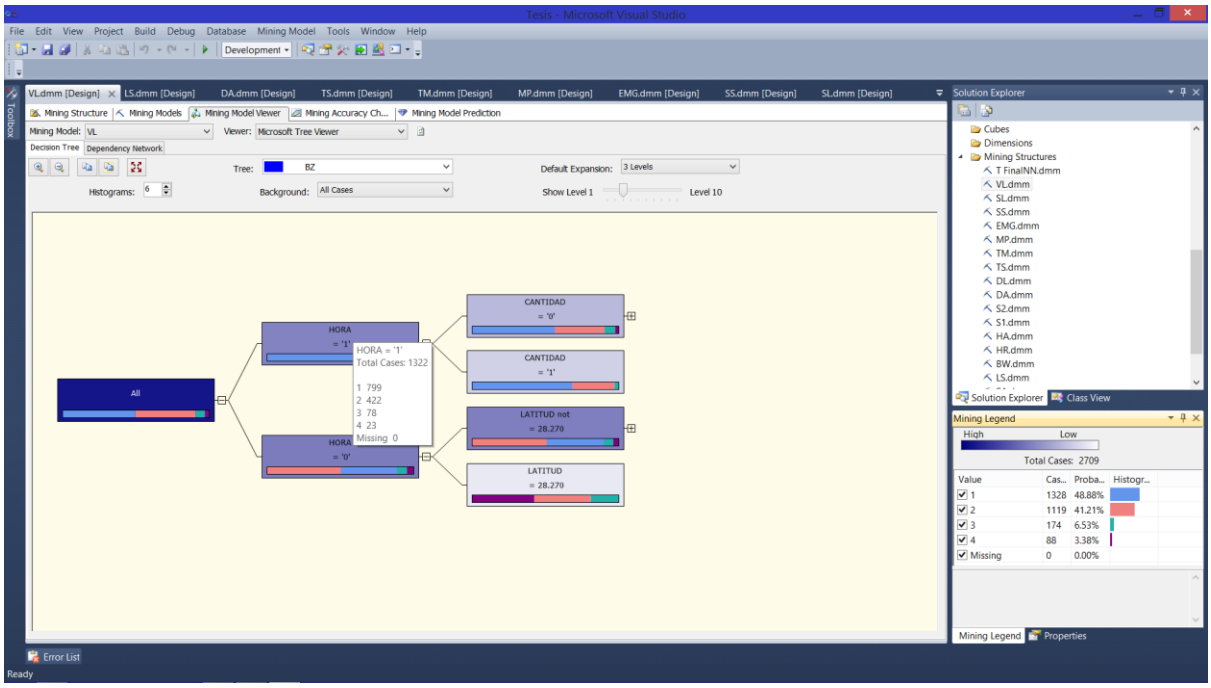


Figura 4.39, Visor del modelo.

El modelo nos muestra el resultado del minado en forma de un árbol, donde se despliegan los atributos más importantes para la distribución de una condición en particular, asimismo nos indica con color de fondo distinto la abundancia de registros generados por el modelo. Para conocer los atributos o condiciones que afectan la incidencia o ausencia de una especie, (el atributo CANTIDAD), elegiremos el árbol de este atributo desde la opción Tree. Nos interesa que muestre aquellos registros en los que el atributo CANTIDAD es igual a 1, indicando incidencia, así que elegiremos ese valor desde la opción Background. Esto nos cambiará el fondo del árbol de manera que podamos observar gráficamente cómo se distribuye la especie en un árbol de distribución (Figura 4.40).

El visor también tiene opciones para determinar la cantidad de niveles que se mostrarán. El número de niveles varía en función del modelo y de la cantidad de atributos discriminantes que se encuentren para un determinado resultado (En este caso el árbol tiene 7 niveles) Se puede cambiar el número de niveles mostrados desde la herramienta Show Level. También podemos configurar cuántos niveles se mostrarán por default en la herramienta Default Expansion. En nuestro caso se seleccionó All Levels, de manera que siempre se muestren todos los atributos que discriminan la incidencia de la especie (Figura 4.41).

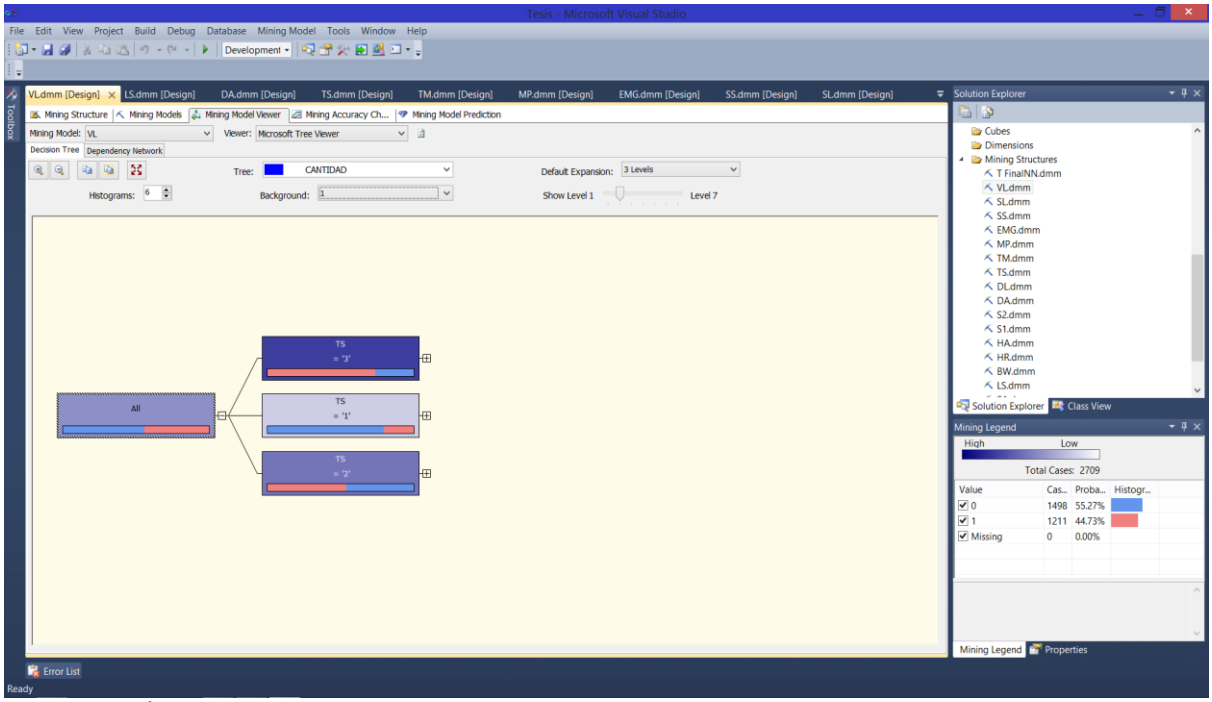


Figura 4.40, Árbol de distribución.

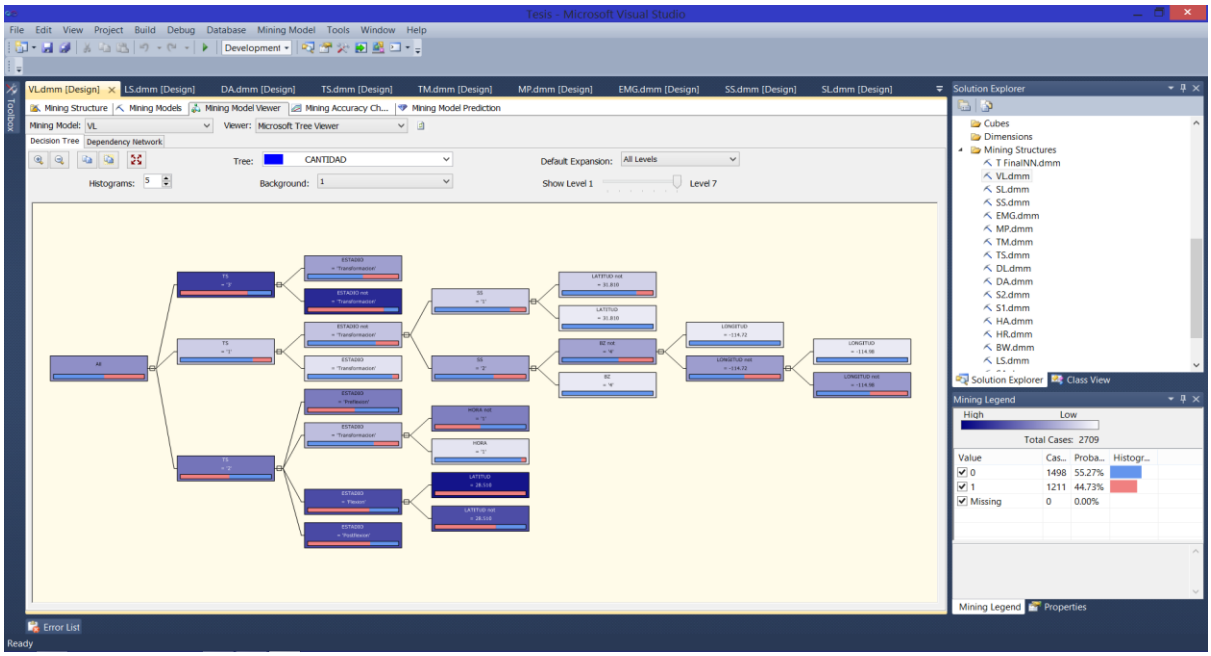


Figura 4.41, Árbol expandido.

Cada hoja de el árbol de distribución contiene información detallada acerca de la cantidad de casos que se generaron en base al modelo y la probabilidad de que los valores de el atributo que elegimos se presenten, esto lo podemos observar desde el cuadro Mining Legend en la sección inferior derecha (Figura 4.42).

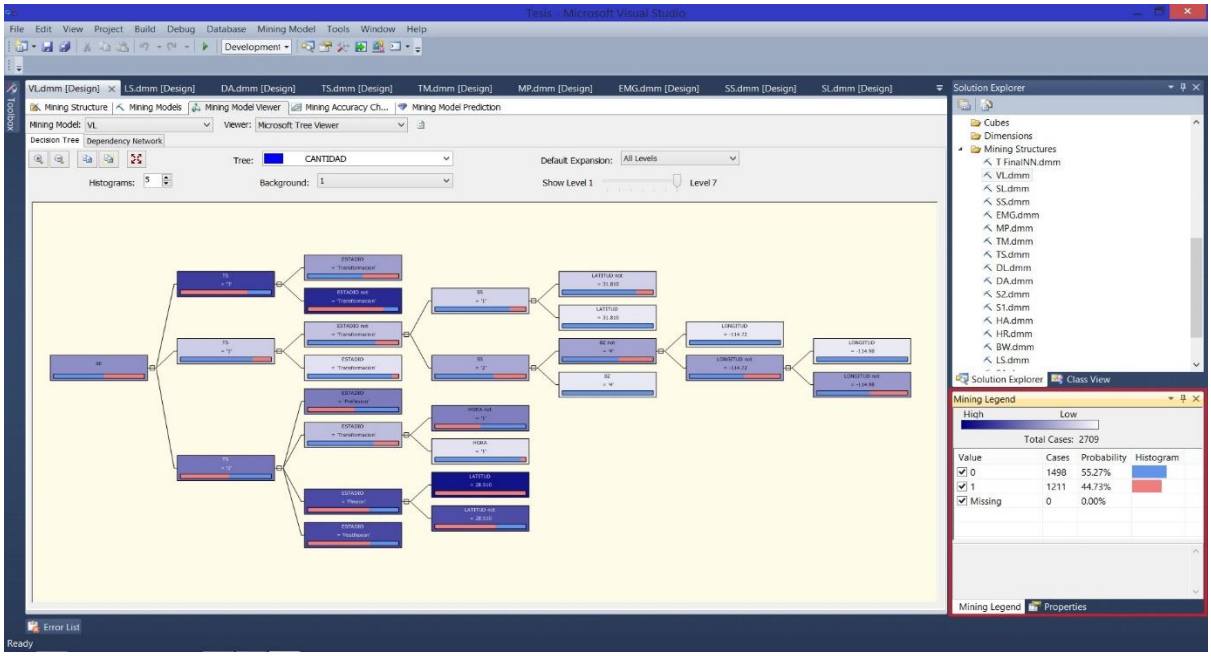


Figura 4.42, Mining Legend.

Aquí podemos observar el total de registros que se generaron en base a la construcción del modelo, divididos entre los posibles valores del atributo, cada uno con su probabilidad asociada y con el histograma que se muestra dentro de la hoja del árbol, asimismo se muestra en un degradado de color azul, el color de fondo de cada hoja y que es lo que indica, como vemos mientras más azul más alta la probabilidad (Figura 4.43).

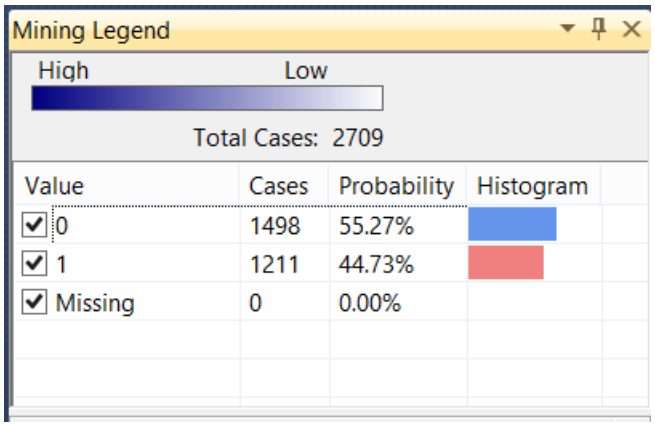


Figura 4.43, Mining Legend a detalle.

Una vez realizado lo anterior, podremos analizar de manera gráfica los resultados del procesamiento de minado del modelo que creamos para esta especie.

Capítulo 5
Resultados y Conclusiones.

5.1 Resultados.

Del listado de atributos integrados en las bases de datos de la Colección ICTIOPLANCTON se realizó una segregación de aquellos que se referían de forma directa a las variables ambientales que se han identificado como significativas en la distribución de las larvas de la mayoría de las especies de peces registradas para la zona oceánica y costera de la Península de Baja California. La bibliografía indica que los atributos de temperatura y salinidad en la columna de agua son importantes para el correcto desarrollo de los huevos y larvas de peces, y a su vez, están asociados a condiciones de abundancia o escasez de alimento y/o densidad de predadores y/o competidores de las larvas de peces. Ambos factores que tienen influencia en la sobrevivencia larval, expresada en su abundancia.
De igual manera, estos atributos difieren de acuerdo a la ubicación geográfica de las estaciones muestreadas (latitud) y a su relación con las zonas costera y oceánica (longitud).

Una vez hecha la discriminación de atributos se realizó un proceso, ETL que nos permitió limpiar y cambiar la estructura de la información para transformar las hojas de cálculo en una base de datos para SQL Server 2012 el diagrama se muestra en la Figura 5.1.

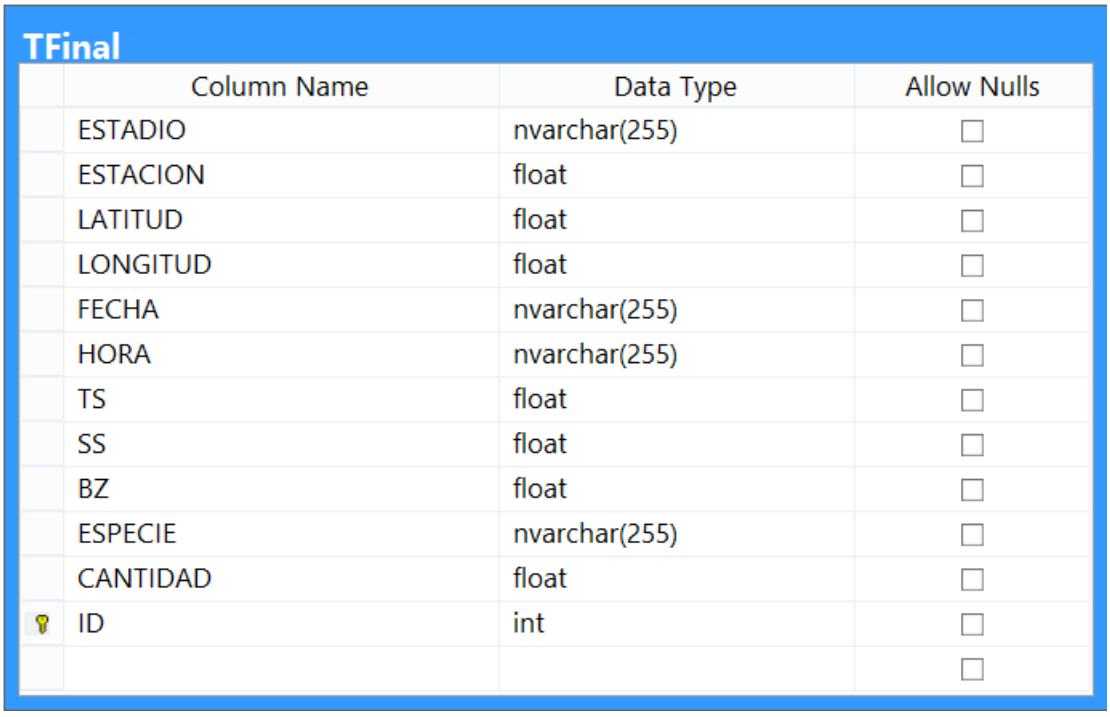


Figura 5.1, Diagrama E-R de la base de datos.

Posteriormente se creó un modelo de minería de datos, utilizando el módulo Analysis Services de SQL Server 2012 con el algoritmo de árboles de decisión de Microsoft.

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporcionado por Analysis Services para el modelado de predicción de atributos discretos y continuos como los que se tienen en la colección ICTIOPLANCTON, razón por la que fue el algoritmo elegido para este trabajo. Los atributos continuos y discretos que se especificaron se muestran en la Figura 5.2.

Columns	Content Type	Data Type
BZ	Discrete	Double
CANTIDAD	Discrete	Double
ESPECIE	Discrete	Text
ESTACION	Discrete	Double
ESTADIO	Discrete	Text
FECHA	Continuous	Text
HORA	Discrete	Text
ID	Key	Long
LATITUD	Continuous	Double
LONGITUD	Continuous	Double
SS	Discrete	Double
TS	Discrete	Double

Figura 5.2, Atributos continuos y discretos.

Y la configuración de atributos de entrada y de predicción se muestra en la Figura 5.3.

BZ	Predict
CANTIDAD	Predict
ESPECIE	Predict
ESTACION	Predict
ESTADIO	Predict
FECHA	Input
HORA	Predict
ID	Key
LATITUD	Input
LONGITUD	Input
SS	Predict
TS	Predict

Figura 5.3, Atributos de entrada y predicción.

Los atributos de entrada y predicción fueron configurados de esta manera debido a un proceso de discretización mostrado en la sección de desarrollo.
Otra razón del uso de este algoritmo es la presentación gráfica de los resultados a través de árboles de decisión, que permiten hacer un análisis rápido desde el primer acercamiento. Esta capacidad nos permitió dar resultados de 2 formas distintas, la primera es un árbol de distribución en el cual se muestran en forma de árbol, los atributos que afectan la aparición de un valor en un atributo dado y el número de registros generados por el modelo en una escala de colores lo cual permite un análisis más rápido para el experto. La otra forma de mostrar resultados fue la creación de árboles de probabilidades, que nos muestran más a detalle cada una de las probabilidades de que los valores del atributo elegido se presenten o no en cada una de sus hojas, en este caso nos muestra probabilidades de que el atributo CANTIDAD tenga el valor 1 y 0 (que indica incidencia o ausencia respectivamente).
Una ventaja de este algoritmo sobre los demás es su poca complejidad computacional, permitiéndonos explotar información de cualquier tipo en relativamente poco tiempo.

Identificación de los patrones de distribución de las larvas de peces por estadio de desarrollo en diferentes escenarios ambientales.

Vinciguerria lucetia

Una de las especies que se destaca por su alta abundancia y distribución en el Océano Pacífico es ***Vinciguerria lucetia*** (pez linterna de Panamá). A pesar de su amplia distribución esta especie no es considerada de importancia comercial, sin embargo en algunas regiones se les captura para la elaboración de harinas de pescado. Los adultos se encentran en profundidades de 100 a 500 m (17), mientras que sus larvas se encuentran desde la superficie hasta aproximadamente 200 m (18).

Si bien la especie no es de valor comercial, si se considera de valor ecológico debido a su cualidad como especie indicadora ya que se ha observado que la abundancia de sus larvas, aun cuando éstas están presentes en el ambiente durante todo el año, cambia en cuanto a su distribución en distintas condiciones ambientales. Frente a las costas de la Península de Baja California, la distribución de sus larvas es abundante y uniforme durante los meses de verano y otoño, mientras que durante el invierno y la primavera son más escasas y sus abundancias se concentran en la región sur. Esta especie por tanto nos indica cambio estacional en el ambiente entre meses fríos y meses cálidos. Así mismo, se ha visto que su abundancia disminuye notablemente en años influenciados por el evento oceanográfico "La Niña" que constituye básicamente en un enfriamiento por debajo del promedio de la superficie del mar (2).

La variable de temperatura superficial parece ser la principal en afectar a la distribución de las larvas de esta especie. Sin embargo, una exploración más detallada de la información que se tiene sobre su distribución y de las variables ambientales en el área podría indicarnos si el comportamiento es el mismo en cada uno de los estadios de desarrollo de la especie, lo cual aportaría una información más fina sobre lo que ocurre en el ambiente.

De la aplicación del modelo de minería de datos en la información de las capturas de las larvas de esta especie, se observa que, como lo indican los antecedentes, el factor que determina la distribución de sus abundancias es la temperatura superficial (TS) (Figura 5.4 y 5.5). El modelo además nos indica que cuando los valores de TS están dentro de un rango considerado como "alto" (3), la mayoría de las larvas se encuentran en un estadio temprano (preflexión y flexión) o medio (postflexión) del desarrollo (19), mientras que un número reducido se encuentra en estadio de transformación (19) (Figura 5.4 y 5.5). Esto es coherente, ya que las larvas en estadio de transformación son aquellas que se están preparando para el hábitat del adulto, que en este caso son las aguas profundas, por lo que la temperatura superficial no sería un factor importante para su distribución.

El modelo nos muestra también que cuando la TS no está en un rango de valores "alto", sino "medio" (2) y "bajo" (1), son otros los factores que tienen influencia secundaria en la distribución de las larvas que no están en transformación (Figura 5.4 y 5.5). Cuando los valores de TS son bajos, -lo que representaría las zonas norteñas del área de estudio y/o las épocas climáticas frías del periodo estudiado- la distribución de las larvas está determinada por valores medios de salinidad superficial (SS), así como valores bajos y medios de biomasa zooplanctónica (BZ) (Figura 5.4 y 5.5). Cuando éstas condiciones se cumplen, las mayoría de larvas tienden a estar concentradas en la zona oceánica, en valores de LONGITUD mayores a -114.72 (Figura 5.4 y 5.5). Por otro lado, cuando los valores de SS son bajos, la distribución de las larvas está además

limitada por la LATITUD, lo cual está en coincidencia con la naturaleza transicional de la zona de estudio (2), en donde se espera que las especies de origen tropical-subtropical se distribuyan hacia el sur del área, mientras que las especies de origen templado se distribuirán hacia el norte.

Tratándose de TS dentro del rango de valores "medios", el modelo nos muestra información más desglosada de las larvas de la especie por estadio de desarrollo (Figura 5.4 y 5.5), lo que significa que más variables tienen influencia en su distribución. Las larvas en preflexión y transformación se encuentran en menor número, y de éstas, las de transformación se encuentran presentes en el área de estudio principalmente durante la noche (Figura 5.4 y 5.5). Las larvas en preflexión claramente estarían en desventaja en un ambiente frío o templado, ya que al ser una especie de afinidad tropical-subtropical, sus larvas tempranas tendrían mejores posibilidades de desarrollarse en ambientes cálidos. Por lado, las larvas en transformación están muy relacionadas con la hora del día, ya que esta especie tiene una alta capacidad de migración vertical, con la cual los adultos, juveniles y larvas en desarrollo avanzado tienden a subir a la superficie durante las horas de noche para alimentarse.

Los estadios de flexión y postflexión, son estadios en donde ocurre el desarrollo de las aletas, por lo cual el organismo va adquiriendo una mayor capacidad para desplazarse hacia sitios donde tengan una mayor oportunidad de sobrevivencia, generalmente sitios donde exista una mayor cantidad de presas de que alimentarse o un número menor de depredadores. La diversidad y abundancia de presas y depredadores depende de factores distintos en el ambiente, por lo que es lógico que no se vea una asociación clara de la distribución de estas larvas a una variable específica, excepto por la latitud que, nuevamente atribuido a la afinidad tropical-subtropical de esta especie (17), concentra a las larvas hacia el sur del área de estudio (Figura 5.4 y 5.5).

Diogenichthys laternatus

Los mictófidos (peces linterna) son una familia de peces de aguas profundas pero que presentan importantes migraciones verticales hacia la superficie, principalmente durante la noche. Varias especies de esta familia han sido consideradas por la FAO (Organización de las Naciones Unidas para la Agricultura y la Alimentación) como un grupo muy importante por su potencial pesquero ya que representan grandes biomasas en prácticamente todos los océanos del mundo (20).

Dentro de ésta familia, las larvas de ***Diogenichthys laternatus*** se distribuyen durante todo el año frente a las costas de la Península de Baja California y hasta Centro y Sur América, especialmente en los meses más cálidos (18). Siendo tan amplia su distribución es difícil concretar cuáles son las variables ambientales que más tienen influencia en ella.

Al analizar los resultados de la minería de datos, se observa que el componente latitudinal es importante en la distribución de las larvas de *D. laternatus* (Figura 5.6 y 5.7), lo que podría estar indicando zonas preferenciales para el desove y la crianza, es decir, larvas en distintos estadios de vida se encuentran en ubicaciones distintas, quizás como resultado de la deriva de estas larvas a distintos ambientes o, al parecer, como una estrategia para evitar la competencia por recursos. Esto último podría suponerse al observar que gran parte de las larvas en desarrollo.

temprano o intermedio que se encuentran agrupadas en la zona norte (LATITUD >=25.902 y <27.504; Figura 5.6 y 5.7), donde hay mayor abundancia de alimento, se encuentran asociadas a altos valores de biomasa zooplanctonica (BZ), mientras que las larvas en transformación se encuentran asociadas a la hora del día, es decir, dado la característica de migración vertical que

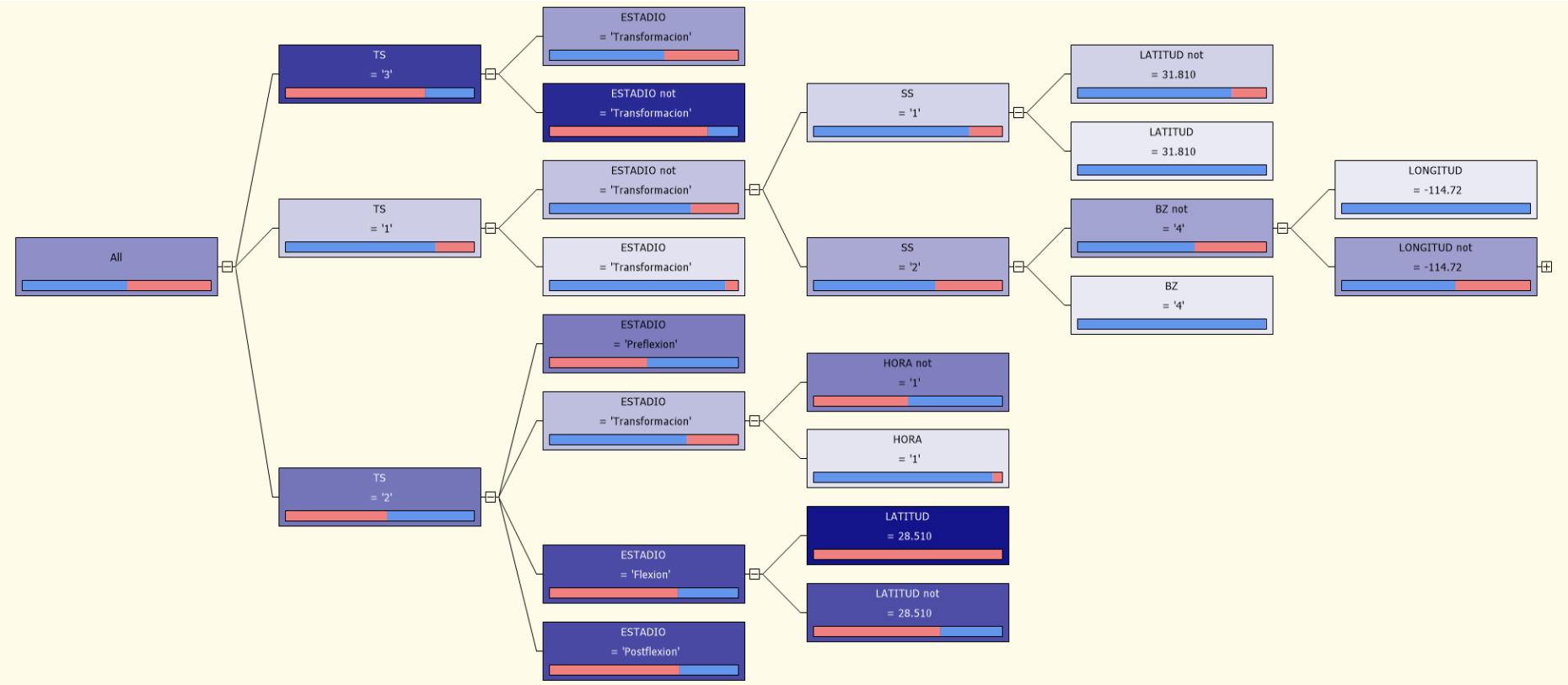


Figura 5.4 Árbol de distribución, especie Vinciguerria Lucetia.

Figura 5.5 Árbol de probabilidades, especie Vinciguerria Lucetia

tienen los adultos de la especie, éstas larvas parecen estar migrando hacia la superficie durante la noche (HORA not = 1; Figura 5.6 y 5.7) para alimentarse y evitar la competencia con larvas más pequeñas o de otras especies durante el día.

Así mismo, la temperatura es un factor importante para la distribución de las larvas en general en latitudes más norteñas (LATITUD >29.907 y >=27.504 <29.907; Figura 5.6 y 5.7). Esto tiene coherencia con el origen subtropical de la especia que, al disminuir la temperatura hacia las regiones templadas (norteñas), su abundancia va a disminuir y su reproducción va a estar restringida a que existan condiciones favorables del ambiente, siendo una de las principales la temperatura superficial del mar.

Por otro lado, las latitudes más sureñas (LATITUD <25.902; Figura 5.6 y 5.7), presentan claramente en general condiciones óptimas para la reproducción de los adultos y el desarrollo de las larvas, esto acorde, de nuevo, con la naturaleza subtropical de la especie.

Hygophum atratum

Al igual que en *D. laternatus*, la latitud es un factor determinante para las larvas de ***Hygophum atratum***. Ambas especies pertenecen a la misma familia de peces y son de afinidad tropical-subtropical (17) Analizando los resultados se observa que la distribución preferencial de las larvas de *H. atratum* es hacia el sur (LATITUD <25.902; Figura 5.8 y 5.9), en donde, al igual que en el caso de las larvas de *D. laternatus*, las condiciones ambientales en general son las idóneas para el desarrollo de estas larvas. En latitudes superiores, ya empiezan a ser relevantes otros factores que condicionan su distribución, en este caso la Salinidad Superficial (SS) y la BZ (Figura 5.8 y 5.9). Nuevamente las larvas en estadio de transformación se distinguen de resto, lo que nos indica, como en los casos anteriores, que su ambiente óptimo tiende a ser diferente del de los estadios menos desarrollados (Figura 5.8 y 5.9).

Triphoturus mexicanus

Triphoturus mexicanus es un mictófido de afinidad subtropical (17), su distribución es por tanto más restringida que la de las anteriores especies. La aplicación del modelo de minería de datos sobre la información que se tiene de las larvas colectadas, destaca la importancia de la TS en la distribución de estas (Figura 5.10 y 5.11). Las larvas menos desarrolladas se distribuyen acorde al rango de temperaturas más bajas (“TS =1”; Figura 5.10 y 5.11), mientras que las larvas más desarrolladas se encuentran en valores más elevados de temperatura (“TS not = 1”; Figura 5.10 y 5.11). Otro factor importante para la distribución de las larvas fue lo hora del día en la que fueron colectadas, colectándose un mayor número de larvas durante la noche (“HORA=0”; Figura 5.10 y 5.11). El modelo confirma lo que se observa para la distribución de los adultos, restringidos por las regiones climáticas (temperaturas de latitudes subtropicales) y su desplazamiento en la columna de agua (migraciones nocturnas) (18).

Familia Batilágidae (Leuroglossus stilbius, Bathylagoides wesethi)

A diferencia de las especies anteriores, las especies de la familia Batilágidae (peces lengua) se distribuyen en toda la columna de agua. Estos peces son considerados por la FAO como especies de potencial pesquero por su amplia distribución y abundancia y su alto valor alimenticio (20), sin embargo su pesquería no se ha desarrollado a nivel industrial. A pesar de no existir muchas especies distribuidas en el Pacífico Mexicano (18), se ha observado que los adultos no responden

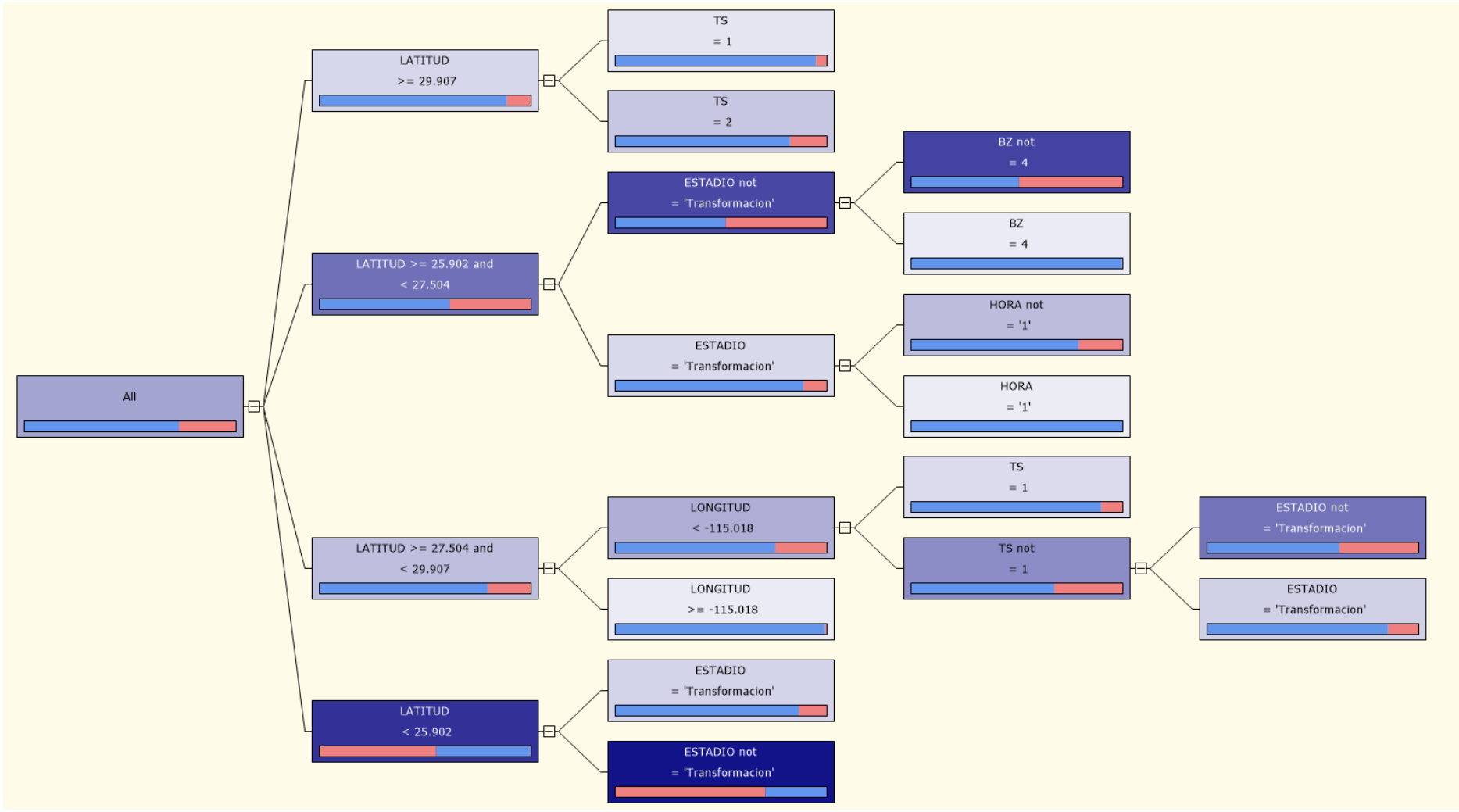


Figura 5.6 Árbol de distribución, especie Diogenichthys laternatus

Figura 5.7 Árbol de probabilidades, especie Diogenichthys laternatus

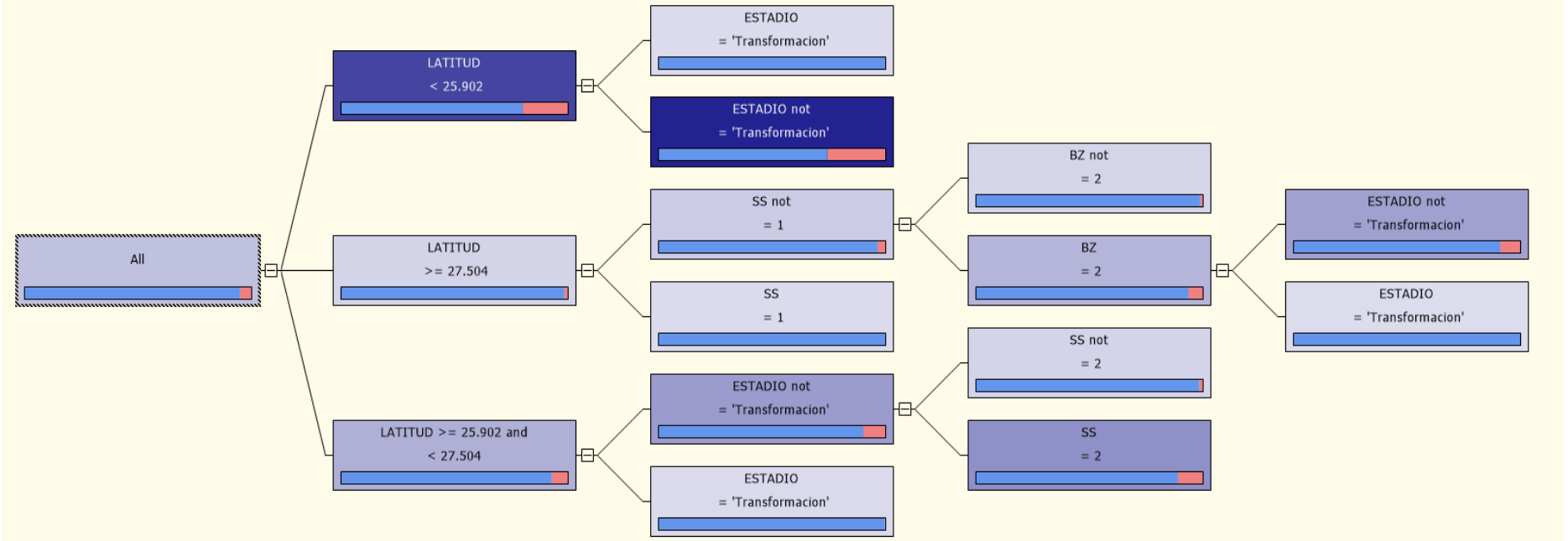


Figura 5.8 Árbol de distribución, especie Hygophum atratum

Figura 5.9 Árbol de probabilidades, especie Hygophum atratum

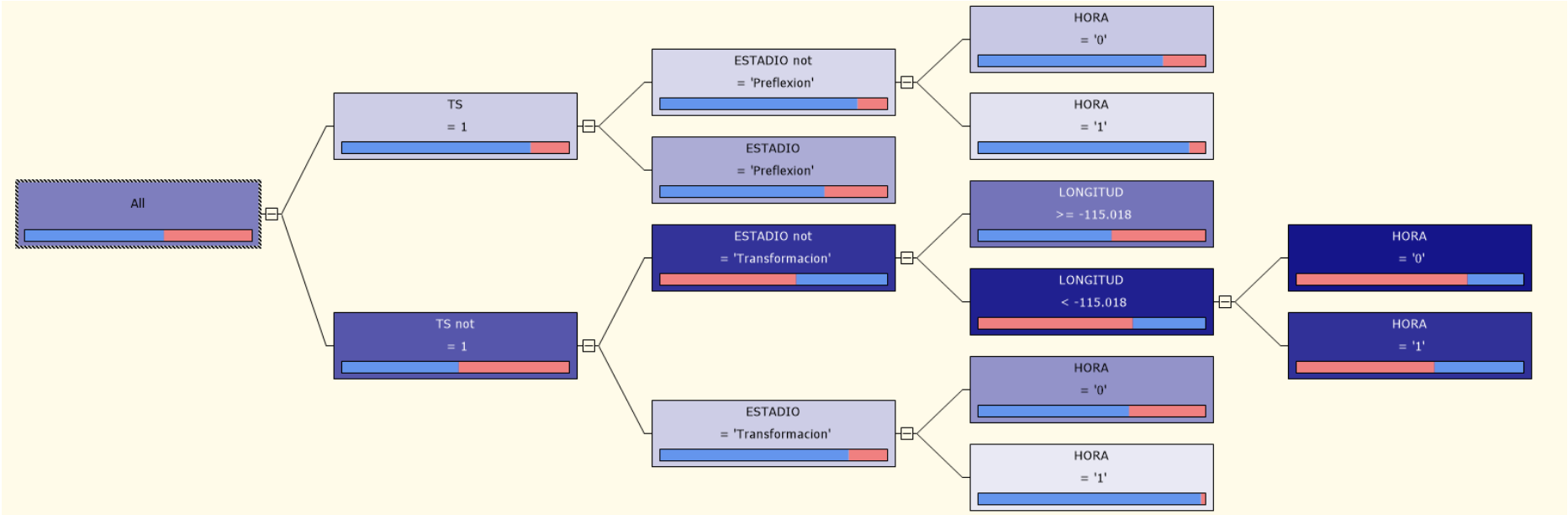


Figura 5.10 Árbol de distribución, especie Triphoturus mexicanus

Figura 5.11 Árbol de probabilidades, especie Triphoturus mexicanus

de igual manera ante las mismas variables ambientales. Esto se evidencia al aplicar el modelo sobre las bases de datos, en donde se observa que para las larvas de ***Leuroglossus stilbius***, de afinidad templada (17), la TS es un factor definitivo para la distribución de sus larvas (Figura 5.12 y 5.13) y la BZ el factor secundario para la distribución de las larvas menos desarrolladas (Figura 5.12 y 5.13). Mientras que para las larvas de ***Bathylagoides wesethi***, de aguas más cálidas (17), la cercanía a la costa (LONGITUD <-115.685; Figura 5.14 y 5.15) es un factor determinante para la distribución de las larvas tempranas (ESTADIO = Preflexión, ESTADIO = Flexión; Figura 5.14 y 5.15), y no es tan importante para larvas más desarrolladas (ESTADIO = Postflexión, ESTADIO = Transformación; Figura 5.14 y 5.15).

Familia Sebastidae (Sebastes tipo 1, Sebastes tipo 2)

Al aplicar el modelo a las bases de datos de larvas de especies costeras se observa que estas, aun perteneciendo a la misma familia y teniendo la misma zona de distribución, tienen un comportamiento de distribución afectado por distintas variables ambientales. Las larvas de los peces piedra (Familia Sebastidae) reflejan este comportamiento, ***Sebastes* tipo 1** y ***Sebastes* tipo 2** presentan exactamente la misma distribución en el área de estudio (17), sin embargo las de la especie 1 están relacionadas con la región geográfica y la hora del día (Figura 5.16 y 5.17), mientras que las de la especie 2 se relacionan directamente con los valores de temperatura, prefiriendo los valores más bajos (Figura 5.18 y 5.19). Se ha reportado en la bibliografía, que aun cuando los adultos de la especie se distribuyen en las mismas zonas, éstos tienden a exhibir diferentes estrategias reproductivas precisamente para evitar la competencia entre sus larvas (18). Los resultados que arroja el modelo podrían estar reflejando, en la distribución asociada a las variables, el resultado de estas distintas estrategias reproductivas.

Especies costeras (Engralius mordax, Sardinops sagax)

Especies costeras de importancia comercial, y por tanto ampliamente estudiadas, generalmente se caracterizan por la migración de los adultos para reproducirse hacia sitios específicos a lo largo de la costa de Estados unidos y la Península de Baja California, de manera que algunas de sus poblaciones se han aislado geográficamente generando, en algunos casos, subespecies (17).

Al aplicar el modelo sobre las bases de datos de algunas de estas especies, se observa que los atributos de LATITUD y LONGITUD segregan la distribución de las larvas en el área de estudio. En el caso de ***Engraulis mordax*** (anchoveta de California) al parecer se caracterizan dos sitios de desove, uno en latitudes mayores y con mayor abundancia larval a latitudes mayor o igual a 29.907 (Figura 5.20 y 5.21) y otro sitio, con valores larvales menores, a latitudes menores (Figura 5.20 y 5.21). Esto es congruente con la distribución de las poblaciones más grandes de esta especie, que se localizan hacia las costas de California, donde desovan todo el año (20), mientras que poblaciones más pequeñas desovan principalmente durante primavera al sur del área de estudio.

Por otro lado, ***Sardinops sagax*** (sardina de California) también tiende a formar cardúmenes costeros muy grandes, con preferencia a reproducirse en aguas oceánicas (20). Al aplicar el modelo sobre la base de datos de las larvas, se observan dos agrupaciones de ellas: unas hacia la región oceánica (LONGITUD >=-115.018; Figura 5.22 y 5.23) en donde se concentran la mayoría de las larvas, y la mayoría de estas en desarrollo temprano (Figura 5.22 y 5.23), y otro grupo hacia la zona costera, con menores abundancias (Figura 5.22 y 5.23). La TS es también un factor para la distribución de estas larvas, y aquellas que están en la región oceánica se encuentran por lo general asociados con temperaturas en rangos altos (“TS =3”; Figura 5.22 y

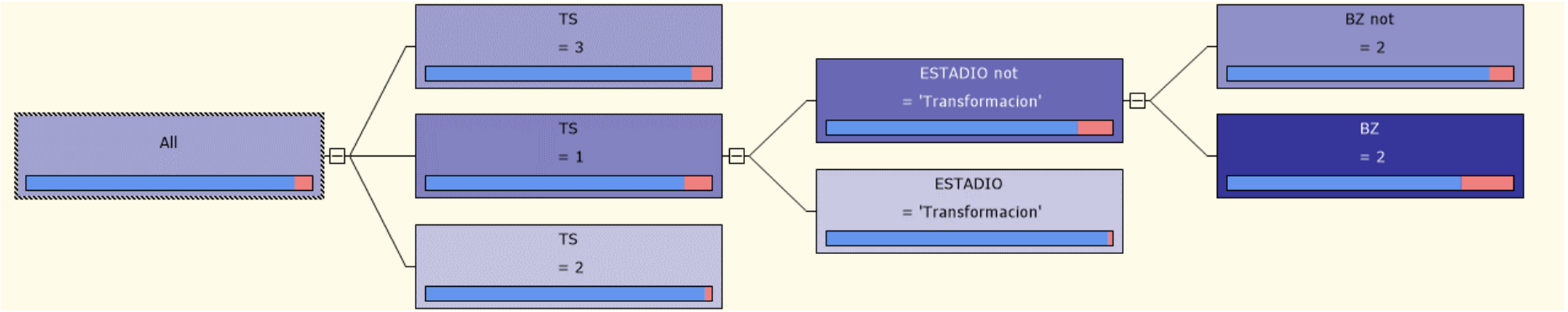


Figura 5.12 Árbol de distribución, especie Leuroglossus stilbius

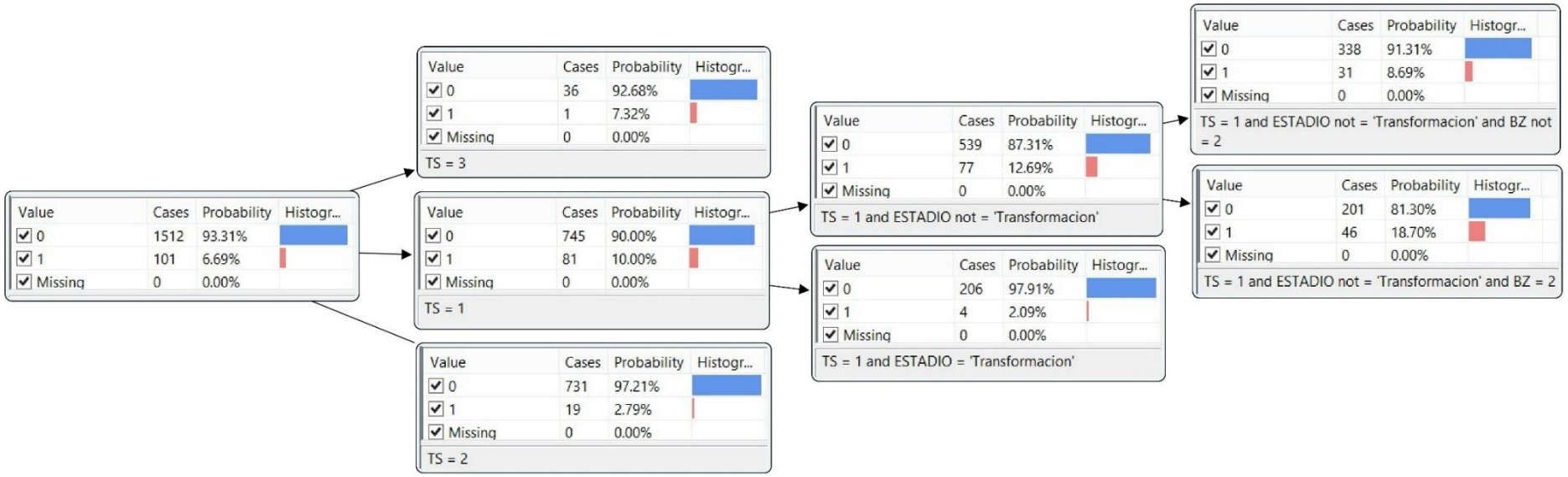


Figura 5.13 Árbol de probabilidades, especie Leuroglossus stilbius

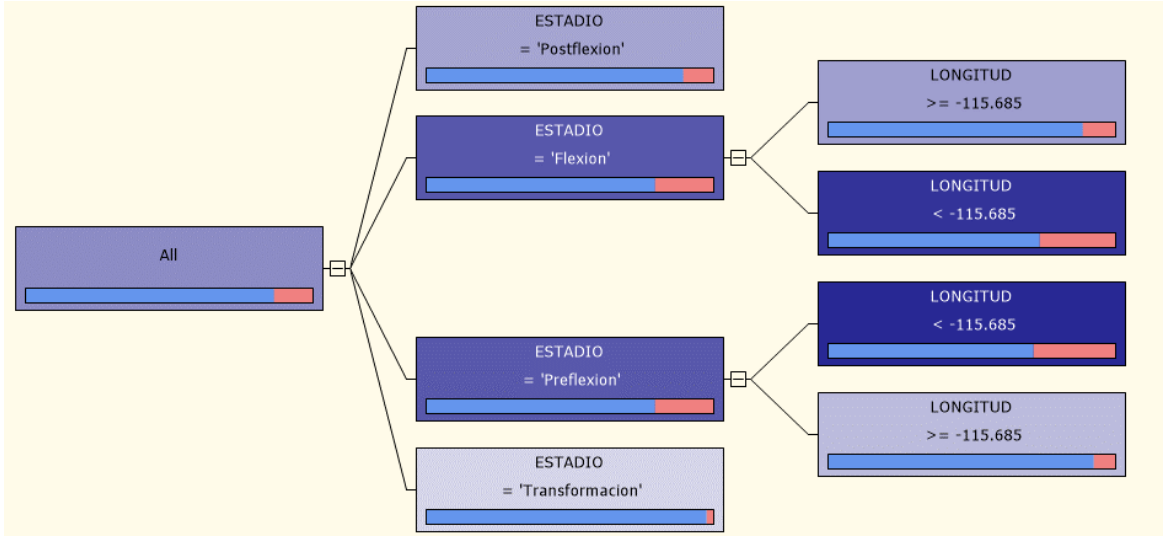


Figura 5.14 Árbol de distribución, especie Bathylagoides wesethi

Figura 5.15 Árbol de probabilidades, especie Bathylagoides wesethi

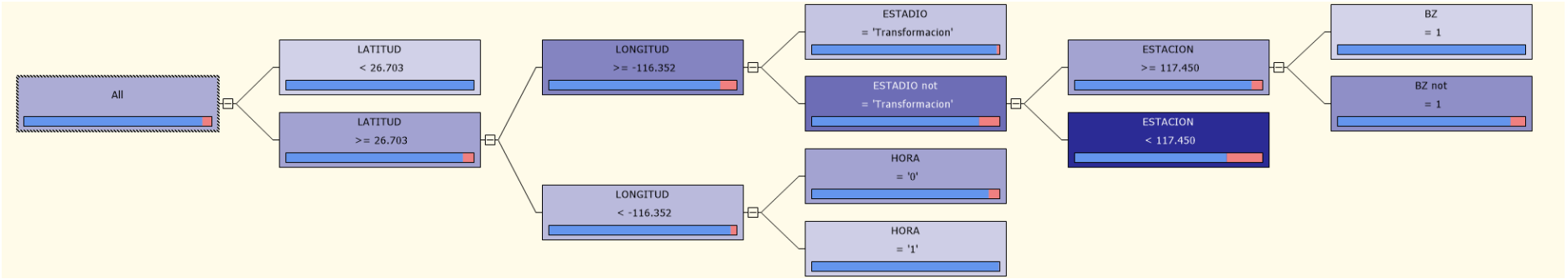


Figura 5.16 Árbol de distribución, especie Sebastes sp1

Figura 5.17 Árbol de probabilidades, especie Sebastes sp1

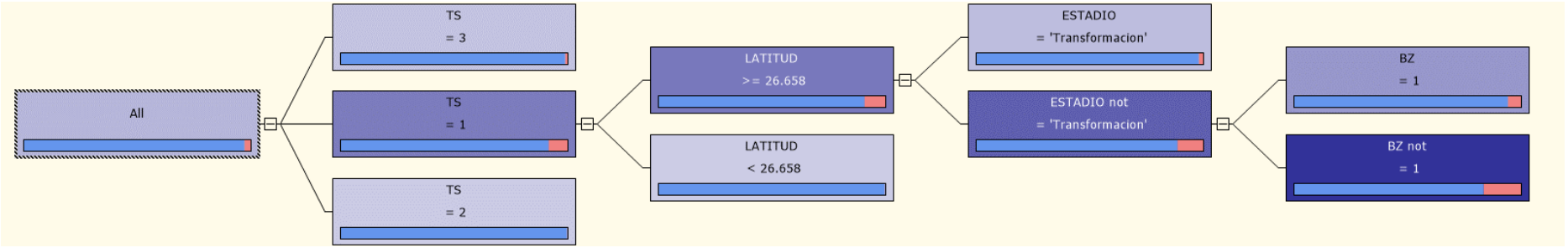


Figura 5.18 Árbol de distribución, especie Sebastes sp2

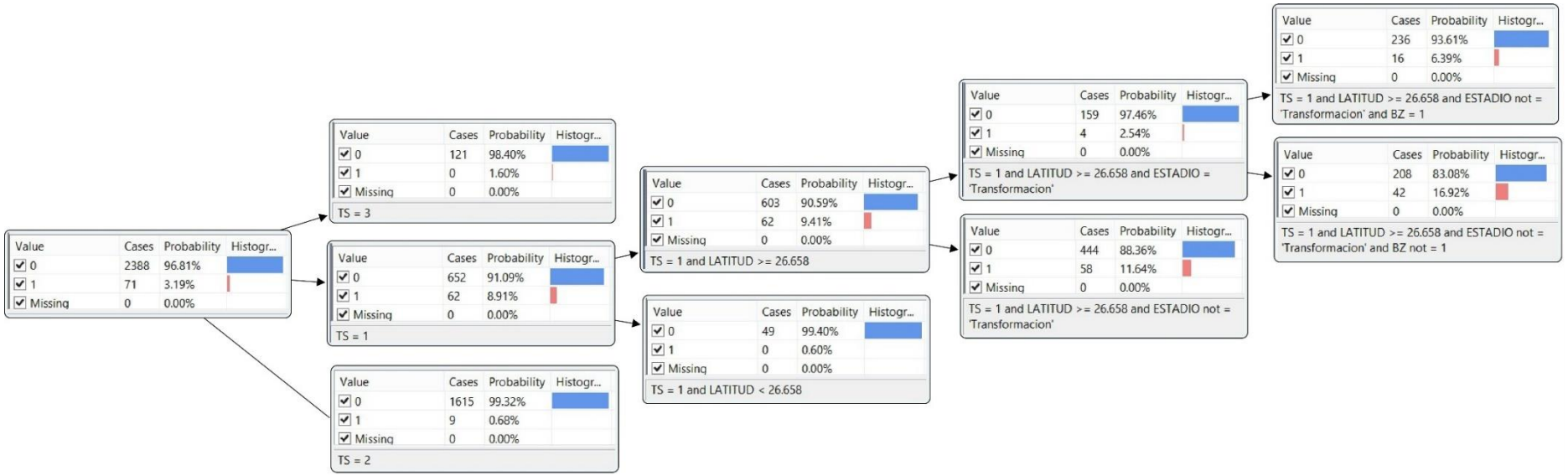


Figura 5.19 Árbol de probabilidades, especie Sebastes sp2

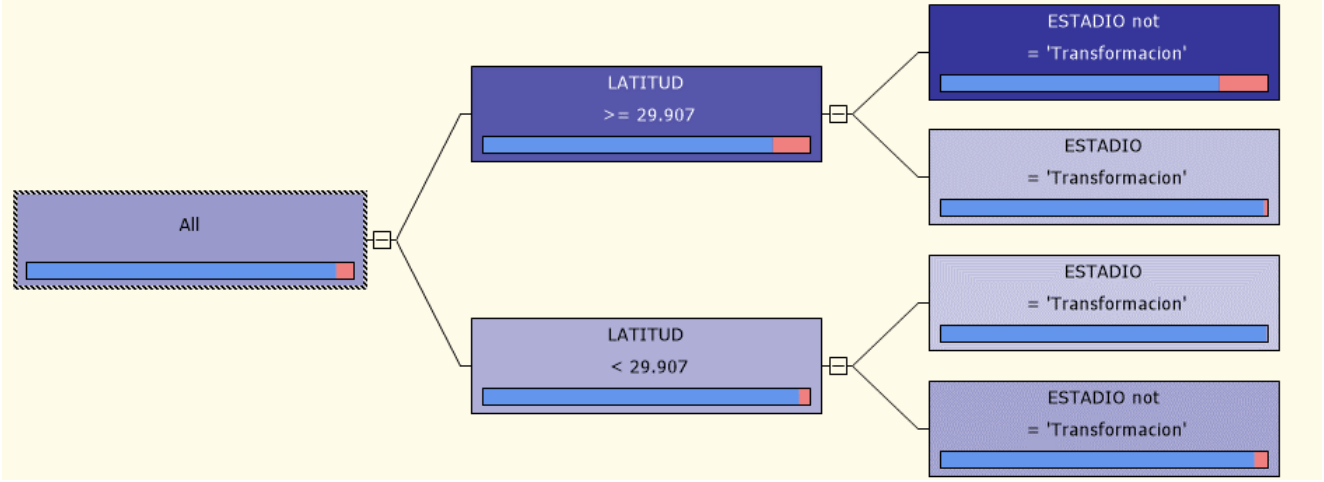


Figura 5.20 Árbol de distribución, especie Engralius mordax

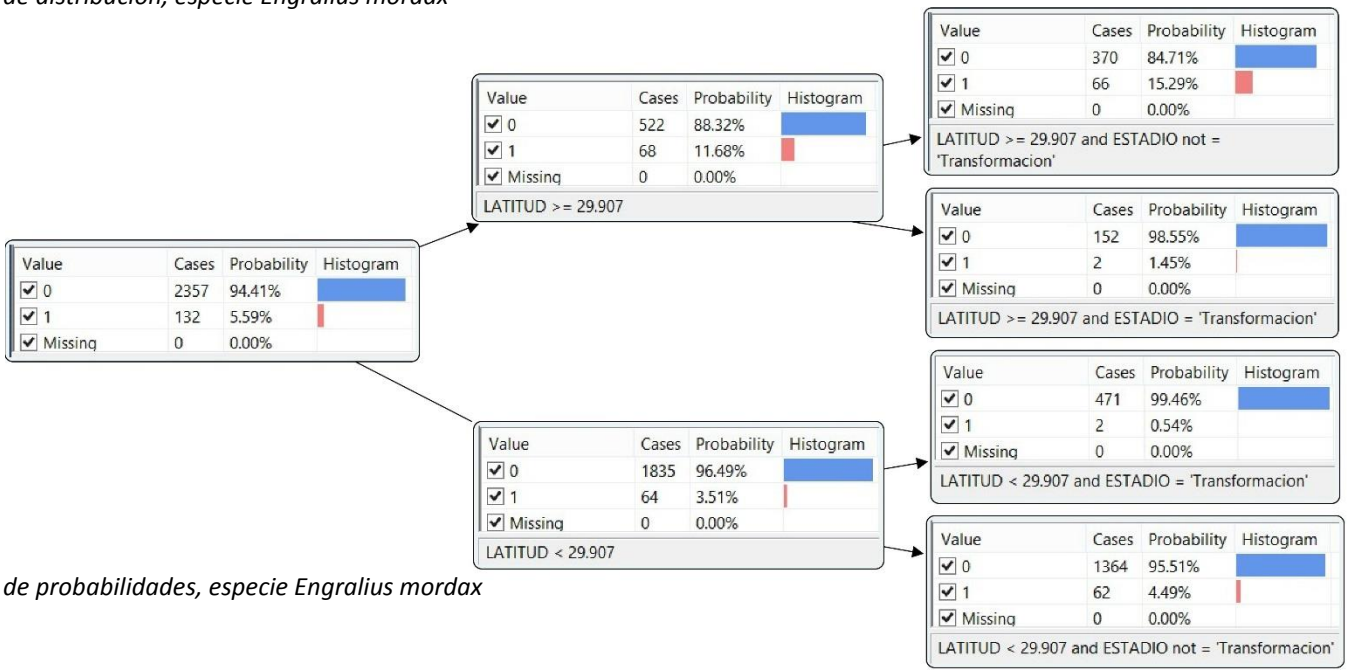


Figura 5.21 Árbol de probabilidades, especie Engralius mordax

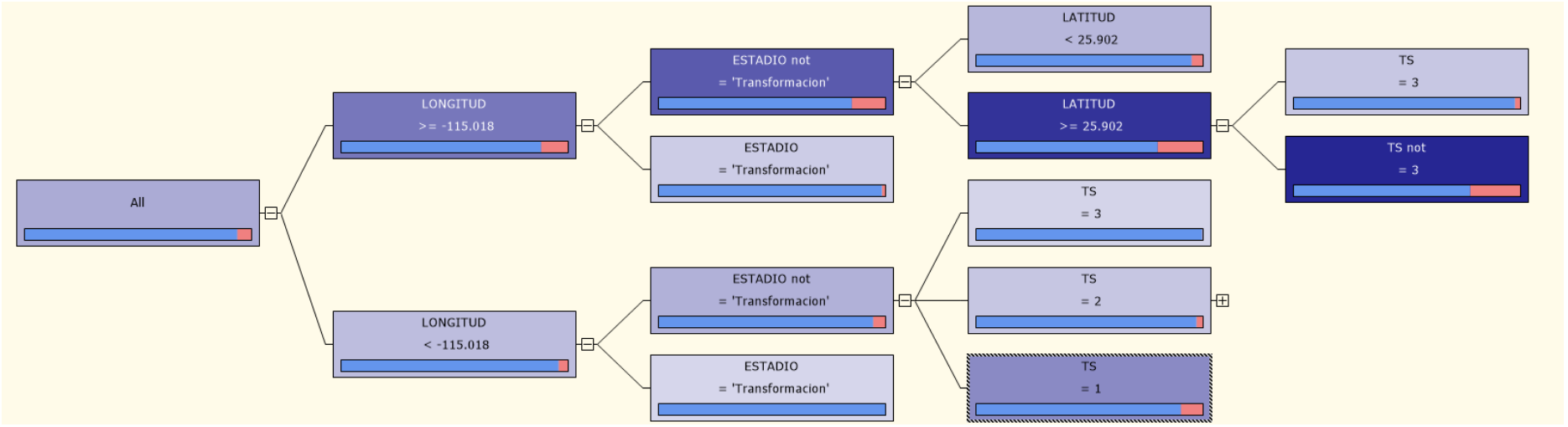


Figura 5.22 Árbol de distribución, especie Sardinops sagax

Figura 5.23 Árbol de probabilidades, especie Sardinops sagax

5.23), mientras que aquellas que se encuentra más cercanas a la costa se asocian a zonas donde la temperatura es más baja (Figura 5.22 y 5.23). Por lo general, zonas costeras con temperaturas bajas se asocian a zonas de "surgencia", donde las aguas más frías y nutritivas del fondo emergen hacia la superficie, lo que crea zonas preferenciales para la alimentación de las larvas de muchas especies marinas (21).

Synodus lucioceps

Synodus lucioceps (pez lagarto) es una especie que vive estrechamente asociada al fondo marino hasta profundidades de aproximadamente 200 m (20). Al analizar los datos, se observa que la mayoría de las larvas tienden a distribuirse hacia la zona oceánica (LONGITUD <=-112.984; Figura 5.24 y 5.25), mientras que tres grupos más pequeños, en distintos estadios de desarrollo, se distribuyen en diferentes latitudes hacia la zona costera, asociados también con los valores de salinidad superficial (SS) (Figura 5.24 y 5.25) y secundariamente con la TS (Figura 5.24 y 5.25). La distribución de los grupos costeros podría estar asociada a la cualidad de los adultos de distribuirse en "parches" a lo largo de la zona costera (18), habitando sitios en donde las salinidades pueden ser muy altas debido a condiciones inter-mareales o de desecación.

Otras especies (Trachurus symetricus, Merluccius productus, Diogenichthys atlanticus, Stomias atriventer)

Bases de datos de larvas de otras especies fueron analizadas mediante este modelo, tanto costeras ***Trachurus symetricus*** (Figura 5.26 y 5.27), que habitan en el fondo oceánico***: Merluccius productus*** (Figura 5.28 y 5.29) y en toda la columna de agua: ***Diogenichthys atlanticus*** y ***Stomias atriventer*** (Figura 5.30 y 5.31, Figura 5.32 y 5.33). Los resultados no fueron tan concluyentes como los descritos de las especies anteriores, sin embargo esto parece atribuirse principalmente a la falta de variables del ambiente con las cuales se pueda contrastar la distribución de las larvas de estas especies.

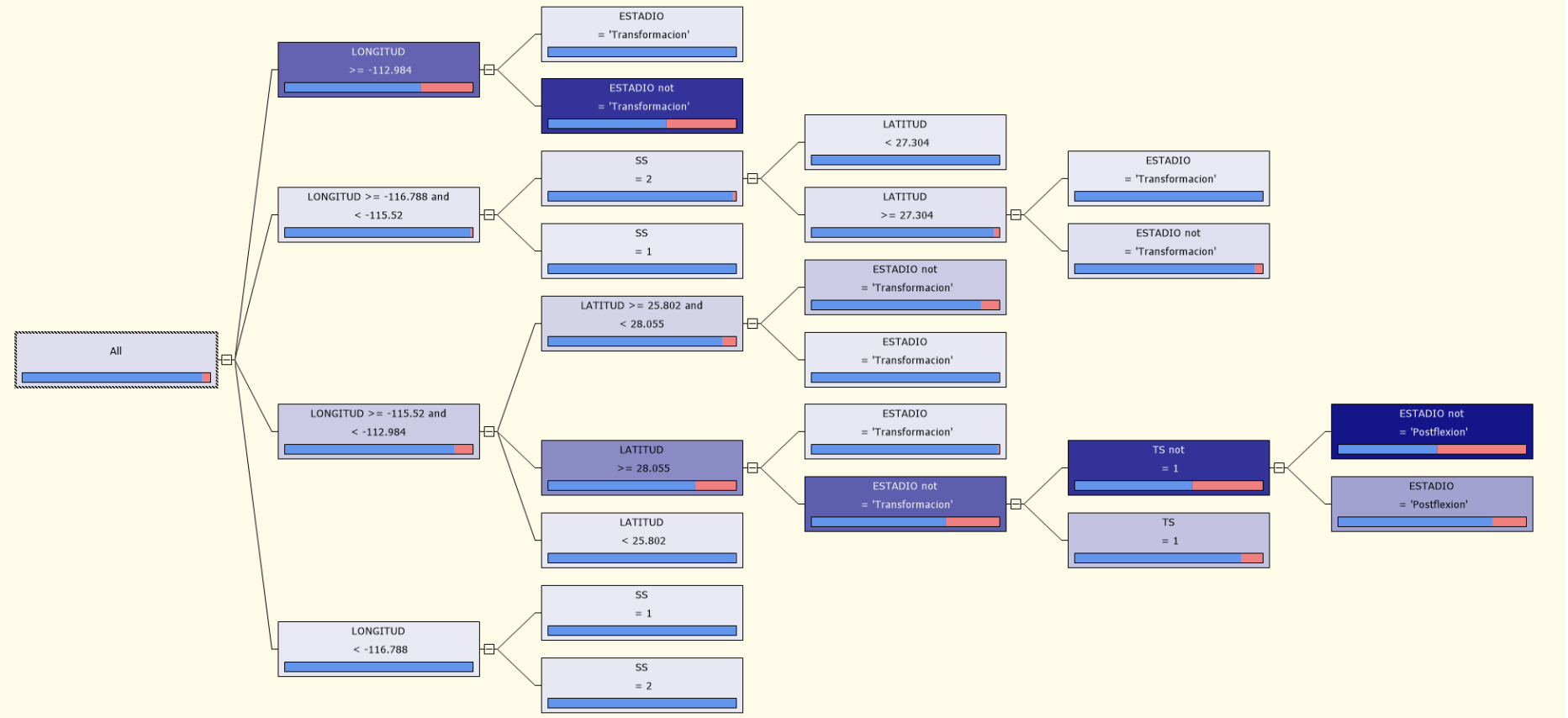


Figura 5.24 Árbol de distribución, especie Synodus lucioceps

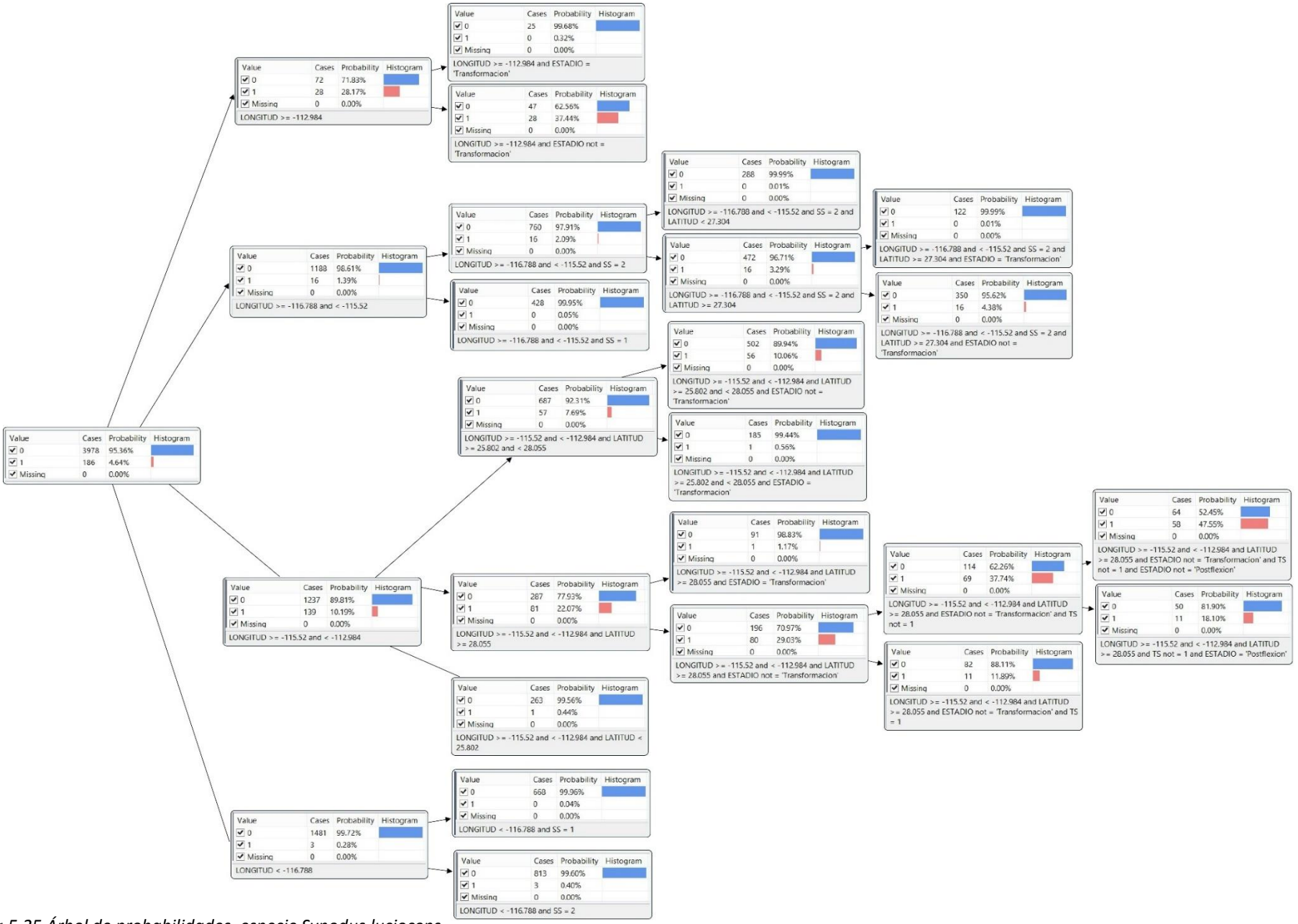


Figura 5.25 Árbol de probabilidades, especie Synodus lucioceps

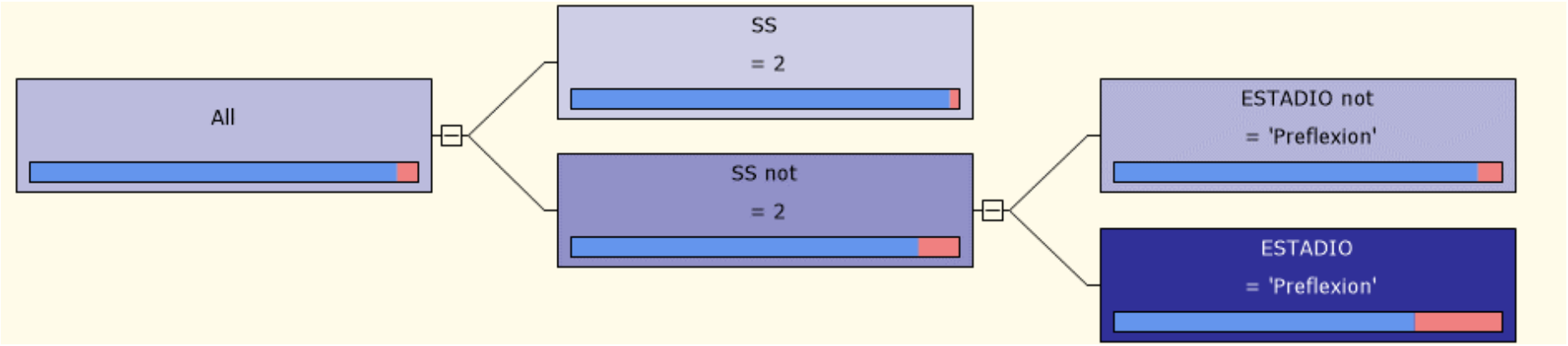


Figura 5.26 Árbol de distribución, especie Trachurus symetricus

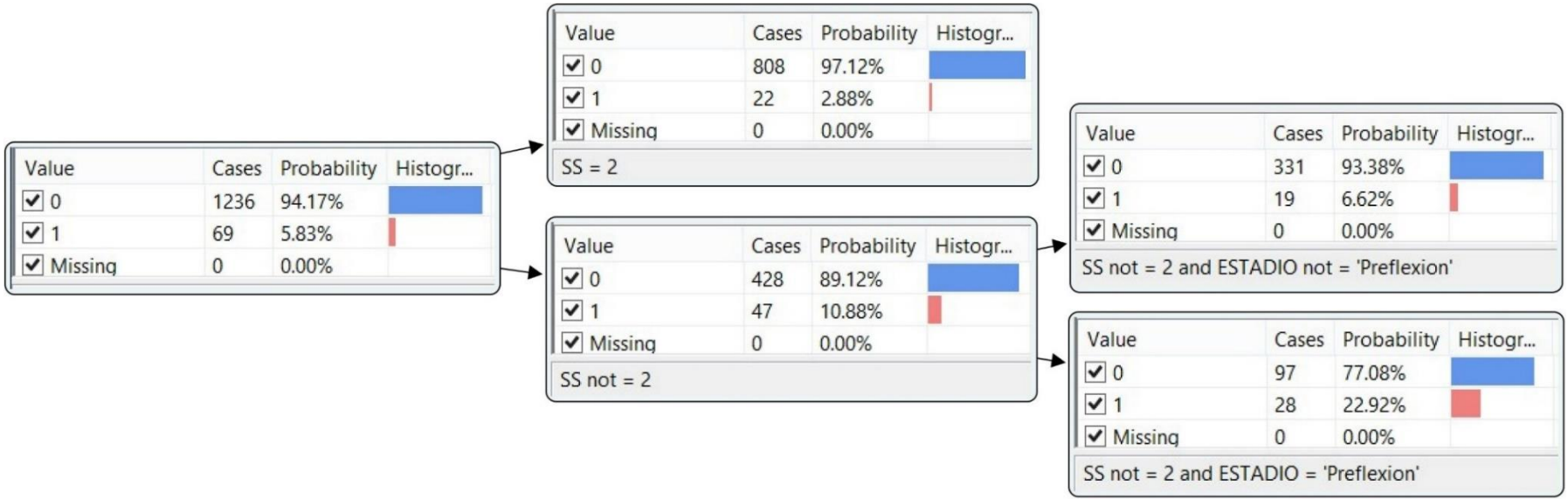


Figura 5.27 Árbol de probabilidades, especie Trachurus symetricus

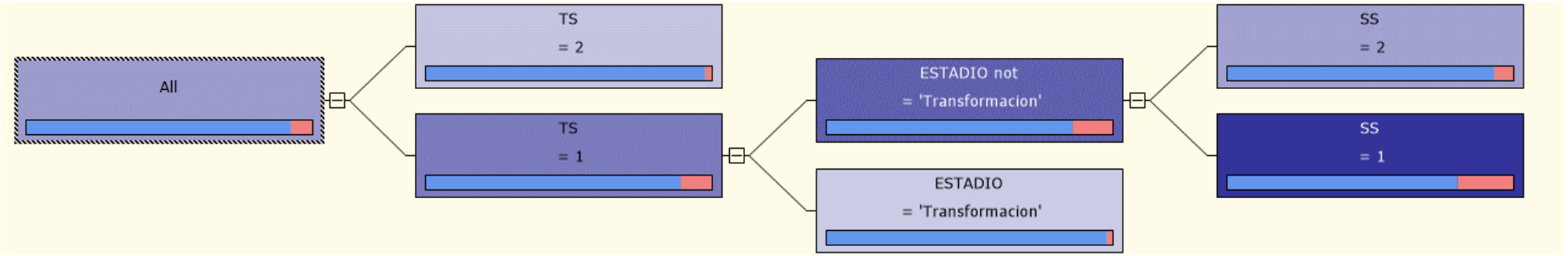


Figura 5.28 Árbol de distribución, especie Merluccius productus

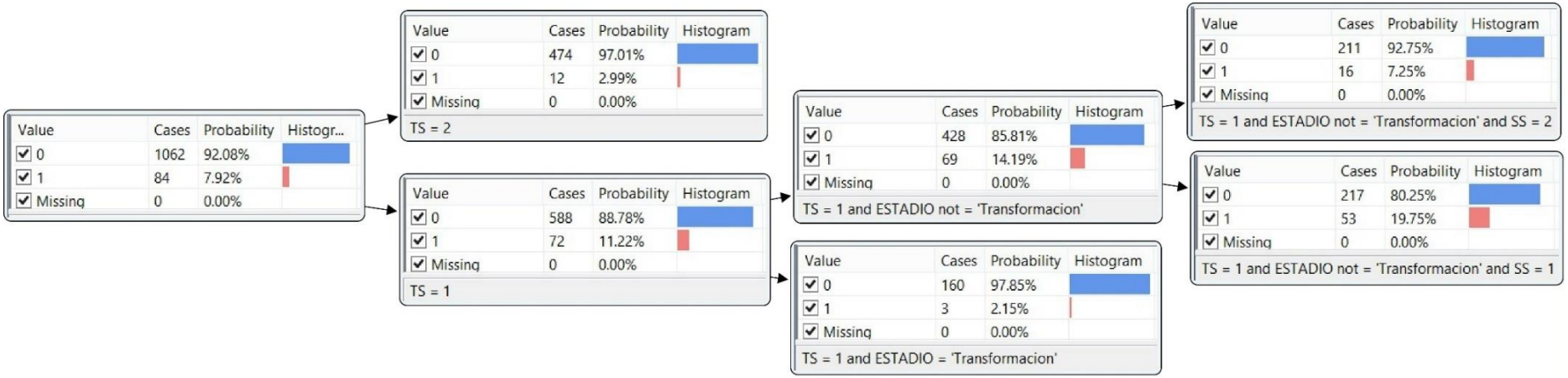


Figura 5.29 Árbol de probabilidades, especie Merluccius productus

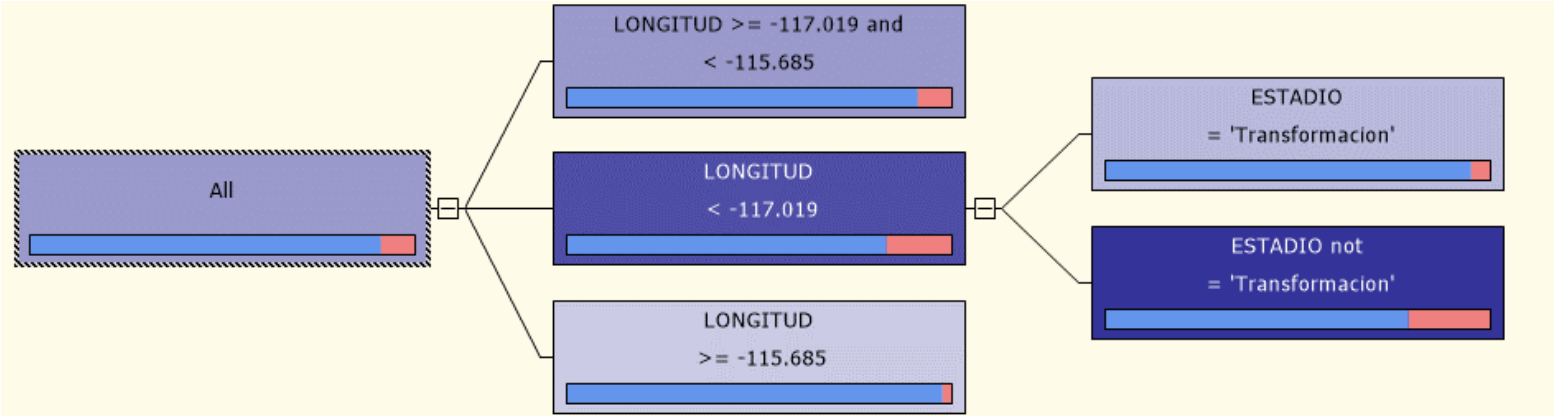


Figura 5.30 Árbol de distribución, especie Diogenychtis atlanticus

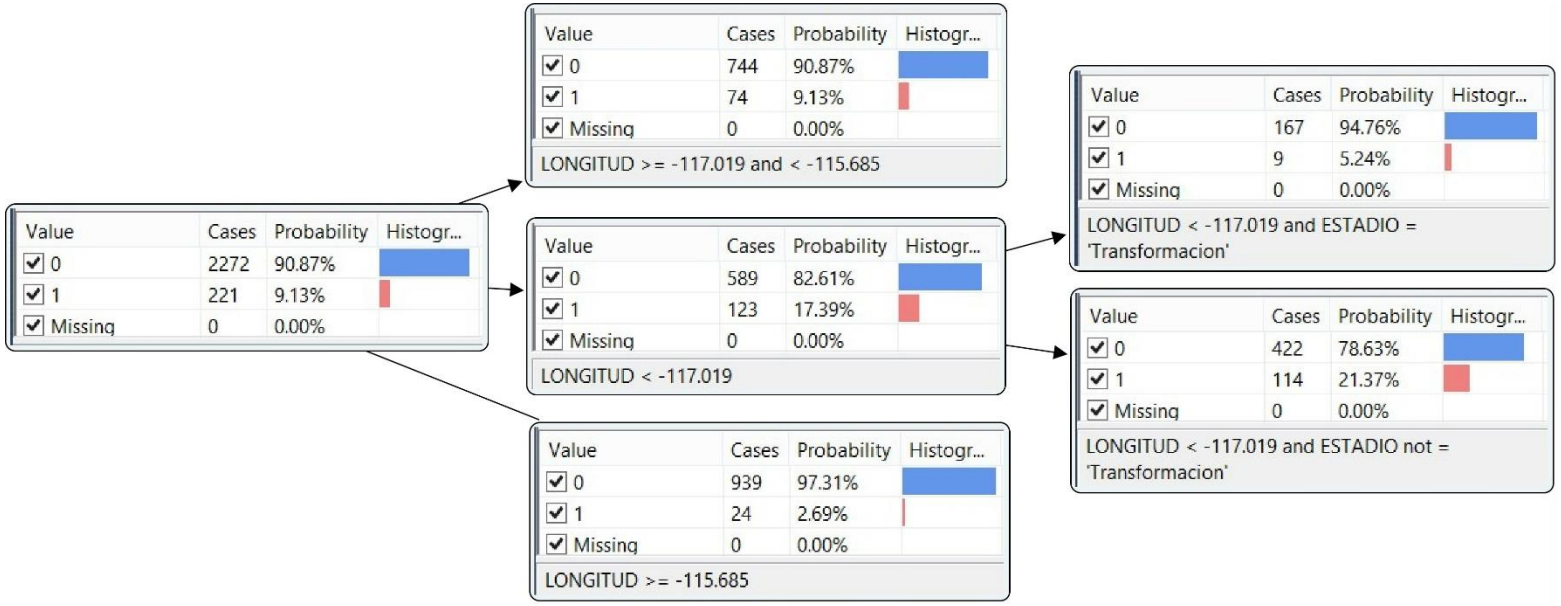


Figura 5.31 Árbol de probabilidades, especie Diogenychtis atlanticus

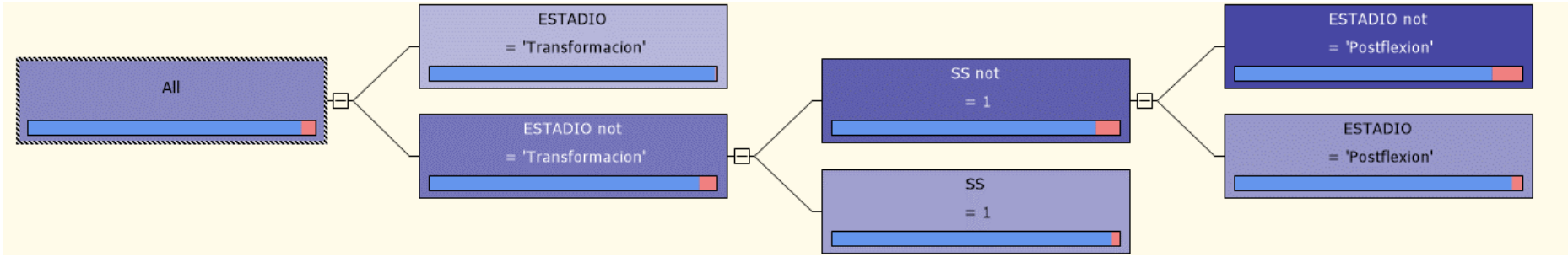


Figura 5.32 Árbol de distribución, especie Stomias atriventer

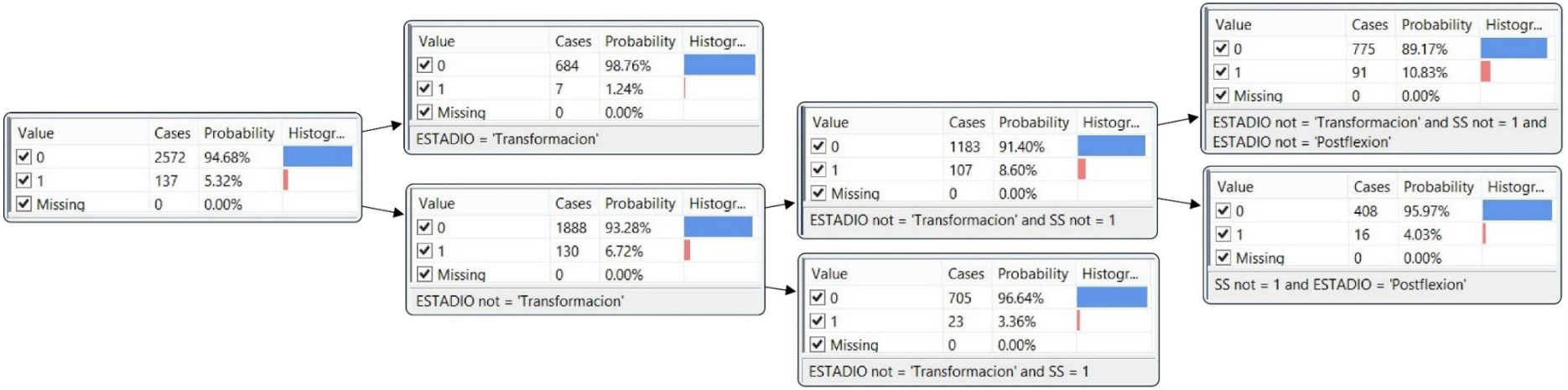


Figura 5.33 Árbol de probabilidades, especie Stomias atriventer

5.2 Conclusiones

Se cumplió con los objetivos planteados en el proyecto:

- La selección de los atributos en el modelo se basó principalmente en el conocimiento que se tiene sobre la distribución de los adultos y larvas de las especies más abundantes que se encuentran registradas en las bases de datos de la Colección ICTIPLANCTON. Aun así, es posible que, al momento de la recolecta de las variables ambientales, se sub o sobre-estime la importancia de su relación con la distribución de las especies, y esto es algo que se debe tomar en cuenta en el momento de la interpretación de resultados.
- Como producto de este trabajo, se generó una base de datos para SQL Server 2012 a partir de la información contenida en las bases de datos de la Colección ICTIOPLANCTON de CICIMAR-IPN contenida en hojas de cálculo a través de un proceso ETL, en el que se seleccionó un conjunto de datos inicial y se realizó una discriminación de atributos en base a su importancia para la incidencia o ausencia de una especie en determinado estadio de desarrollo. Como parte del proceso ETL se limpiaron y estructuraron los datos y posteriormente se creó un modelo de minería de datos que permitió explotar la información biológica no lineal a través de un algoritmo de árboles de decisión, que generó resultados positivos del proceso de minado que se aplicó.
- La generación de este modelo permitió agilizar el análisis inicial de los datos contenidos en las bases de datos de la Colección ICTIOPLANCTON. Los resultados expuestos muestran una alta correlación con lo que se conoce del habitad de las especies de peces analizadas. En gran medida los resultados que destaca el modelo suprimen la necesidad de hacer análisis estadísticos “a priori” sobre los datos para visualizar cada uno de los atributos que tendrá influencia en la distribución de las especies. A partir del análisis de los resultados del modelo se pueden tomar decisiones sobre qué tipo de análisis “a posteriori” se aplicará con mayor eficacia en cada caso.

5.3 Trabajo Futuro

Debido a que el trabajo se concentró en la limpieza, selección y procesamiento de la información, en la creación de la base de datos el modelo de minería, no se construyó una aplicación que englobe todos los procesos en uno solo. Se propone como trabajo futuro el desarrollo de una aplicación de este tipo, de manera que el manejo del modelo sea más amigable para el usuario.

Asimismo, sería deseable generar una herramienta que permita al usuario introducir la información generada y mapear geográficamente la distribución de las especies contra la distribución de los atributos (o variables ambientales), de manera que el proceso de análisis sea más visual y se simplifique su entendimiento y comparación.

Anexo: Glosario de términos biológicos.

Abundancia: Número de individuos por unidad de área, distancia o tiempo durante el esfuerzo de observación o recolecta.

Ambiente: Es el conjunto de elementos físico-químicos, geológicos y biológicos interrelacionados, que producen los diferentes recursos que requieren los organismos para perpetuarse a través del tiempo.

Biomasa: Cantidad de materia orgánica que forma parte de los organismos. Se expresa en unidades de volumen, de peso fresco o peso seco, o en unidades de energía.

Crucero Oceanográfico: Expedición de trabajo que se lleva a cabo en una embarcación para realizar mediciones oceanográficas.

Comunidad: Grupo de organismos pertenecientes a taxa distintos que ocurren en el mismo hábitat o área, que interactúan mediante relaciones tróficas y espaciales. Típicamente está caracterizada por la referencia a una o más especies dominantes.

Diversidad: Número de especies de una comunidad o muestra; riqueza de especies o medida del número de especies y su abundancia relativa en la comunidad.

Epipelágico: Se refiere a los organismos que viven en la columna de agua entre la superficie y los 200 m de profundidad, en el océano abierto.

ENSO: (El Niño Oscilación del Sur) Desequilibrio océano-atmosférico en donde la presión atmosférica cambia, provocando el debilitamiento de los vientos alisios del Este, por lo que favorece una invasión anormal de aguas cálidas del trópico hacia latitudes altas, afectando principalmente las costas del Pacífico Oriental.

Estadio: Etapa o fase de un proceso, desarrollo o transformación.

Evento: Variación no periódica de un fenómeno (no necesariamente predecible), que se presenta con magnitud y duración variables. Ejemplo, los eventos de fuertes vientos llamados "Nortes", el fenómeno de "El Niño".

Hábitat: Conjunto de recursos y condiciones ambientales definidos espacio-temporalmente que determinan la presencia, supervivencia y reproducción de una población o especie.

Ictioplancton: La fase planctónica de la mayoría de los peces; comprende tanto a los huevos como a las larvas de éstos. Sus desplazamientos dependen principalmente de las corrientes de agua.

Indicadores biológicos: son las plantas o animales que aportan información fidedigna sobre el estado de conservación o características físicas, químicas y/o biológicas de un ecosistema particular.

Larva: estadio que precede a la eclosión del huevo; es diferente en forma y pigmentación al juvenil y al adulto, y debe sufrir una etapa de transformación antes de asumir las características del adulto.

Masas de agua: Un volumen de agua usualmente identificado por valores típicos de temperatura y salinidad que le son característicos y que permiten distinguirlo de las aguas circundantes, Su formación ocurre en la interfaz con la atmósfera y por la mezcla de dos o más tipos de agua.

Pelágico: Se refiere a los organismos que viven en la columna de agua, independientes del fondo oceánico.

Plancton: Conjunto de organismos, tanto animales como vegetales, que habitan a la deriva en la columna de agua. Su capacidad de movimiento es insuficiente para evitar ser transportados pasivamente por las corrientes.

Profundidad de la capa de mezcla: Es el espesor de la capa superficial que es mezclada por el efecto del viento y la advección; en la práctica se caracteriza por la distribución vertical de temperatura y salinidad.

Surgencia: Ascenso de aguas superficiales, más frías y con mayor concentración de nutrientes, que reemplazan las aguas superficiales en zonas restringidas del océano. Las surgencias más importantes que se presentan en el océano son las llamadas surgencias costeras, las cuales son provocadas por vientos hacia el ecuador en los océanos con frontera oriental.

Variable ambiental: descriptor físico, químico, geológico y/o biológico que permite identificar una característica del ambiente.

Zooplancton: Animales que forman parte del plancton. Comunidad de animales que flotan libremente en el agua, incapaces de moverse en contra de las corrientes.

Bibliografía

1. Ekman, S. 1953. *Zoogeography of the sea*. Sidwick and Jackson Ltd. London. 417 pp.
2. Moser, H. G., P. E. Smith & L. E. Eber. 1987. Larval fish assemblages in the California Current Region, 1954-1960, a period of dynamic environmental change. *CalCOFI Rep.*, 27: 97-127.
3. Hammond, A., A. Adriaanse, E. Rodenburg, D. Bryant y R. Woodward. 1995, Environmental indicators: A systematic approach to measuring and reporting on Environmental Policy & Performance in the context of Sustainable Development. World Resources Institute.
4. Jiménez-Rosenberg, S. P. A. y G. Aceves-Medina. 2009, Indicadores Biológicos en el ambiente marino. Oceánides.
5. Gomez-Flechoso, A.J.1998, Induccion de Conocimiento con incertidumbre en Bases de Datos Relacionales Borrosas. http://www.gsi.dit.upm.es/~anto/tesis/html/stateart.html
6. Zhu, X. y I. Davidson. 2007, Knowledge Discovery and Data Mining: Challenges and Realities. IGI Global.
7. Silberchatz, A. 2001, Database system concepts 4e. McGraw-Hill.
8. Date, C. J. 2001, Introducción a los sistemas de bases de datos. Prentice Hall.
9. Haag, S. y M. Cummings. 2012, Management Information Systems for the information age. McGraw-Hill/Irwin.
10. Torres-Jiménez, J. 2011, Breve Introducción a las Bodegas de Datos. http://www.tamps.cinvestav.mx/~jtj/courses/dbs/slides/Bodegas%20de%20datos.pdf
11. Gupta, A. y I. Mumick. 1999, Materialized views: techniques, implementations, and applications. The MIT Press.
12. Chaudhuri, S. y U. Dayal. 1997, An overview of data warehousing and OLAP technology. SIGMOD Record.
13. Tamayo M. y F. J. Moreno. Comparing the MOLAP the ROLAP storage models.
14. Stanek, W. R. 2012, Microsoft SQL Server 2012 Pocket Consultant.
15. Microsoft.com. 2013, Database System, Performance and Scalability, SQL Server 2012 Business Intelligence Editions.
16. Jacobson R. 2000, Microsoft SQL Server 2000 Analysis Services Step by Step. Microsoft Press.
17. Froese, R. and D. Pauly. Editors. 2009. FishBase. Electronic publication accessible at http://www.fishbase.org. Electronic version August 2014.
18. Moser, H. G. (Ed.). 1996. The early stages of the fishes in the California Current Region. *CalCOFI Atlas,* 33.
19. Hubbs, C. L. 1943. Terminology of early stages of fishes. *Copeia*, 260
20. Fisher, W., F. Krupp, W. Schneider, C. Sommer, K.E. Carpenter y V.H. Niem. 1995. Guía FAO para la identificación de especies para los fines de la pesca. Pacífico centro-oriental, Volumen II y III. Vertebrados. Roma. 1167 pp.
21. Ekman, S. 1953. Zoogeography of the sea. Sidwick and Jackson Ltd. London. 417 pp