ESCUELA POLITECNICA NACIONAL

FACULTAD DE INGENIERIA DE SISTEMAS

UNIDAD DE TITULACION

TÍTULO DEL TRABAJO DE TITULACIÓN:

DISEÑO E IMPLEMENTACION DE UN SISTEMA DE INTELIGENCIA DE NEGOCIO BI,

SOBRE LA BASE DE INFORMACION DE EDUCACION CONTINUA EN ECUADOR

PARA ANALIZAR LA OFERTA Y DEMANDA MEDIANTE INDICADORES ESTATICOS

Y DINAMICOS.

TRABAJO DE TITULACION PREVIO A LA OBTENCION DEL GRADO DE MAGISTER EN SISTEMAS DE INFORMACION MENCION INTELIGENCIA DE NEGOCIOS Y ANALITICA DE DATOS MASIVOS.

Autor: Ing Christian Vinicio Báez Espinosa

christian.baez@epn.edu.ec

Director: PhD. Julián Galindo

julian.galindo@epn.edu.ec

Quito, junio de 2023

APROBACION DEL DIRECTOR

Como director del trabajo de titulación Diseño e implementación de un sistema de inteligencia de negocios BI, sobre la base de información de educación continua en Ecuador para analizar la oferta y demanda mediante indicadores estáticos y dinámicos, desarrollado por Christian Vinicio Báez Espinosa, estudiante de la Maestría en Sistemas de Información mención Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral

Firmado electrónicamente por:
JULIAN ANDRES
GALINDO LOSADA

PhD. Julián Galindo

DIRECTOR.

DECLARACION DE AUTORIA.

Yo, Christian Vinicio Báez Espinosa, declaro bajo juramento que el trabajo descrito aquí es de mi autoría; el mismo no has sido previamente presentado en ningún grado o calificación profesional; y que las referencias bibliográficas que se incluyen en el documento fueron consultadas.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según la ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Firmado electrónicamente por:
CHRISTIAN VINICIO BAEZ ESPINOSA

Ing. Christian Vinicio Báez Espinosa.

DEDICATORIA

A mi círculo familiar compuesto por padres y hermanos.

A mis sobrinos quienes están comenzando su vida estudiantil.

Los amo.

Christian Vinicio Báez Espinosa

AGRADECIMIENTO.

A mi madre que supo guiarme para ser una persona productiva, a mis hermanos que apoyaron la decisión de continuar estudiando y a mi director de tesis Julián Galindo por su tiempo y dedicación en la guía de la tesis.

Christian Vinicio Báez Espinosa

ÍNDICE DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN.

El objetivo del presente trabajo es diseñar un Sistema de Inteligencia de negocios que nos permita conocer la oferta y la demanda de la educación continua en base a indicadores estáticos y dinámicos.

La metodología seleccionada para este proyecto fue HEFESTO 2.0; y en su fase de análisis de requerimientos se consiguió identificar las preguntas a responder las cuales sirvieron como base para diseñar las perspectivas y los indicadores dinámicos y estáticos asociados a la oferta y la demanda (sección 2.1.1).

En la fase de integración de datos se diseñaron los ETL que extraen información de documentos xls con la herramienta RapidMiner para su correspondiente carga en la base de datos del DW; los indicadores dinámicos se crearon mediante modelos de aprendizaje automático considerando criterios como el peso de las variables, velocidad de ejecución y el menor porcentaje de error (sección 2.4).

Para poder representar la información del DW se diseñó DashBoards con componentes visuales que permiten una clara visualización de la información de los indicadores dinámicos y estáticos por medio de la herramienta PowerBI (sección 2.5).

Se realizaron pruebas de usabilidad SUS (System Usability Scale) para evaluar el DashBoard; seis personas con un conocimiento intermedio en TICs realizaron la evaluación, el resultado se ubicó en un rango de usabilidad aceptable (sección 3.1).

Palabras clave: inteligencia de negocios, predicción, data warehouse, machine learning, educación continua, Hefesto.

ABSTRACT

The objective of this work is to design a Business Intelligence System that allows us to know the supply and demand of continuing education based on static and dynamic indicators.

The methodology selected for this project was HEFESTO 2.0; and in its requirements analysis phase, it was possible to identify the questions to be answered, which served as the basis for designing the perspectives and the dynamic and static indicators associated with supply and demand (section 2.1.1).

In the data integration phase, the ETLs that extract information from xls documents with the RapidMiner tool were designed for their corresponding loading in the DW database; the dynamic indicators were created by means of automatic learning models considering criteria such as the weight of the variables, speed of execution and the lowest percentage of error (section 2.4).

In order to represent the DW information, DashBoards were designed with visual components that allow a clear visualization of the information from the dynamic and static indicators through the PowerBI tool (section 2.5).

SUS (System Usability Scale) usability tests were performed to evaluate the DashBoard; six people with an intermediate knowledge of ICTs carried out the evaluation, the result was located in an acceptable usability range (section 3.1).

Keywords: business intelligence, prediction, data warehouse, machine learning, hefesto.

1. INTRODUCCION.

En el Ecuador la Educación Continua es una actividad docente universitaria donde la misión es vincularse con el medio mediante programas de capacitación que permitan educar de forma permanente a las personas.

Esta actividad es desarrollada exclusivamente por instituciones de educación superior las cuales deben estar legalmente registradas y autorizadas a funcionar por la Secretaria Nacional de Educación Superior SENECYT.

Actualmente tenemos acreditadas las siguientes instituciones entre públicas y privadas:

- 59 universidades.
- 110 institutos.

No se cuenta con información referente a precios de los cursos, modalidades de estudio, ofertas de capacitación y demás información relacionada a educación continua; de tal forma que fue necesario recopilar estos datos desde las páginas web de las principales universidades.

Hay que considerar que algunas universidades tienen sedes en varias ciudades del país, sin embargo, la oferta de capacitación es variada de tal forma que tratar de conocer cuáles son las necesidades de los potenciales clientes, que horarios, que precios, que cursos, que modalidad de estudio, etc. Son algunas de las preguntas que tratamos de averiguar mediante el siguiente proyecto de titulación.

La metodología seleccionada para el desarrollo del trabajo fue Hefesto 2.0, ya que se caracteriza por su facilidad de entendimiento y por tener fases donde se definen claramente los objetivos que se quieren implementar, se realizó un procedimiento de limpieza y calidad de datos a la información disponible de la oferta y la demanda utilizando diferentes herramientas como Python, Excel y NotePad++; en la fase de integración de datos utilizamos RapidMiner Studio y en la etapa de visualización utilizamos PowerBI.

La información recolectada para la oferta se basa en una investigación de mercado realizada por una empresa en el año 2021, está información fue proporcionada por el PHD. Julian Galindo, la misma está formada de:

- Costos de los cursos.
- Oferta de capacitación de las universidades.
- Duración de los cursos.
- Modalidades de estudio ofertadas.
- Horarios ofertados.

Encuesta relacionada a las preferencias de capacitación en cuanto al tiempo, horarios, precios, modalidad, curso de interés y áreas de estudio, esta encuesta se realizó a un número de 100 personas de distinto perfil profesional tanto de empresas públicas y privadas mediante un formulario de Google [20].

Al consolidar toda esta información en un base de datos, vamos a tener la capacidad de poder realizar un análisis que nos permita responder las preguntas planteadas y de esta manera conocer que sucede con la educación continua en el Ecuador.

1.1. Objetivo general.

Realizar un análisis de la oferta y demanda de la educación continua en Ecuador, mediante indicadores estáticos y dinámicos y con el diseño e implementación de un BI.

1.2. Objetivos específicos.

- Obtener los requerimientos del negocio.
- Obtener la información de demanda y oferta en educación continua.
- Diseñar el DW considerando el proceso ETL extracción, transformación y carga.
- Diseñar los tableros de información en base a los requerimientos del negocio.
- Definir indicadores dinámicos y estáticos asociados a la oferta y demanda.
- Implementar el diseño de la solución en un ambiente de pruebas.

1.3. Alcance.

El alcance del presente proyecto implica el diseño de un sistema de Inteligencia de Negocios (BI), en base a la información recolectada, no está contemplado la implementación de este dentro de un ambiente de producción.

1.4. Marco Teórico.

Dentro de esta sección se realiza una descripción de las diferentes temáticas tratadas en el desarrollo del proyecto, como es: Dato, Información, Metodología Hefesto, Inteligencia de Negocios (BI), DW, en base a las definiciones encontradas de varios autores.

1.4.1. Dato.

Es la mínima unidad de información que puede representar un valor cualitativo o cuantitativo de una entidad física o imaginaria, por ejemplo: peso, color, activo, etc.; El dato por sí solo carece de valor y es necesario tratarlo y procesarlo para poder obtener información relevante.

1.4.2. Ciclo de vida del Dato.

Los datos son en la actualidad el activo más importante que tiene una empresa; de tal manera que conocer cuál es su ciclo de vida es de vital importancia para obtener buenos resultados al momento de procesar la información. El ciclo de vida del dato DML se divide en cinco

etapas, este debe pasar por cada una de ellas para poder cumplir su ciclo, tal como lo indica la **Figura 1**.

Figura 1. Ciclo de vida del dato.

1.4.2.1. Creación.

La primera fase es donde el dato se crea, en esta etapa es donde definimos como se crean, su origen, su arquitectura, que tipo de dato se va a elaborar y como se lo va a conservar.

1.4.2.2. Almacenamiento.

En esta etapa se enfoca en la forma en que vamos a almacenar los datos según sea su tipo, arquitectura, origen y de acuerdo a sus características se almacenaran en diferentes secciones.

Dependiendo de si son datos Estructurados su almacenamiento se realizará en bases de datos relacionales del tipo SQL, donde gracias a las tablas e índices podemos ubicar la información que necesitemos.

En el caso de tener datos No Estructurados su almacenamiento se realizará en bases de datos del tipo NoSQL como por ejemplo MongoDB, este tipo de almacenamiento están orientados ambientes donde se requiere guardar datos masivos a gran escala.

1.4.2.3. Utilización.

Es una de las etapas más críticas del ciclo de vida, pues es aquí donde se hace uso de la información para poder generar conocimiento y conocer de lo que está sucediendo mediante la generación de gráficas, cuadros de mando y para en lo posterior generar analítica. En esta etapa el dato se utiliza para la toma de decisiones.

1.4.2.4. Archivo del dato.

Después de haber transcurrido algún tiempo desde la creación y utilización del dato en algunas ocasiones estos ya no son útiles para el trabajo diario, sin embargo, es importante mantener un respaldo para en algún momento volver a consultarlo y utilizarlo cuando se lo requiera. Una estrategia de DML bien definida indica claramente donde, cuando y durante que tiempo se archivan los datos.

1.4.2.5. Eliminación del dato.

Es la etapa final del ciclo de vida del dato; las empresas eliminan los datos innecesarios para poder tener espacio de almacenamiento para nuevos datos activos, la organización elimina la información que ya no tiene importancia para la misma, en base a políticas establecidas en el gobierno de datos.

1.4.3. Inteligencia de negocios.

La Inteligencia de negocios la podemos definir como un conjunto amplio de aplicaciones, tecnologías y procesos para recolectar, almacenar, acceder y analizar información que les permita a las empresas poder tomar mejores decisiones [1]. Existen elementos que están muy correlacionados con BI, datos, información y conocimiento; los datos se transforman en información y estos a su vez en conocimiento. El dato por sí solo carece de valor y no aporta con algún significado, mientras que la información es el resultado de un procesamiento de datos y que luego de este proceso aportan con relevancia, propósito y contexto, finalmente el conocimiento es una mezcla de experiencia, valores y know-how que sirve como guía para la incorporación de nuevas experiencias y es útil para la acción [2].

1.4.4. Base de Datos.

Las bases de datos aparecieron en la medida que las empresas empezaron a darse cuenta del enorme potencial y ventaja competitiva que podían obtener al tener la capacidad de gestionar su información. Estas permiten almacenar grandes volúmenes de información, en la actualidad cualquier negocio sin importar su tamaño organiza su información con la ayuda de una Base de Datos.

1.4.4.1. Bases de datos relacionales.

Una base de datos relacional almacena la información en tablas, las cuales están organizadas en filas y columnas, las tablas tienen una relación lógica de acuerdo con la información que va a almacenar.

Estas bases de datos son las adecuadas cuando la información que deseamos almacenar es del tipo estructurada, tal como lo indica la **Figura. 2**.

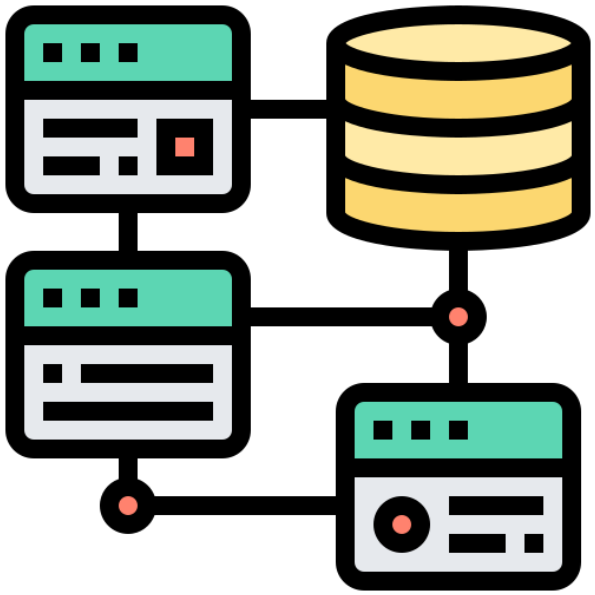


Figura 2. Base de datos relacional.

1.4.4.2. Bases de datos NoSQL.

En los últimos años, las nuevas características de los datos llevaron al desarrollo de nuevos sistemas de gestión de bases de datos denominados NoSQL[14].

Actualmente las empresas deben tener la capacidad de gestionar un flujo de información enorme; si consideramos que estos datos provienen de distintas fuentes, no son estructurados y su tamaño es variable, entonces almacenar esta información en una base de datos relacional no es la mejor opción. Ejemplos de este tipo de información lo tenemos en los datos que

generan los sismógrafos de una estación sísmica o la información que genera algún dispositivo de IoT.

Para estos escenarios se necesita una Base de datos del tipo NoSQL, en donde la velocidad de almacenamiento y recuperación de la información es de vital importancia debido a la cantidad de información que se procesa por segundo. En el mercado existen diferentes tipos de bases de datos [2] que se agrupan en base a cuatro características principales. En la **Figura 3** se muestra estas características.

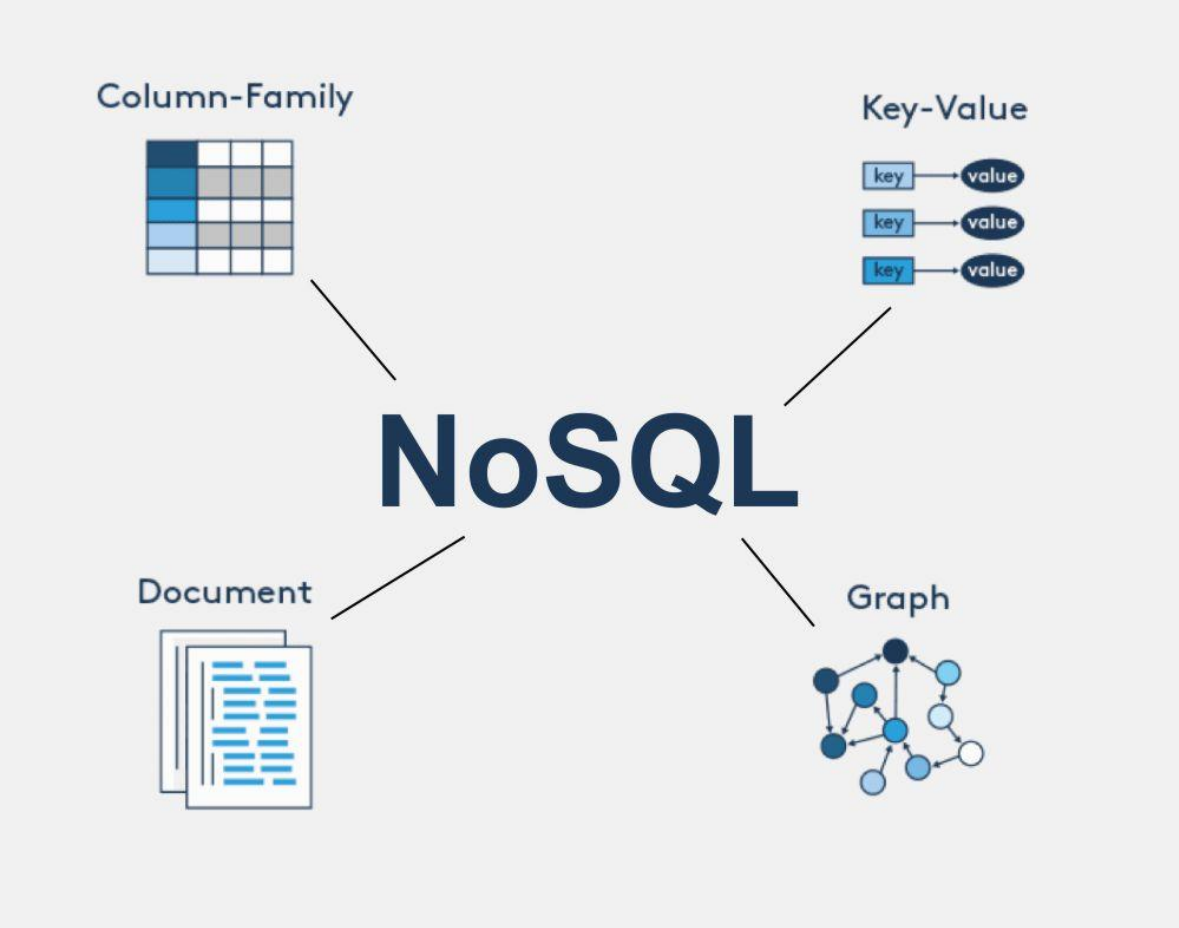


Figura 3. Tipos de bases de datos NoSQL.

1.4.5. Integración de Datos.

La Integración de Datos, es el proceso de combinar información que provienen de distintas fuentes como por ejemplo hojas de Excel, archivos de texto, otras bases de datos, etc. Al combinar estos datos en una única fuente, tenemos la posibilidad de contar con una unidad de información sobre la cual podemos obtener información para los distintos departamentos de la empresa.

Para manejar estos desafíos, la integración de datos es clave, especialmente donde los datos vienen en formatos estructurados y no estructurados y deben integrarse a partir de fuentes dispares almacenadas en sistemas gestionadas por diferentes departamentos [15].

Si la cantidad de datos no es grande, la integración la podemos realizar de forma manual, pero el momento que los datos crecen este proceso se vuelve complejo y es recomendable realizarlo de forma automatizada mediante el uso de aplicaciones de software creadas para este fin [3].

La Integración de datos debe realizarse de forma muy meticulosa, pues es un punto clave para crear un DW, es el repositorio desde donde se extrae información para la Inteligencia de Negocios que ayuda en la toma de decisiones, la **Figura 4** indica el ciclo que sigue la Integración de Datos.

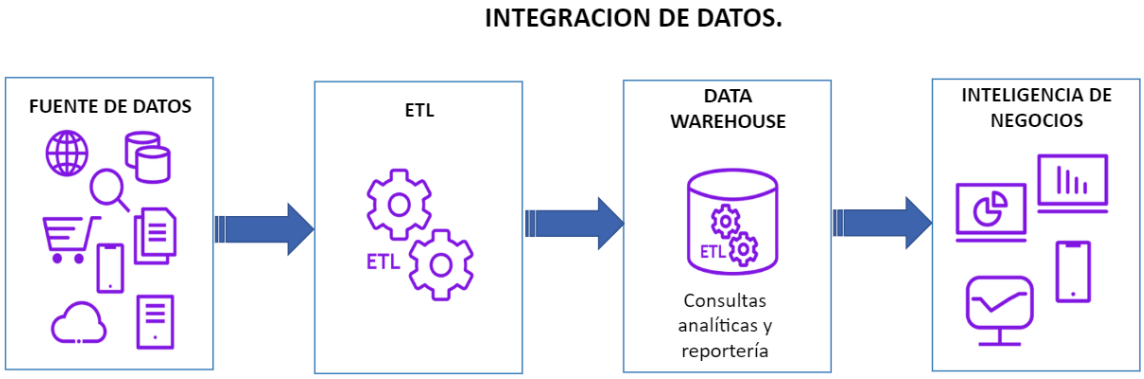


Figura 4. Integración de Datos.

1.4.5.1. ETL (Extracción, Transformación y Carga).

El proceso ETL es una parte de la integración de datos cuya finalidad es recuperar, organizar y gestionar la información para su posterior carga en un DW [13].

El proceso ETL consta de tres fases, la primera es la **Extracción** de la información de las distintas fuentes de datos sin importar si la data es estructurada o no y la información puede venir desde distintas fuentes, la misma se alojará en un área de staging en espera de la siguiente etapa que es la **Transformación** en la cual se logra tener información de calidad, para esto aplicamos técnicas como filtrado, limpieza, eliminación de duplicados, validaciones y demás actividades que nos permiten contar con data uniforme y de alta importancia, podemos decir que esta etapa es una de las más críticas.

La siguiente y última etapa es la **Carga** que consiste en mover los datos desde el área de staging a la base de datos del DW, por lo general se realiza una carga inicial y luego se van realizando cargas de información en distintos periodos con la finalidad de actualizar, eliminar o incrementar información al DW. Este procedimiento las empresas lo realizan de manera automática en horarios donde la transaccionalidad de las operaciones de la base de datos es baja.

1.4.5.2. Data WareHouse

Podemos definirlo como un almacén de datos donde se recopila y almacena información que provienen desde distintas fuentes, este consolidado de información será utilizado por el personal de la empresa que necesita comprender y utilizar estos datos para la toma de decisiones.

Para un DW lo más importante es el soporte a la toma de decisiones y en menor medida el soporte a la transaccionalidad [4]; para la elaboración de un DW debemos seguir los pasos que se muestran en la **Figura 5**.

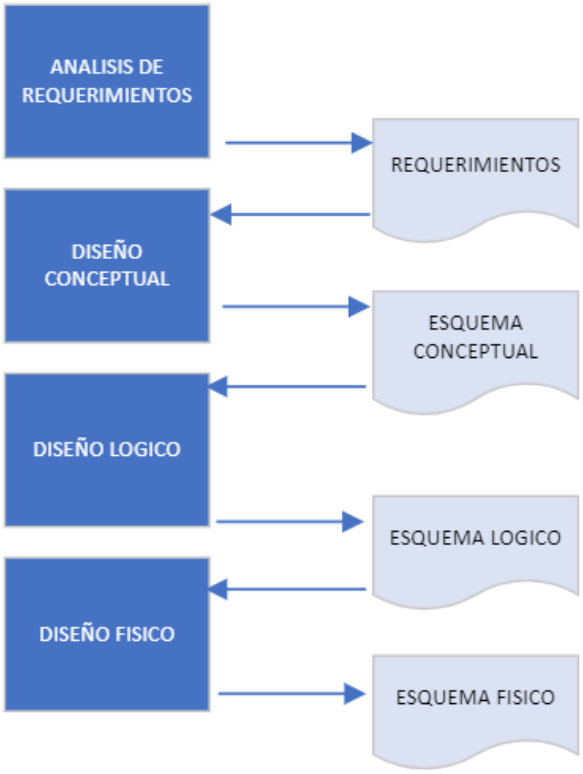


Figura 5 Fases de elaboración del DW.

Análisis de Requerimientos.

En esta etapa se recopila todas las necesidades, documentación e información necesaria para el desarrollo del proyecto, se considera de gran importancia y relevancia pues con un buen análisis se garantiza el éxito del proyecto.

Diseño Conceptual.

El modelo conceptual nos da un alto nivel de abstracción para describir los procesos de almacenamiento y arquitectura que involucra el diseño de la Base de Datos para el DW.

En esta etapa hacemos uso de notación gráfica que expresan las ideas para que sean comprendidas por los diseñadores y por los usuarios [4].

Diseño Lógico.

Una vez que se ha completado el Diseño Conceptual, se procede a elaborar el diseño lógico que consiste en transformar el esquema conceptual en un esquema lógico optimizado para un motor de almacenamiento específico [4].

Diseño Físico.

En la última etapa de construcción del DW, procedemos a la creación del esquema de la BD sobre un motor de Base de Datos seleccionado, será aquí donde podremos almacenar y gestionar la información para su correspondiente explotación aprovechamiento de consultas de analítica y reportería.

1.4.5.3. Business Intelligence.

Existen varias definiciones para BI, todas orientadas al manejo y explotación de la información; de todos estos conceptos encontrados, podemos definir al BI como un término general que comúnmente es utilizado para describir las tecnologías, aplicaciones y procesos para recopilar, almacenar, acceder y analizar datos para ayudar a los usuarios a tomar mejores decisiones [5].

Si consideramos al BI como un Sistema de Información tenemos a nuestra disposición un conjunto de fuentes de datos externas e internas que se cargan en un DW, desde donde se van a ejecutar consultas analíticas que producen resultados para el usuario final, tal como lo indica la **Figura 6.**

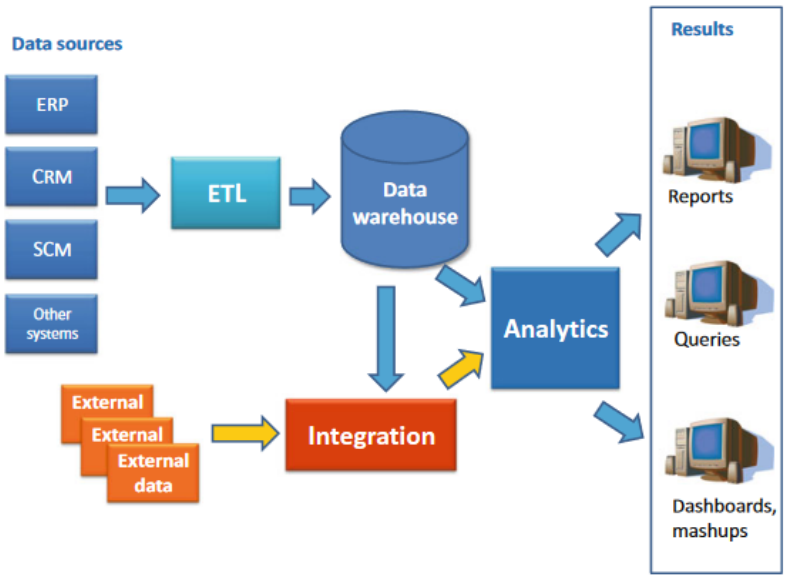


Figura 6. BI como Sistema de información.

El crecimiento de información que se genera en la actualidad obliga a las empresas de cualquier índole a gestionar grandes volúmenes de datos de una manera rápida para poder responder de manera ágil a las expectativas de sus clientes, de tal manera que un BI dentro de una empresa sin importar su tamaño se vuelve una herramienta tecnológica clave.

En el mercado tenemos variedad de competidores de herramientas de software para BI, tal como lo indica el cuadrante de Gartner en la **Figura 7**, lo que nos indica que las empresas siguen beneficiándose del uso del Business Intelligence.

Figura 7. Cuadrante de Gartner sobre aplicaciones de software para BI.

1.4.6. Indicadores.

Un indicador representa una unidad de medida o magnitud y que a través del análisis de este podemos tomar acciones en función de los valores que tenga.

1.4.6.1. Indicadores estáticos.

Este tipo de indicadores hacen referencia a un análisis descriptivo, tratan de responder a preguntas como que sucedió o que está pasando [10].

1.4.6.2. Indicadores dinámicos.

Se enfoca en el resultado de un análisis predictivo, trata de responder la pregunta ¿Qué es probable que suceda? [10].

1.4.7. Python.

Es un poderoso lenguaje de programación que es muy fácil de aprender, es muy eficiente en el manejo de estructuras de datos, lo que le vuelve un lenguaje apto el campo de la ciencia de datos [11].

1.5. RapidMiner.

Es una plataforma para ciencia de datos que nació a partir de un proyecto en la Universidad de Dormunt en el año 2001; cuenta con varios módulos que nos permiten crear modelos de predicción, ETLS, ingeniería de datos, etc. Se puede descargar una versión académica de forma gratuita para su uso [12].

2. METODOLOGIA.

En este capítulo se realiza una descripción de la metodología utilizada para llevar a cabo el desarrollo del DW, considerando que tenemos algunas alternativas entre las cuales podemos mencionar a Kimball, CRISP-DM, Hefesto. Hemos seleccionado a Hefesto por ser una metodología que está en constante evolución y toma en cuenta todas las aportaciones de la comunidad que la utiliza [6].

Para poder llegar a la culminación exitosa de un proyecto de DW es muy necesario que las fases de la metodología no seas muy extensas y que tampoco compliquen su desarrollo [6]; en base a estas consideraciones HEFESTO requiere de cuatro pasos a desarrollar [7], los cuales los resumimos en la **Figura 8**.

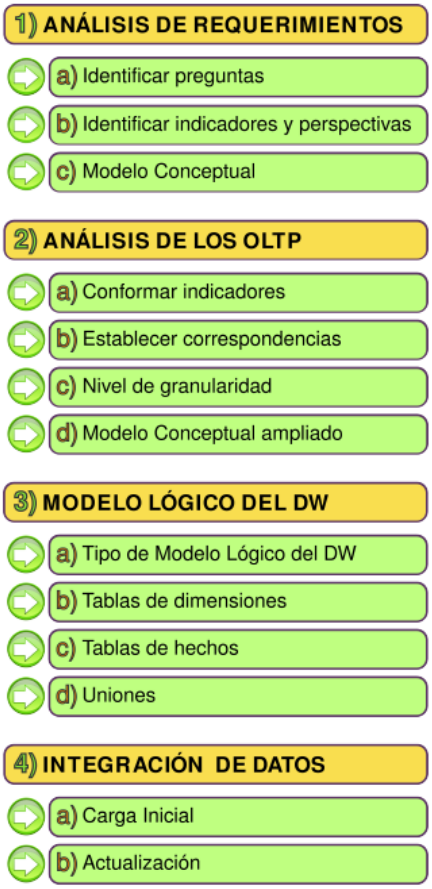


Figura 8. Etapas de la metodología HEFESTO.

Entre las características principales podemos mencionar las siguientes [7]:

- En cada una de las fases se puede distinguir fácilmente los objetivos que se persiguen, así como los resultados esperados; adicionalmente son de fácil comprensión.
- La estructura del DW es de fácil y rápida adaptación, esto se debe a que fue construida en base a los requerimientos de los usuarios.
- La resistencia que presentan los usuarios finales al cambio se ve reducido gracias a que en cada etapa se los considera para determinar el comportamiento y las funciones que se incorporan al diseño del DW.
- Los modelos conceptuales y lógicos que se implementan son de sencilla comprensión y análisis.
- El tipo de ciclo de vida que contenga a la metodología marcan independencia el uno del otro.
- Las herramientas que se utilicen para la construcción del DW son independientes de la metodología.
- La metodología es independiente de las estructuras físicas y su distribución, que contengan el DW.
- Los resultados obtenidos al finalizar una fase se convierten en un nuevo punto inicial para el siguiente paso.
- Se aplica tanto para el DW como para los Data Mart.

En las siguientes líneas vamos a realizar una descripción rápida de las cuatro etapas de HEFESTO aplicadas al desarrollo del proyecto propuesto.

- Análisis de Requerimientos: En esta etapa procedimos a la recopilación de la información del negocio en base a las necesidades específicas de los usuarios, en nuestro caso se utilizó una encuesta para medir las necesidades de capacitación y para

el caso de la oferta se utilizó información de las instituciones de educación superior con data referente a los cursos de educación continua que ofrecen.

- Análisis de los OLTP: En esta fase podemos ya generar el modelo conceptual, definir los indicadores con sus respectivas correspondencias y el nivel de granularidad que se requiera.
- Modelo Lógico del DW: En esta etapa conseguimos la creación del modelo lógico con sus correspondientes tablas de hechos, dimensiones y sus correspondientes uniones.
- Integración de datos: En la etapa final procedemos a la creación de los ETL, para proceder con la carga inicial de los datos previa limpieza de estos y luego continuar con la actualización de la información.

2.1. Análisis de requerimientos.

La educación continua es una actividad de capacitación exclusiva de los institutos, universidades y escuelas politécnicas del Ecuador, en el país tenemos 59 universidades o escuelas politécnicas legalmente acreditadas, al igual que 82 institutos [8] tal como lo indica la **Figura 9**.

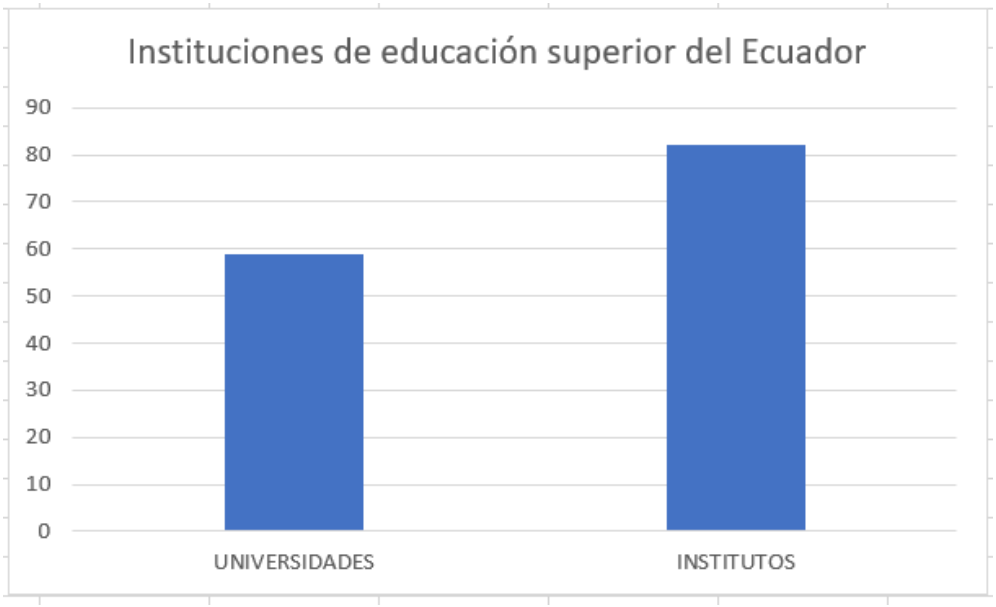


Figura 9. Instituciones de educación superior acreditadas

Sin embargo, no todas las instituciones ofertan el servicio de educación continua, por tal motivo para obtener información de la oferta hemos seleccionado a las principales universidades de Quito, Guayaquil y Cuenca.

De esta segmentación encontramos la siguiente información que la resumimos en la **Tabla 1**.

Tabla 1. Oferta de capacitación.

SEGMENTACION	CANTIDAD
Cursos	502
Horarios	135
Modalidades estudio	9

De acuerdo con la tabla la oferta de cursos es amplia al igual que los horarios y en menor cantidad las modalidades de estudio.

Para conocer la demanda de capacitación se creó una encuesta dirigida a profesionales de distintas áreas, las preguntas las detallamos en la **Tabla 2**.

Tabla 2. Preguntas para medir la demanda

Nro.	**Pregunta**
1	Cuál es su género
2	Cuál es su edad
3	En qué sector trabaja usted
4	En que institución labora
5	En qué departamento o área de la institución desarrolla sus actividades
6	En orden jerárquico, usted pertenece a.
7	En qué ciudad trabaja o desarrolla sus actividades.
8	Qué nivel de formación tiene
9	Qué área de estudios o especialización tiene
10	Qué título profesional tiene
11	Tiene interés en capacitarse

12	Tiene interés en tomar un programa de Educación Continua
13	En su criterio cuál es el valor razonable de un programa de Educación Continua en USD$
14	Qué modalidad de estudio es la más apropiada para usted
15	Que días son más cómodos para usted tomar este tipo de programas de formación de educación continua.
16	Qué horario en los días seleccionados es más cómodo para usted tomar este tipo de programas.
17	En su criterio, cual es el número de horas máximo que debería tener un curso de educación continua
18	Qué tan importante es la educación continua en sus actividades laborales.
19	Considera que la educación continua es útil para su organización
20	Considera que la educación continua es importante para mejorar la provisión de servicios públicos.
21	En qué área usted estaría interesado en tomar un programa de educación continua.
22	Seleccione el programa de educación continua en temas administrativos que sería de su interés.
23	Seleccione el programa de educación continua en temas tecnológicos que sería de su interés.
24	Seleccione el programa de educación continua en temas educativos que sería de su interés.
25	Seleccione el programa de educación continua en temas técnicos que sería de su interés.
26	Seleccione el programa de idiomas que sería de su interés.
27	Otras temáticas de interés
28	Su correo electrónico
29	Su número de teléfono

Las preguntas fueron aplicadas a una muestra de 100 personas de distintos perfiles profesionales, edades y de empresas tanto públicas como privadas.

Tomando como base la información de la oferta y demanda se procedió a identificar las preguntas.

2.1.1. Identificación de preguntas.

Esta etapa es muy crucial para el éxito del proyecto, por tal motivo es necesario contar con la participación de las personas que intervienen en el proceso.

En nuestro caso las preguntas que logramos identificar se representan en la **Tabla 3**.

Tabla 3. Preguntas de negocio.

ORD	**PREGUNTA**
1	Se desea conocer la institución educativa con la mayor cantidad de cursos ofertados.
2.	Se desea saber el mayor número de ofertas de las modalidades de estudio.
3	Se desea conocer el promedio en horas de duración de los cursos ofertados según su modalidad.
4.	Se desea saber el costo promedio de la oferta de los cursos ofertados.
5.	El área de formación con mayor oferta de capacitación
6.	Se desea conocer el mayor número de personas que demandan capacitarse por curso.
7.	Se desea saber el horario de capacitación con mayor demanda.
8.	Se desea conocer el promedio de edad de las personas que demandan capacitarse.
9.	Se desea predecir para el año 2023, la demanda de personas interesadas para los cursos.
10.	Se desea predecir el promedio de los precios de los cursos ofertados para el año 2023.
11.	Se desea predecir el promedio de la duración en horas de los cursos ofertados para el 2023.

Ahora que se han identificado las preguntas, el siguiente paso consiste en identificar los indicadores y perspectivas.

2.1.2. Identificar indicadores y perspectivas.

En base a las preguntas identificadas procedemos a identificar los indicadores y sus correspondientes perspectivas, bajo esta premisa podemos decir que los ***indicadores*** representan valores numéricos y lo que queremos analizar, por ejemplo, cálculos, cantidades, sumatorias, etc.

Por el otro lado tenemos a las ***perspectivas*** que son las entidades desde donde vamos a obtener las respuestas a los indicadores, esto tiene la finalidad de responder a las preguntas planteadas, como ejemplo de perspectivas podemos mencionar a: productos, clientes, sucursales, proveedores, etc.

Para el desarrollo de nuestro proyecto hemos logrado identificar las siguientes perspectivas, en base a la información almacenada en los archivos csv tanto de la oferta como de la demanda, las perspectivas encontradas están representadas en la **Tabla 4.**

Tabla 4. Perspectivas encontradas.

PERSPECTIVA
Curso
Modalidad
Horario
Encuesta
Área Formación
Universidad
Tipo Oferta
Fecha
Días
Sector Trabajo
Institución

Ciudad Trabajo
Nivel Jerárquico
Departamento Trabajo
Área Interés
Nivel Formación
Género
Interesado

Los indicadores que se obtuvieron a partir de las preguntas de negocio están representados en la **Tabla 5**.

Tabla 5. Lista de indicadores.

ORD	**INDICADOR**	**TIPO**
1	La institución educativa con la mayor cantidad de cursos ofertados.	Estático
2	El mayor número de ofertas de la modalidad de estudios.	Estático
3	El promedio de la duración en horas de los cursos ofertados según su modalidad.	Estático
4	El valor promedio del costo de los cursos ofertados.	Estático
5	El mayor número de áreas de estudios con ofertas de cursos.	Estático
6	El mayor número de personas que demandan capacitarse por curso.	Estático
7	El mayor número de demanda del horario de estudios.	Estático
8	El promedio de edad de las personas que demandan capacitarse	Estático
9	Predicción del número de interesados de los cursos demandados para el 2023	Dinámico
10	Predicción del promedio del precio de los cursos ofertados para el 2023.	Dinámico
11	Predicción del promedio de la duración en horas de los cursos ofertados para el 2023.	Dinámico

2.1.3. Modelo conceptual.

El modelo conceptual es la representación a alto nivel de la base de datos que va a gestionar las perspectivas y sus indicadores; la información se la representa mediante Objetos, Relaciones y Atributos.

El objetivo del modelo conceptual es explicar mediante una descripción gráfica el alcance del proyecto a las personas interesadas en la implementación del BI, comprender que es lo que se va a obtener, cuales variables se están utilizando para el análisis y como están relacionados entre sí; en la **Figura 10** podemos visualizar los elementos que intervienen en la elaboración del modelo.

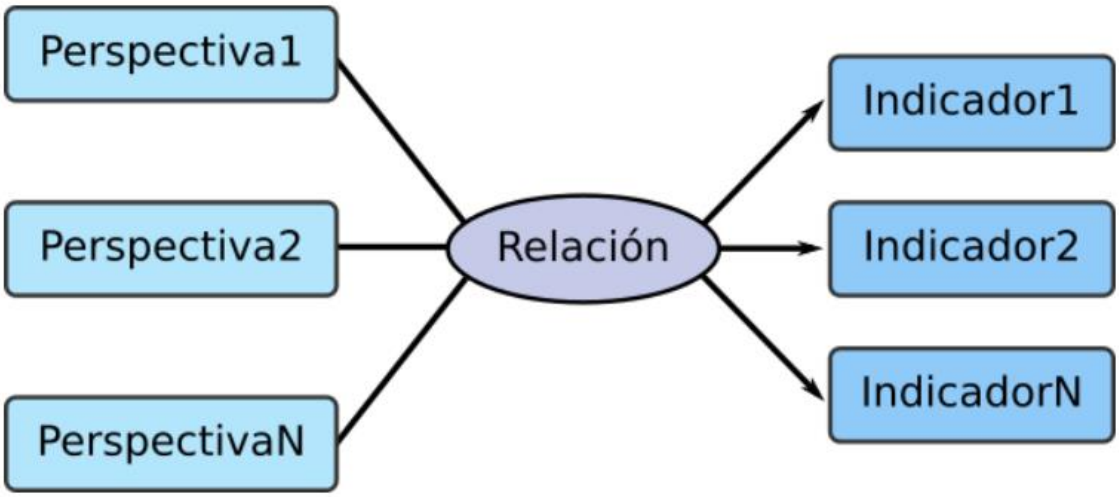


Figura 10. Representación gráfica del modelo conceptual.

El mencionado modelo está formado por las *Perspectivas* y los *Indicadores*, los cuales se interconectan mediante una *Relación* que es el área de estudio sobre la cual queremos desarrollar el BI.

De acuerdo con la información disponible para la elaboración del DW y en base a los indicadores encontrados en la tabla 5; la **Figura 11** representa el modelo conceptual acorde a las perspectivas e indicadores.

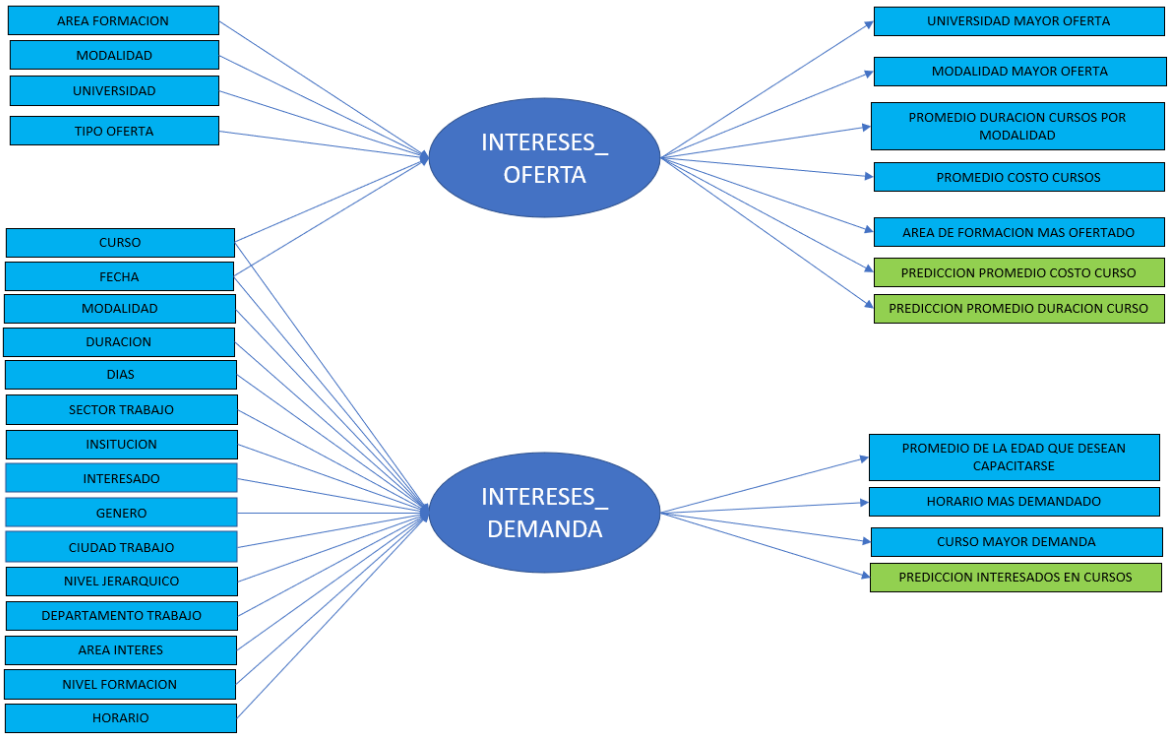


Figura 11. Modelo conceptual.

2.2. Análisis de los OLTP.

El objetivo de este paso es analizar las fuentes OLTP con la finalidad de definir como serán calculados los indicadores y así lograr definir las relaciones entre el modelo conceptual que se creó en el punto anterior con las fuentes de datos disponibles. A continuación, debemos definir los campos que deben estar presentes en las perspectivas para lograr ampliar el modelo conceptual usando la información obtenida.

2.2.1. Conformación de indicadores.

Para la conformación de indicadores es necesario especificar lo siguiente.

- El o los hechos de los que se componen con su respectiva fórmula de cálculo.
- La función de sumarización para su cálculo, por ejemplo: AVG, SUM, COUNT, etc.

Listado de indicadores.

- Universidad con mayor oferta.
    - Hechos: Universidad mayor oferta.
    - Función de sumarización: MAX.
- Modalidad mayor oferta.
    - Modalidad mayor oferta.
    - Función de sumarización: MAX.
- Promedio duración curso ofertado por modalidad.
    - Promedio duración curso por modalidad.
    - Función de agregación: AVG.
- Promedio costo curso por modalidad.
    - Promedio costo curso por modalidad.
    - Función de agregación: AVG.
- Área de formación con mayor oferta.
    - Área formación con mayor oferta.
    - Función de agregación: MAX.
- Predicción del promedio de costo del curso ofertado.
- Predicción del promedio de la duración del curso ofertado.
- Promedio edad interesados en capacitarse.
    - Promedio edad interesado curso.
    - Función de agregación: AVG.
- Horario capacitación con mayor demanda
    - Horario con mayor demanda.

    - Función agregación: MAX.

- Curso capacitación con mayor demanda.
    - Curso con mayor demanda.
    - Función agregación: MAX.
- Predicción número de interesados en curso demandados.

2.2.2. Establecer correspondencias.

Este paso consiste en revisar los DataSource para verificar que estos contengan los datos que necesitamos usar en el DW, una vez revisados debemos establecer la correspondencia con los elementos definidos en el Modelo Conceptual.

Lo que vamos a lograr es tener una visión más amplia de como los indicadores definidos tienen sus correspondencias con las fuentes de datos disponibles, así tenemos la capacidad de garantizar el éxito de nuestro DW.

La información necesaria para la oferta se encuentra en un archivo xls (costos Oferta de Educación continua.xlsx), organizada en columnas, donde cada columna contiene información de las perspectivas del modelo conceptual.

La siguiente gráfica representada en la **Figura 12**, nos indica la columna del archivo xls y su correspondiente asociación con la perspectiva.

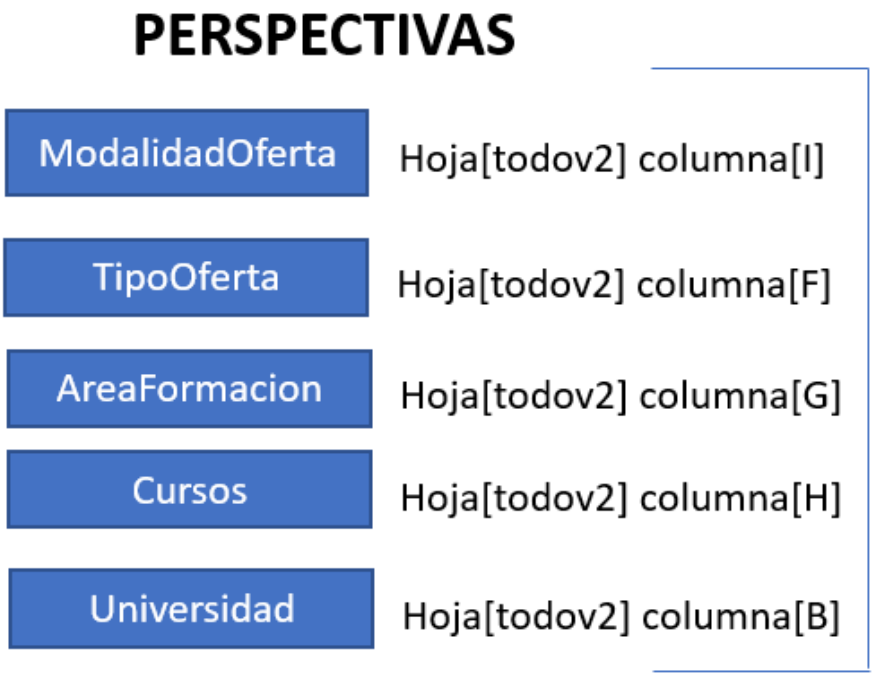


Figura 12. Correspondencia con los DataSources de las perspectivas de la oferta.

Para el caso de la demanda, tenemos un archivo csv (Encuesta de capacitación.csv) que almacena información de la encuesta que nos permitió conocer las necesidades de capacitación continua.

La información de la demanda al igual que la oferta se encuentra organizada mediante columnas, por tal motivo procedemos a ubicar las columnas que representan las perspectivas del modelo conceptual, la **Figura 13** representa de forma gráfica esta información.

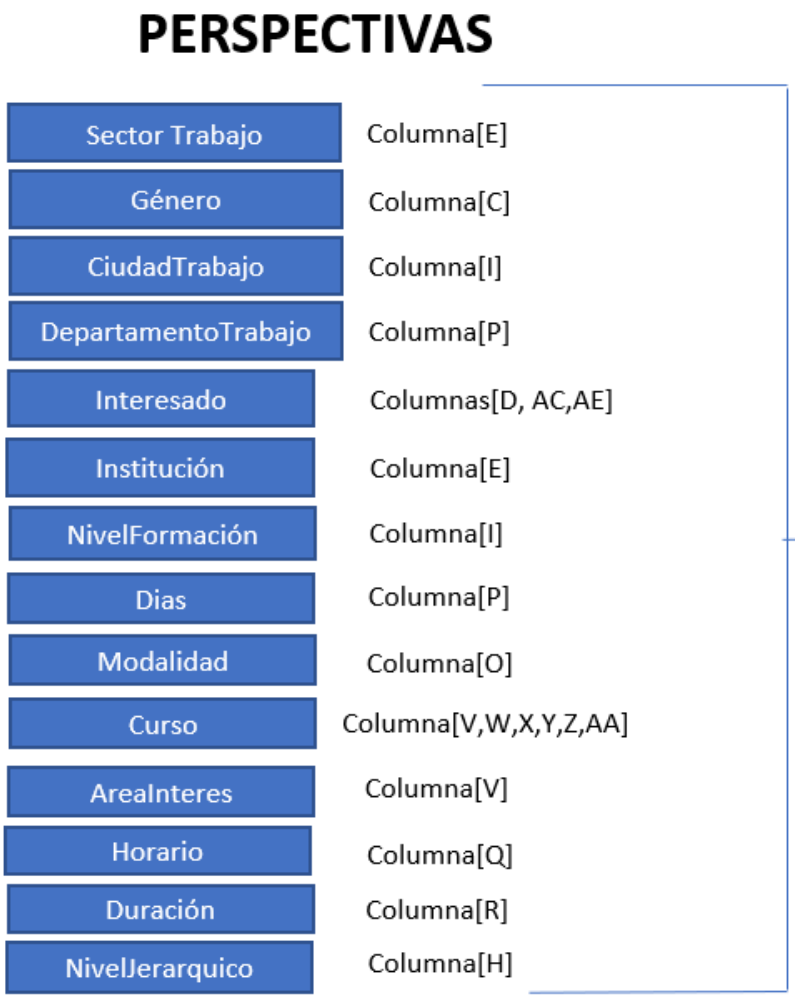


Figura 13. Correspondencia con los DataSources de las perspectivas demanda.

Luego de realizar el mapeo correspondiente hemos obtenido los siguientes indicadores.

- El indicador *Curso_Mayor_Oferta*, se relaciona con el campo curso_id de la perspectiva Curso y se calcula mediante la formula MAX(curso_id).
- El indicador *Modalidad_Mayor_Oferta*, se relaciona con el campo modo_id de la perspectiva modalidad y se calcula mediante la formula MAX(modo_id).
- El indicador *Promedio_Duración_Curso_Ofertado_Por_Modalidad*, se relaciona con el campo *duración_programa* y se calcula mediante la formula AVG(duración_programa).

- El indicador *Promedio_Costo_Curso_Por_Modalidad*, se relaciona con el campo *costo_programa* y se calcula mediante la formula AVG(costo_programa).
- El indicador *Area_Formación_Mas_Ofertado*, se relaciona con el campo *arfo_id* y se calcula mediante la formula MAX(arfo_id).
- El indicador dinámico *predicción_costo_curso*, se relaciona con el campo *costo_programa* y se calcula mediante un algoritmo de machine del tipo Gradient_Boosted_Trees.
- El indicador dinámico *predicción_duracion_curso*, se relaciona con el campo *duración_programa* y se calcula mediante un algoritmo de machine del tipo Gradient_Boosted_Trees.
- El indicador *Promedio_Edad_Demandan_Capacitarse*, se relaciona con el campo *edad* de la perspectiva *Interesado* y se calcula mediante la formula AVG(edad).
- El indicador *Horario_Mas_Demandado*, se relaciona con el campo *hora_id* de la perspectiva *Horario* y se calcula mediante la formula MAX(hora_id).
- El indicador *Curso_Mayor_Demanda*, se relaciona con el campo *curs_id* de la perspectiva *Curso* y se calcula mediante la formula MAX(curs_id).
- El indicado dinámico *interesados_prediccion* se relaciona con el campo *interesados_prediccion* y se calcula utilizando un algoritmo de machine learning.

2.2.3. Nivel de granularidad.

El nivel de granularidad nos permite seleccionar que campos de las perspectivas encontradas nos sirven para poder interpretar la información de los indicadores, en este sentido es de suma importancia el poder conocer el significado de cada campo.

La información encontrada para cada perspectiva se detalla a continuación.

La perspectiva **Area_Formación** tiene los siguientes campos listados en la **Tabla 6.**

Tabla 6. Campos de la perspectiva Área Formación

Campo	Descripción
Valor_área_de_formación	Nombre del área de formación del curso

La perspectiva **Modalidad** tiene los siguientes campos que los detallamos a continuación en la **Tabla 7.**

Tabla 7. Campos de la perspectiva Modalidad.

Campo	Descripción
Valor_modalidad_oferta	El nombre de la modalidad de la oferta del curso

La perspectiva **Universidad** tiene los siguientes campos que los detallamos a continuación en la **Tabla 8.**

Tabla 8. Campos de la perspectiva Universidad.

Campo	Descripción
Valor_universidad	El nombre de la universidad que imparte el curso

La perspectiva **Tipo de Oferta** tiene los siguientes campos, los cuales los detallamos en la **Tabla 9.**

Tabla 9. Campos de la perspectiva Tipo de oferta.

Campo	Descripción
Valor_tipo_oferta	El nombre del tipo de oferta del curso

La perspectiva **Curso** tiene los siguientes campos, los cuales los detallamos en la **Tabla 10.**

Tabla 10. Campos de la perspectiva Curso.

Campo	Descripción
Valor_curso	El nombre del curso.

La perspectiva **Fecha** tiene los siguientes campos, los cuales los detallamos en la **Tabla 11.**

Tabla 11. Campos de la perspectiva Fecha.

Campo	Descripción
Año	Año de la fecha
Semestre uno	Número de semestre
Semestre dos	Número de semestre
Fecha	Fecha compuesta por el año y el semestre

La perspectiva **Modalidad demanda** tiene los siguientes campos, los cuales los detallamos en la **Tabla 12.**

Tabla 12. Campos de la perspectiva Modalidad oferta.

Campo	Descripción
Valor modalidad	Nombre de la modalidad de la demanda del curso.

La perspectiva **Duración** tiene los siguientes campos, los cuales los detallamos en la **Tabla 13.**

Tabla 13. Campos de la perspectiva Duración.

Campo	Descripción
Valor duración curso	Valor de la duración del curso.

La perspectiva **Días** tiene los siguientes campos, los cuales los detallamos en la **Tabla 14.**

Tabla 14. Campos de la perspectiva Días.

Campo	Descripción
Valor días	Valor del día para seguir el curso.

La perspectiva **Sector trabajo** tiene los siguientes campos, los cuales los detallamos en la **Tabla 15.**

Tabla 15. Campos de la perspectiva Sector trabajo.

Campo	Descripción
Valor sector trabajo	Valor del sector de trabajo del interesado en seguir el curso.

La perspectiva **Institución** tiene los siguientes campos, los cuales los detallamos en la **Tabla 16.**

Tabla 16. Campos de la perspectiva Institución.

Campo	Descripción
Valor institución	Nombre de la institución de trabajo del interesado en seguir el curso.

La perspectiva **Género** tiene los siguientes campos, los cuales los detallamos en la **Tabla 17.**

Tabla 17. Campos de la perspectiva Género.

Campo	Descripción
Valor género	Valor del género del interesado en seguir el curso.

La perspectiva **Ciudad Trabajo** tiene los siguientes campos, los cuales los detallamos en la **Tabla 18.**

Tabla 18. Campos de la perspectiva Ciudad Trabajo.

Campo	Descripción
Valor ciudad trabajo	Valor de la ciudad de trabajo del interesado en seguir el curso.

La perspectiva **Nivel jerárquico** tiene los siguientes campos, los cuales los detallamos en la **Tabla 19.**

Tabla 19. Campos de la perspectiva Nivel jerárquico.

Campo	Descripción
Valor nivel jerárquico	Valor del nivel jerárquico del interesado en seguir el curso.

La perspectiva **Departamento de trabajo** tiene los siguientes campos, los cuales los detallamos en la **Tabla 20.**

Tabla 20. Campos de la perspectiva Departamento de trabajo.

Campo	Descripción
Valor departamento de trabajo	Valor departamento de trabajo del interesado en seguir el curso.

La perspectiva **Área interés** tiene los siguientes campos, los cuales los detallamos en la **Tabla 21.**

Tabla 21. Campos de la perspectiva Área interés.

Campo	Descripción
Valor área de interés	Valor del área de interés del interesado en seguir el curso.

La perspectiva **Nivel formación** tiene los siguientes campos, los cuales los detallamos en la **Tabla 22.**

Tabla 22. Campos de la perspectiva Nivel formación.

Campo	Descripción
Valor nivel formación	Valor del nivel formación del interesado en seguir el curso.

La perspectiva **Horario** tiene los siguientes campos, los cuales los detallamos en la **Tabla 23.**

Tabla 23. Campos de la perspectiva Horario.

Campo	Descripción
Valor horario	Valor del horario de estudio del interesado en seguir el curso.

2.2.4. Modelo conceptual ampliado.

Con la información desagregada en los pasos anteriores procedemos a ampliar el modelo conceptual, colocando en las perspectivas los campos encontrados y en los indicadores las fórmulas de cálculo necesarias, lo que detallamos en la **Figura 14.**

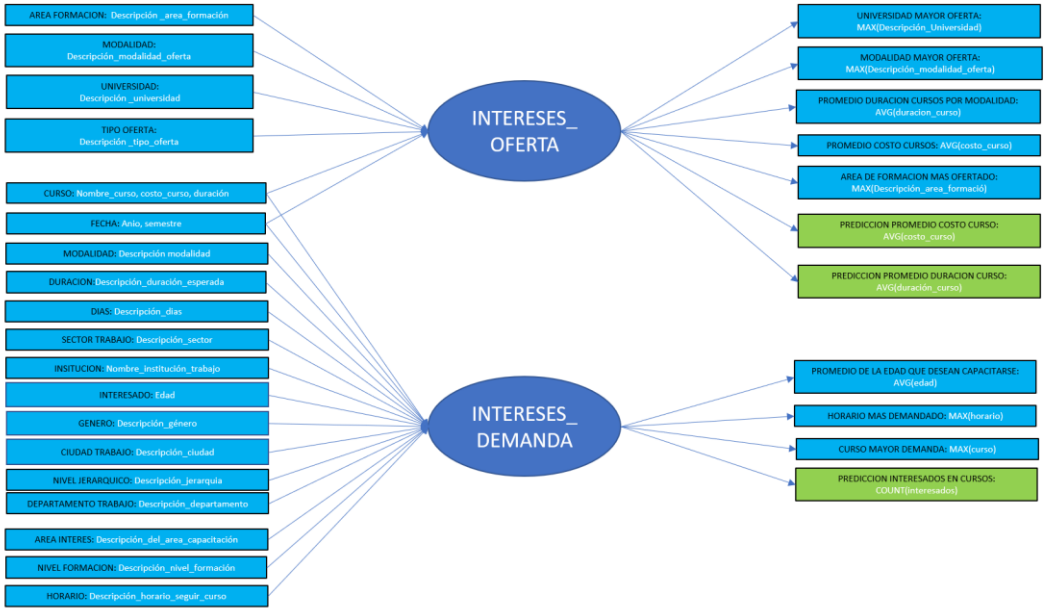


Figura 14. Modelo conceptual ampliado

2.3. Modelo lógico del DW.

En función del modelo conceptual ampliado que se creó en el punto anterior, procedemos a la creación de las dimensiones, las tablas de hechos y sus correspondientes uniones.

2.3.1. Tipo de modelo lógico del DW.

El modelo seleccionado para el DW es del tipo constelación, se optó por este modelo debido a que necesitamos analizar la oferta y demanda al mismo tiempo; adicional nos permite la reutilización de tablas de dimensiones Curso y Fecha.

2.3.2. Tablas de dimensiones.

Las perspectivas y sus correspondientes dimensiones están representadas en la **Tabla 24.**

Tabla 24. Perspectivas y sus correspondientes dimensiones

PERSPECTIVA	DIMENSION

Área Formación. • Descripción_area_formación	Dimension_AreaFormacion • arfo_id • arfo_valor
Modalidad. • Descripción_modalidad_oferta	Dimension_ModalidadOferta • modo_id • modo_valor
Universidad. • Descripción _universidad	Dimension_Universidad • univ_id • univ_nombre
Tipo Oferta. • Descripción _tipo_oferta	Dimension_TipoOferta • tiof_id • tiof_valor
Curso. • Nombre_curso. • Costo_curso. • Duración	Dimension_Curso • curs_id • curs_nombre
Fecha • Año. • Semestre	Dimesion_Fechas • fech_id • fech_anio • fech_semestre_uno • fech_semestre_dos • fech_fecha
Modalidad • Descripción modalidad	Dimencion_ModalidadDemanda • modd_id • modd_nombre
Duración • Descripción_duración_esperada	Dimencion_DuracionDemanda • durd_id • durd_valor
Días	Dimencion_Dias

• Descripción_dias	• dia_id • dia_valor
Sector Trabajo • Descripción_sector	Dimension_SectorTrabajo • sect_id • sect_valor
Institución • Nombre_institución_trabajo	Dimension_Institucion • inst_id • inst_nombre
Interesado • Edad	Dimension_Interesado • dema_part_edad
Genero • Descripción_género	Dimension_Genero • gene_id • gene_valor
Ciudad trabajo • Descripción_ciudad	Dimension_CiudadTrabajo • ciud_id • ciud_valor
Nivel jerarquico • Descripción_jerarquia	Dimencion_NivelJerarquico • jera_id • jera_valor
Departamento trabajo • Descripción_departamento	Dimension_Departamento • depa_id • depa_valor
Area interés • Descripción_del_area_capacitación	Dimension_AreaInteres • arin_id • arin_nombre
Nivel formación • Descripción_nivel_formación	Dimension_NivelFormacion • nifo_id • nifo_nivel
Horario	Dimencion_Horario

• Descripción_horario_seguir_curso	• hora_id • hora_valor

De forma gráfica las tablas de dimensiones quedarían representadas en la **Figura 15.**

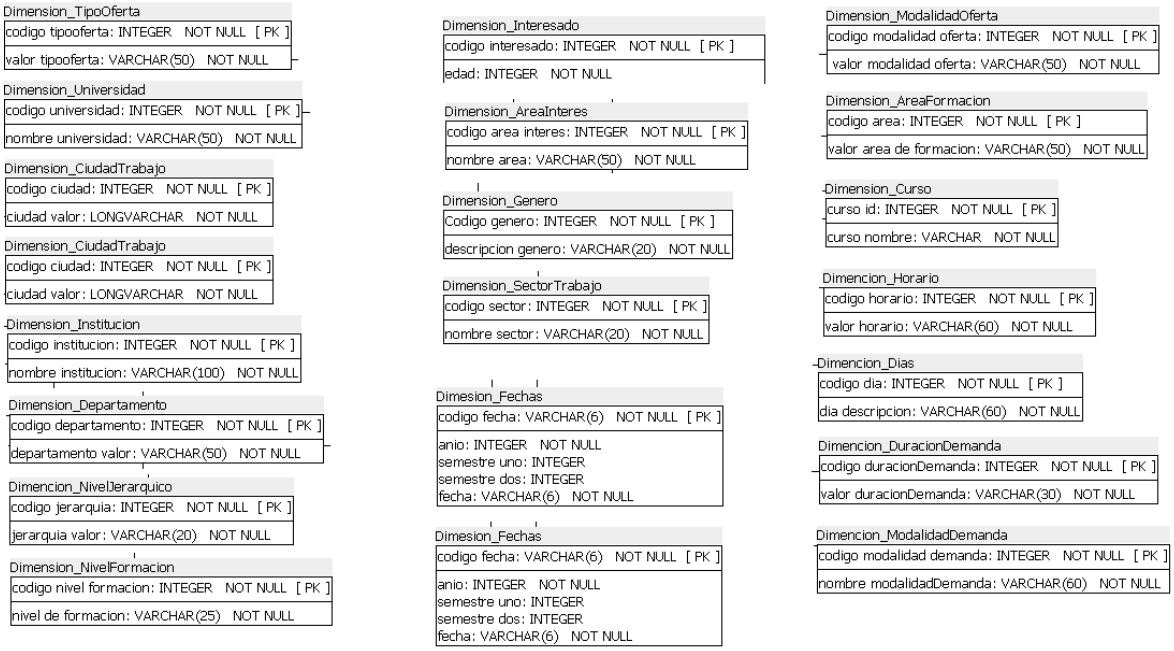


Figura 15. Gráfica de las dimensiones

2.3.3. Tabla de hechos.

Las tablas de hechos van a contener la información que nos permitirá construir los indicadores definidos.

En nuestro caso de estudio necesitamos conocer información referente a la **Oferta** y la **Demanda** al mismo tiempo, de tal manera que vamos a tener dos tablas de hechos.

La tabla **Hechos_Oferta**, tiene la siguiente estructura, tal como indica la **Figura 16**.

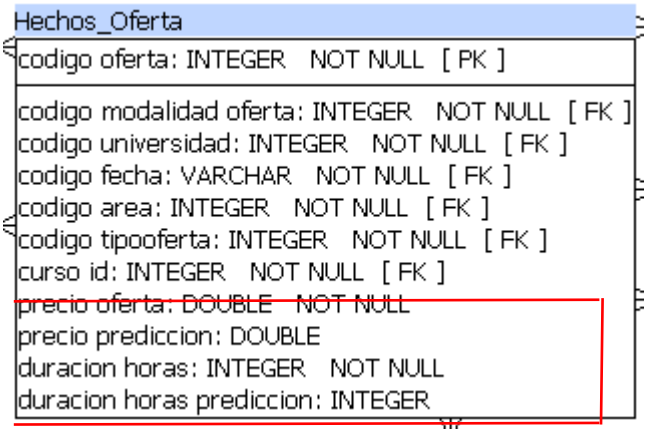


Figura 16 Indicadores de la tabla de hechos Hechos_Oferta

Los indicadores duración_horas_predicción, precio_predicción corresponde a los indicadores dinámicos, los cuales van a ser calculados mediante un modelo de predicción.

La tabla **Hechos_Demanda**, tiene la siguiente estructura, tal como indica la **Figura 17**.

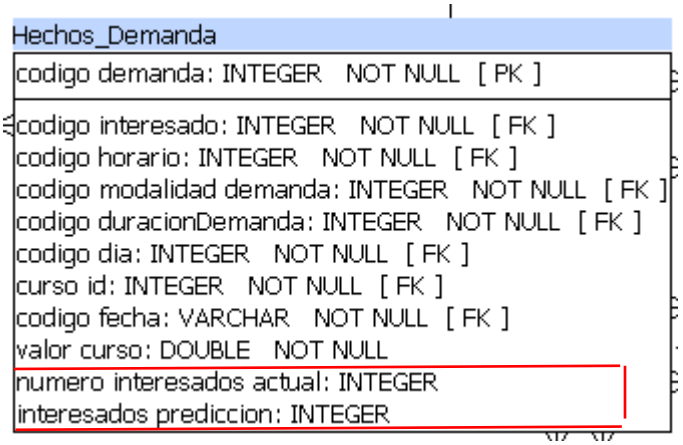


Figura 17 Indicadores de la tabla de hechos Hechos_Demanda

El indicador interesados_prediccion va a ser calculado mediante un modelo de predicción.

2.3.4. Uniones.

En esta etapa debemos establecer las uniones correspondientes de las tablas de dimensiones con las tablas de hechos, para nuestro caso las dimensiones **Curso**, **Fecha** se van a unir con las tablas de hechos **Hechos_Oferta** y **Hechos_Demanda.**

El diagrama de uniones de las dimensiones y hechos está representado en la **Figura 18.**

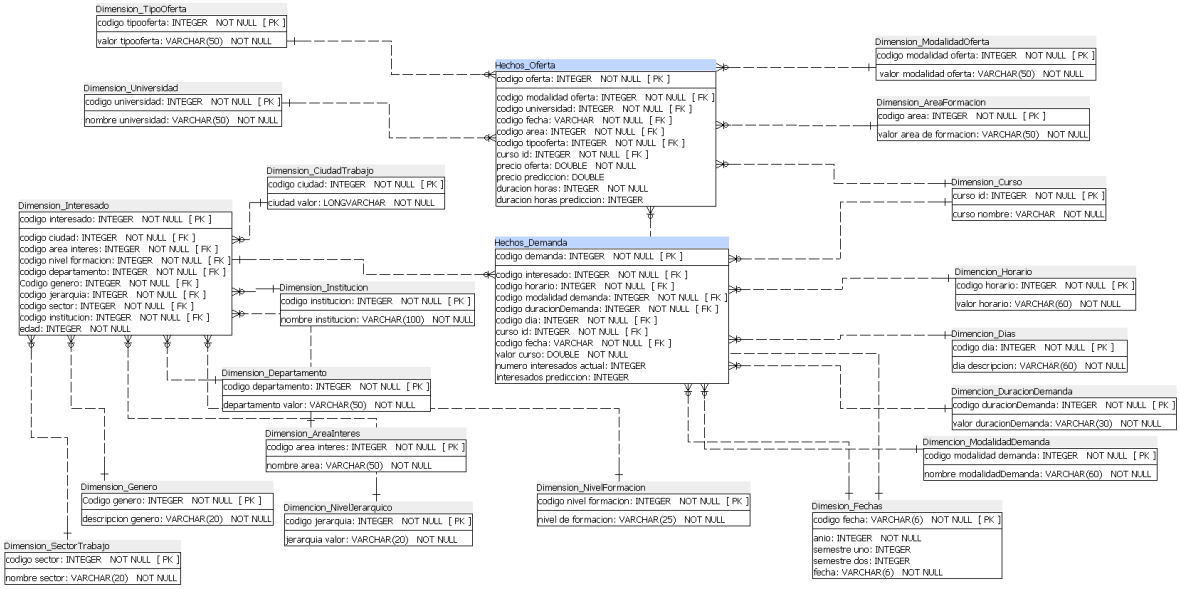


Figura 18. Diagrama de uniones

2.4. Integración de datos.

En esta fase vamos a describir los pasos que se siguieron para poder cargar la información dentro del DW.

Hay que indicar que antes de proceder con la subida de los datos es necesario realizar procesos de Extracción, Transformación y por último la fase de Carga.

En los pasos de Extracción y Transformación fue necesario realizar actividades de limpieza y calidad de datos, utilizando herramientas como Notepad++, Excel, Python, RapidMiner.

Los ETL fueron diseñados en dos etapas, la primera etapa se enfoca en las dimensiones y tabla de hechos referentes a la Oferta

El procedimiento de extracción, transformación y carga de datos de la información de la Oferta lo tenemos representado en la **Figura 19.**

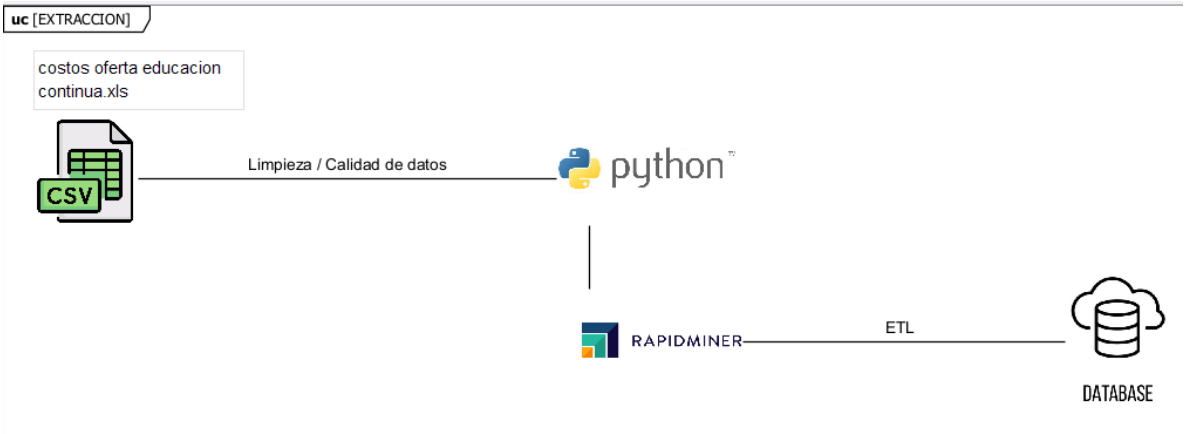


Figura 19. Proceso ETL para tabla Hechos_Oferta.

La información para las dimensiones asociadas a la tabla Hechos_Oferta, se encuentran en columnas del documento **costos Oferta de Educación continua.xlsx**; la extracción se realizó con Python y un extracto de esta actividad está plasmada en la **Figura 20.**

La dinámica se repite para las dimensiones Dimension_TipoOferta, Dimension_ModalidadOferta, Dimension_AreaFormacion, Dimension_curso.

```
In [1]: import numpy as np
        import pandas as pd

        Universidades

In [7]: df=pd.read_excel('costos Oferta de Educación continua.xlsx',sheet_name='oferta')
        df.head()

Out[7]:
```

	Universidad	Tipos de Oferta	Areas de Formación
0	udla universidad de las americas	Maestrias	Administracion y negocios
1	NaN	NaN	Educacion
2	NaN	NaN	Salud
3	NaN	Diplomados	Administracion y negocios
4	NaN	NaN	Tecnologia e informacion

```
In [22]: df.shape
Out[22]: (89, 3)

In [29]: dfu=pd.DataFrame(df['Universidad'].value_counts(dropna=False)).reset_index()
         dfu.rename(columns={"index":'Universidad',"Universidad":'Cantidad'},inplace=True)
         dfu.head()

Out[29]:
```

	Universidad	Cantidad
0	NaN	74
1	udla universidad de las americas	1
2	uide universidad internacional del ecuador	1
3	universidad de los hemisferios	1
4	utpl Universidad Particular de Loja	1

```
In [25]: dfu.to_csv('Universidad.csv')
```

Figura 20 Extracción de datos de columnas para crear las dimensiones

Para la carga de información a las tablas Dimension_TipoOferta, Dimension_ModalidadOferta, Dimension_AreaFormacion, Dimension_curso, diseñamos un proceso automático en RapidMiner, el mismo está representado en la **Figura 21.**

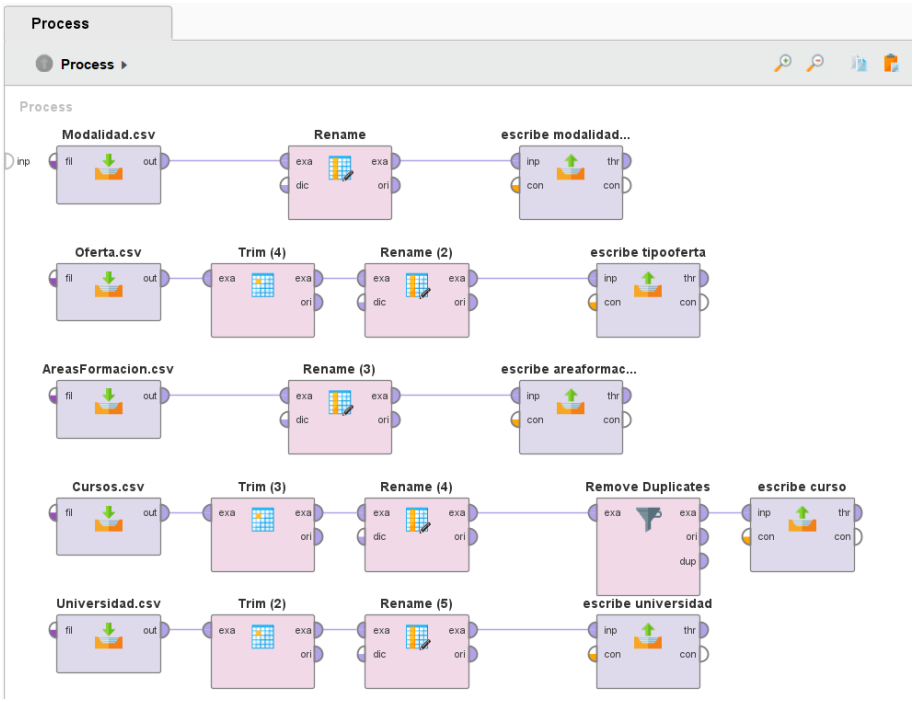


Figura 21. Carga de información en tablas de dimensiones.

Para poder cargar información en la tabla Hechos_Oferta, fue necesario realizar un proceso de JOIN utilizando RapidMiner, esta actividad esta plasmada en la **Figura 22.**

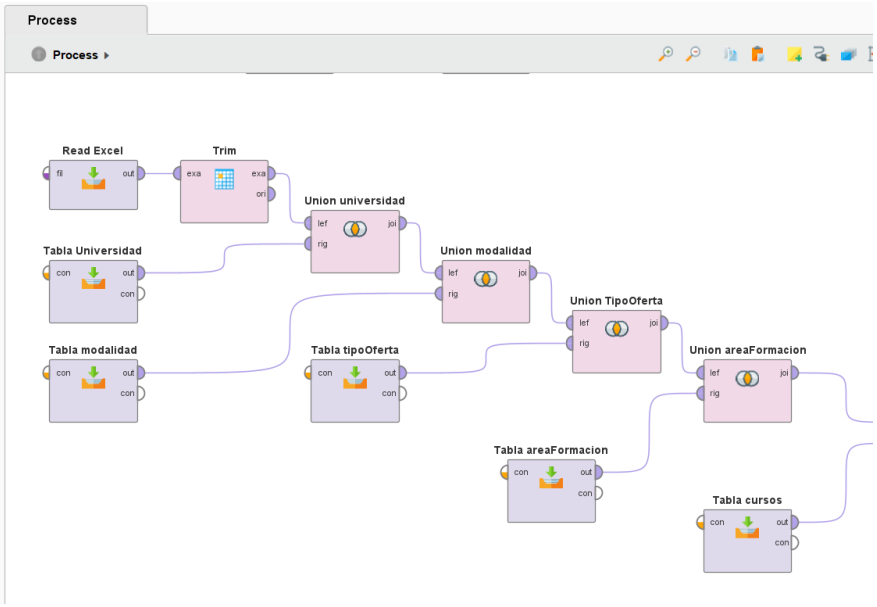


Figura 22. Proceso JOIN para unir información de tablas.

Una vez realizado el proceso de JOIN, procedemos a cargar la información en la tabla Hechos_Oferta, nos apoyamos diseñando un proceso con RapidMiner, el cual está representado en la **Figura 23.**

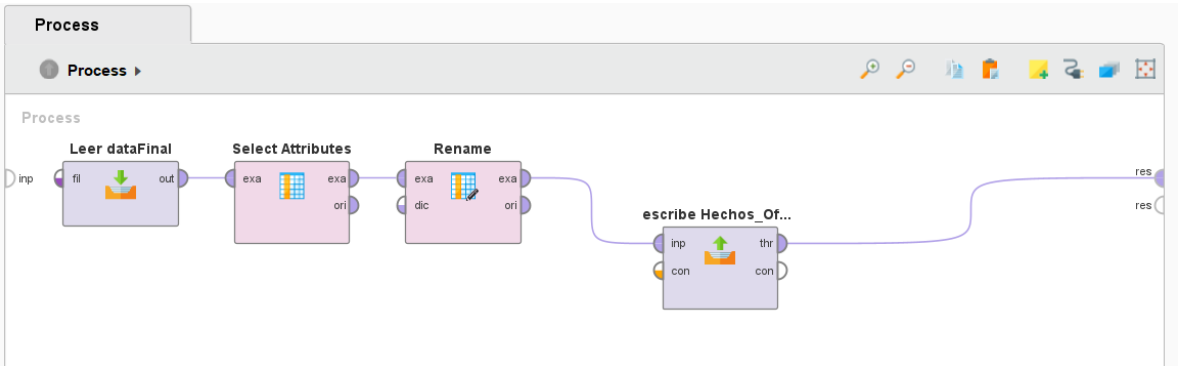


Figura 23. Carga de información a tabla Hechos_Oferta

La segunda fase tiene que ver con los ETL necesarios para cargar información en la tabla Hechos_Demanda y sus correspondientes tablas de dimensiones asociadas, este procedimiento lo podemos visualizar en la **Figura 24.**

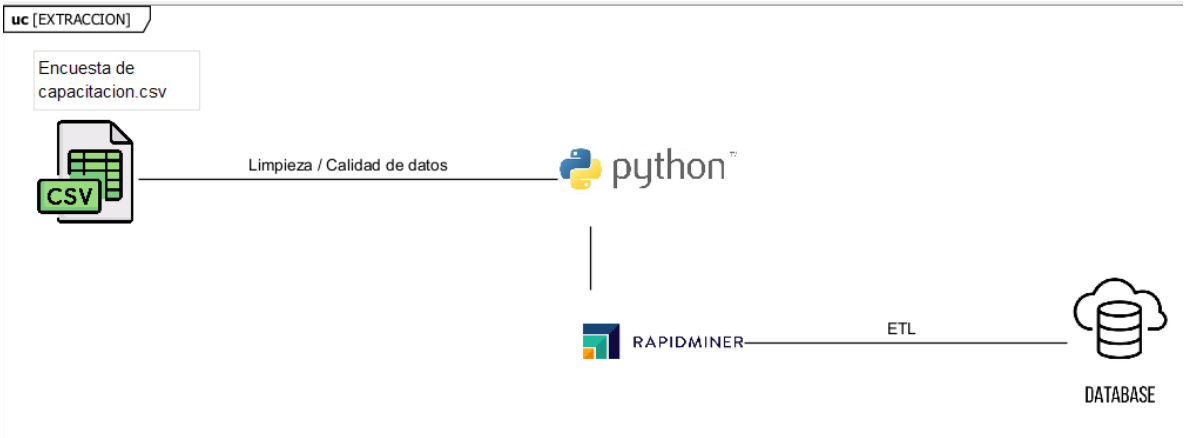


Figura 24. Proceso ETL para tabla Hechos_Demanda

Se procedió a realiza la limpieza de la información del documento Encuesta de capacitación.csv, para esto se utilizó un proceso elaborado en RapidMiner, la **Figura 25** representa el proceso.

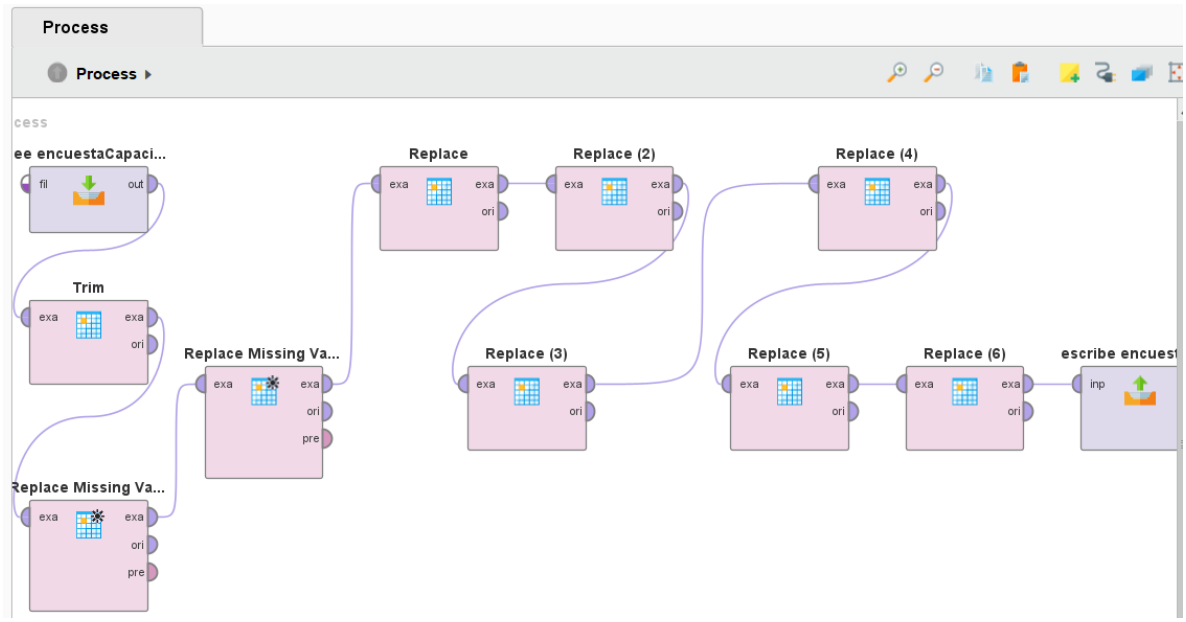
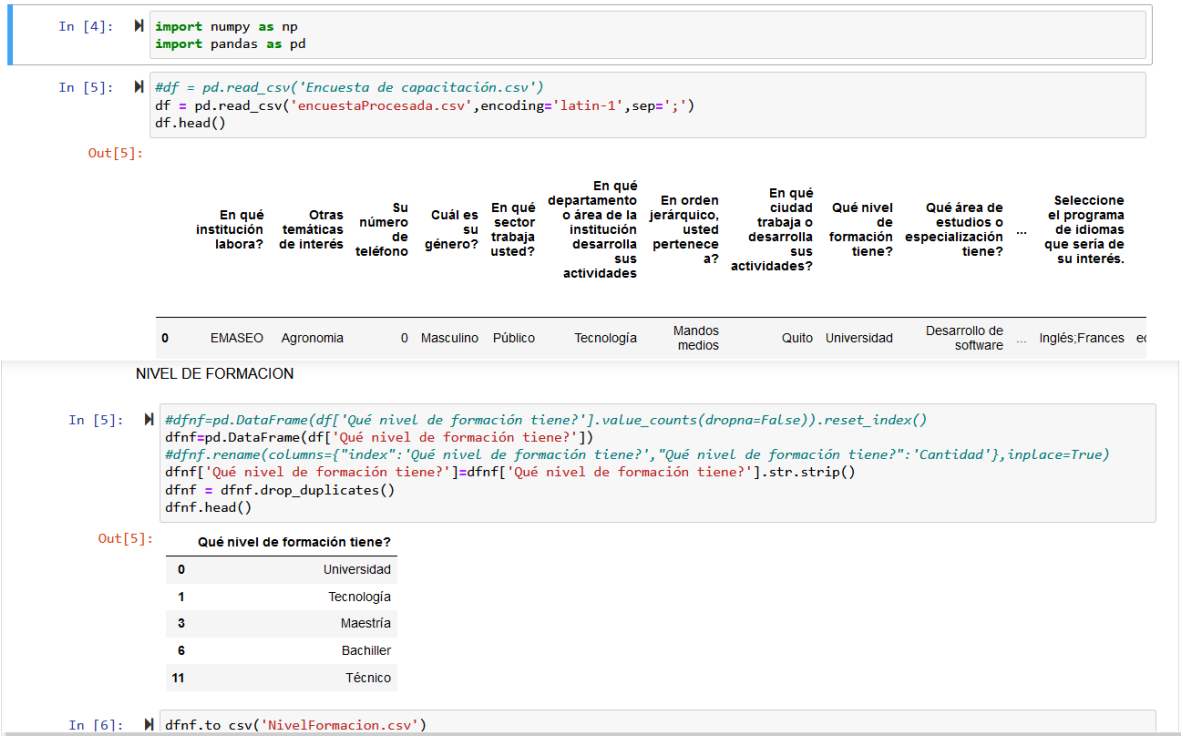


Figura 25. Limpieza de datos.

Una vez ejecutado el proceso de limpieza de datos, procedemos a extraer la información para poder cargar información en las tablas de dimensiones asociadas a la tabla **Hechos_Demanda**.

Esto lo conseguimos con un script de la herramienta Python, el mismo esta representado en la **Figura 26.**



```
In [4]:
import numpy as np
import pandas as pd
```

```
In [5]:
#df = pd.read_csv('Encuesta de capacitación.csv')
df = pd.read_csv('encuestaProcesada.csv',encoding='latin-1',sep=';')
df.head()
```

Out[5]:

	En qué institución labora?	Otras temáticas de interés	Su número de teléfono	Cuál es su género?	En qué sector trabaja usted?	En qué departamento o área de la institución desarrolla sus actividades	En orden jerárquico, usted pertenece a?	En qué ciudad trabaja o desarrolla sus actividades?	Qué nivel de formación tiene?	Qué área de estudios o especialización tiene?	...	Seleccione el programa de idiomas que sería de su interés.
0	EMASEO	Agronomia	0	Masculino	Público	Tecnología	Mandos medios	Quito	Universidad	Desarrollo de software	...	Inglés;Frances

NIVEL DE FORMACION

```
In [5]:
#dfnf=pd.DataFrame(df['Qué nivel de formación tiene?'].value_counts(dropna=False)).reset_index()
dfnf=pd.DataFrame(df['Qué nivel de formación tiene?'])
#dfnf.rename(columns={"index":'Qué nivel de formación tiene?',"Qué nivel de formación tiene?":'Cantidad'},inplace=True)
dfnf['Qué nivel de formación tiene?']=dfnf['Qué nivel de formación tiene?'].str.strip()
dfnf = dfnf.drop_duplicates()
dfnf.head()
```

Out[5]:

	Qué nivel de formación tiene?
0	Universidad
1	Tecnología
3	Maestría
6	Bachiller
11	Técnico

```
In [6]:
dfnf.to csv('NivelFormacion.csv')
```

Figura 26. Script extracción información para tablas de dimensión.

Procedemos a la carga de información para las dimensiones Dimension_SectorTrabajo, Dimension_Genero, Dimencion_NivelJerarquico, Dimension_AreaInteres, Dimension_Departamento, Dimension_Institucion, Dimension_CiudadTrabajo, Dimension_NivelFormacion, Dimencion_ModalidadDemanda, Dimencion_DuracionDemanda, Dimencion_Dias, Dimencion_Horario.

Utilizamos un proceso de Rapidminer, representado en la **Figura 27**.

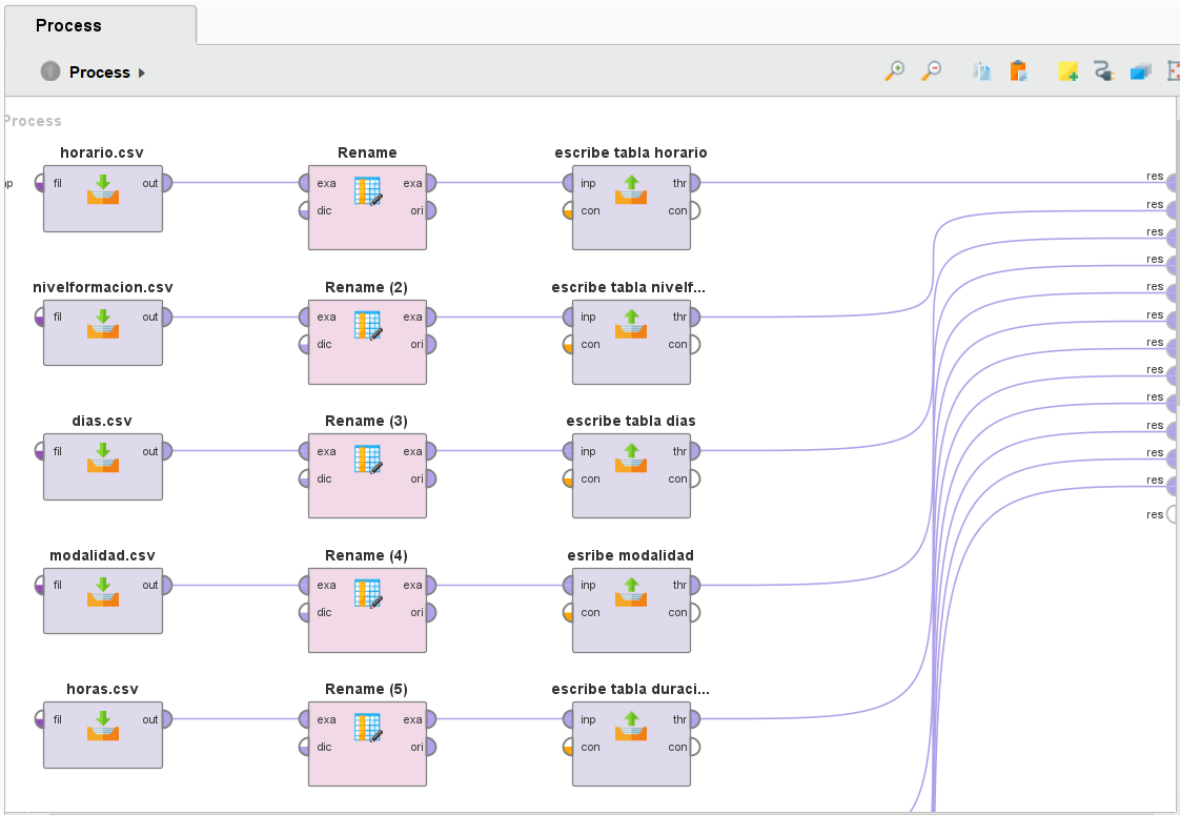


Figura 27. Carga información en tablas de dimensiones.

Para cargar información en las tablas Dimension_Interesado y Hechos_Demanda, fue necesario elaborar un proceso JOIN con Rapidminer, el cual lo detallamos en la **Figura 28**.

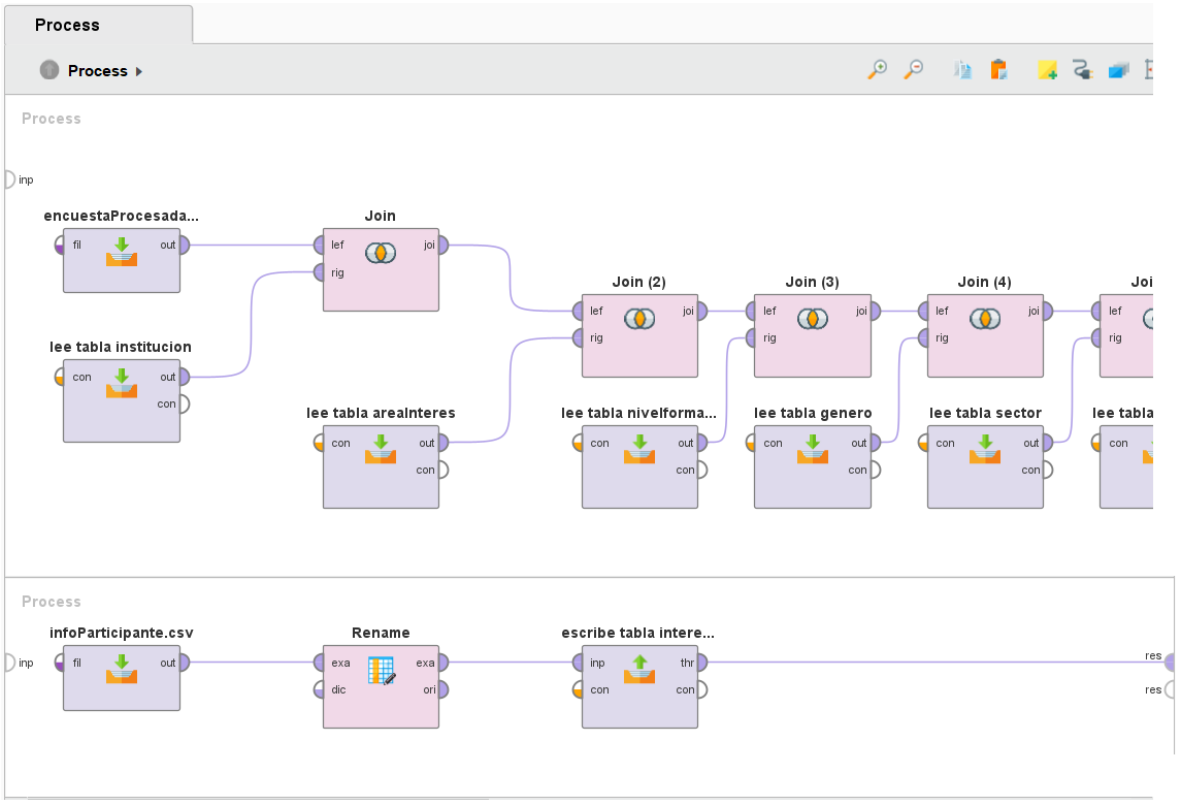


Figura 28. Proceso JOIN y escritura en Dimension_participante

La carga de información en la tabla Hechos_demanda, se describe en la **Figura 29.**

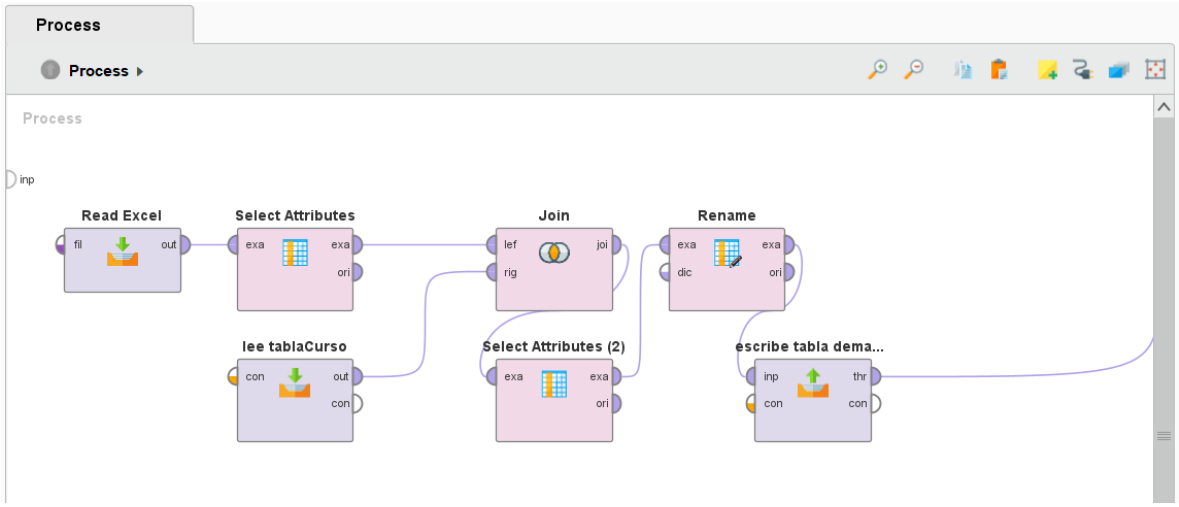


Figura 29. Carga información tabla Hechos_demanda

2.4.1. Indicadores dinámicos.

Para poder generar información para los indicadores dinámicos se creó modelos predictivos mediante la herramienta RapidMiner gracias a su módulo llamado Auto Model.

Las siguientes figuras nos muestran el proceso seguido para generar el modelo para el indicador dinámico prediccion_duracion_curso.

Comenzamos seleccionando los datos para preparar el modelo tal como lo indica la **Figura 30.**

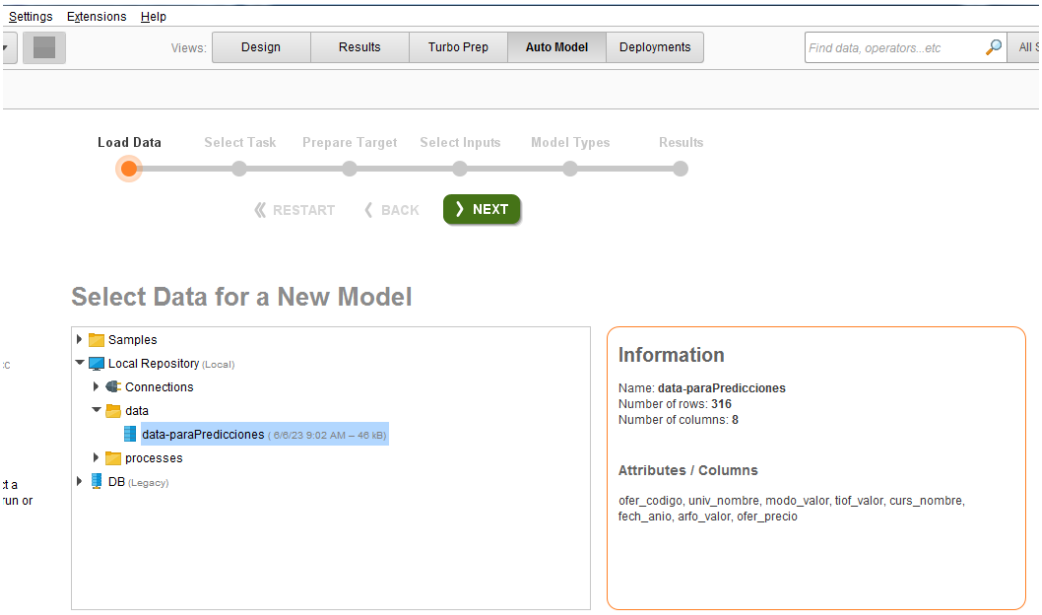


Figura 30. Selección de datos para el modelo

A continuación, vamos a seleccionar la columna de la que deseamos realizar la predicción tal como indica la **Figura 31.**

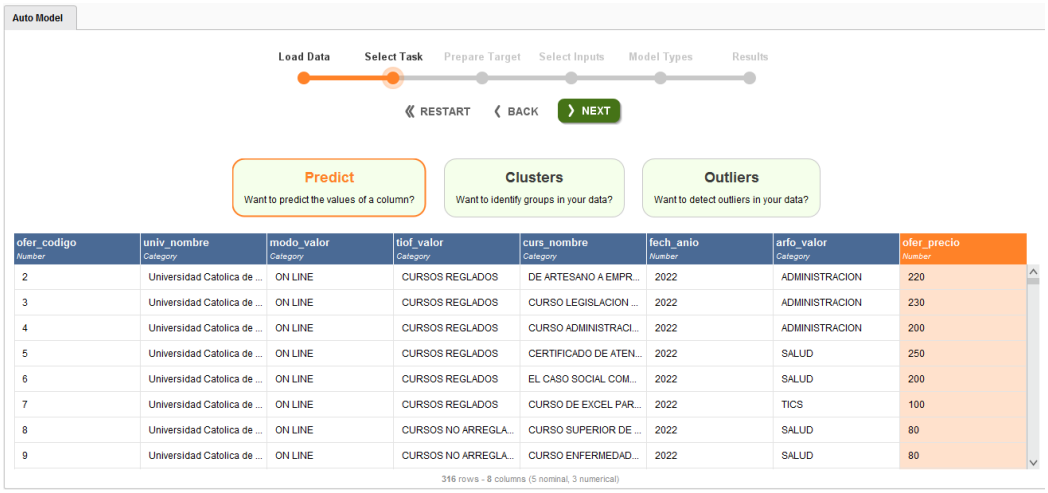


Figura 31. Selección de la columna para la predicción.

Comenzamos entonces la preparación de los objetivos ya sean para un proceso de regresión o clasificación, la **Figura 32.**

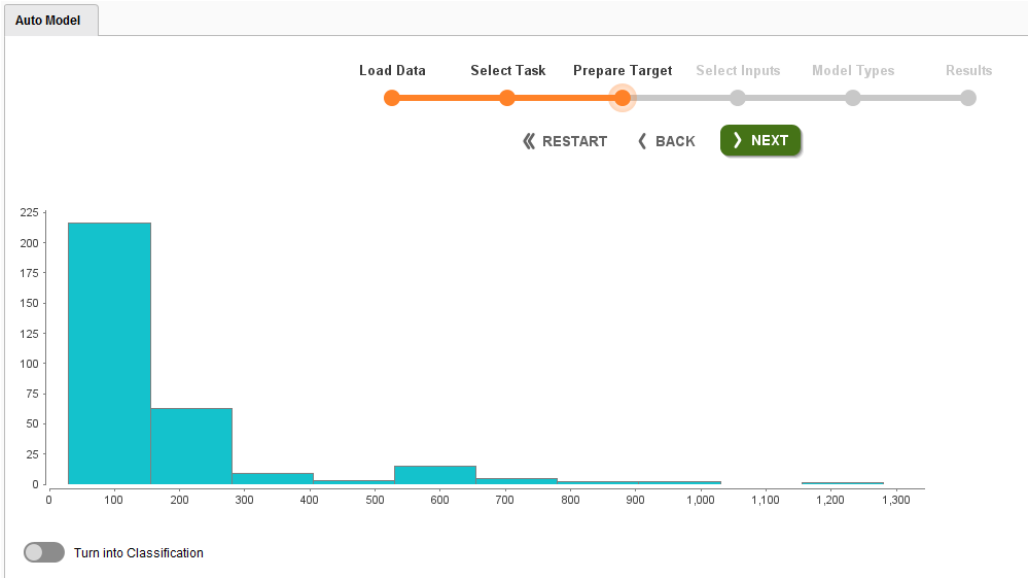


Figura 32. Preparación de objetivos.

Una consideración importante a la hora de seleccionar los atributos que van a intervenir en el entrenamiento es que deben tener una baja correlación y también es recomendable contar con campos cuya información sea numérica, la **Figura 33** nos indica los atributos seleccionados para el entrenamiento.

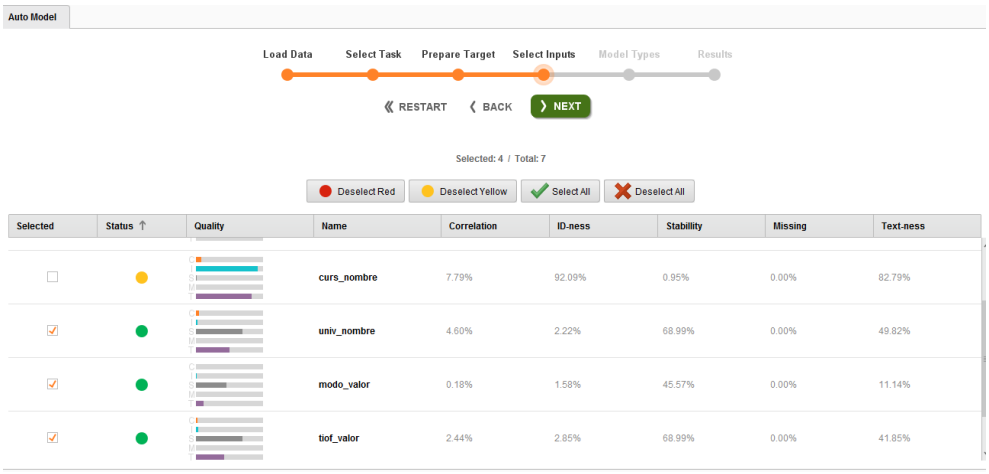


Figura 33. Selección de atributos para el entrenamiento

RapidMiner nos muestra una lista de modelos con los cuales podemos realizar nuestro proceso de entrenamiento, tal como indica la **Figura 34.**

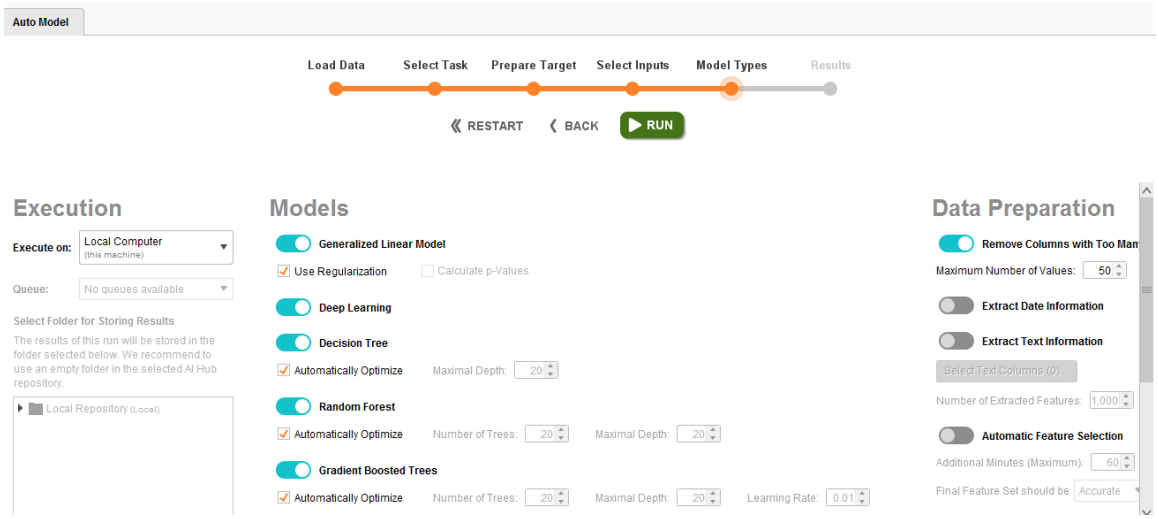


Figura 34. Modelos disponibles para entrenamiento

A continuación, se procede a realizar el entrenamiento y obtener las puntuaciones de los modelos en donde podemos observar en tiempo real como se van generando en base a las métricas seleccionadas en el paso anterior, una vez finalizado este proceso se enlista y generar unos puntajes para que el usuario pueda seleccionar el modelo con el cual desea trabajar, esto se refleja en la **Figura 35.**

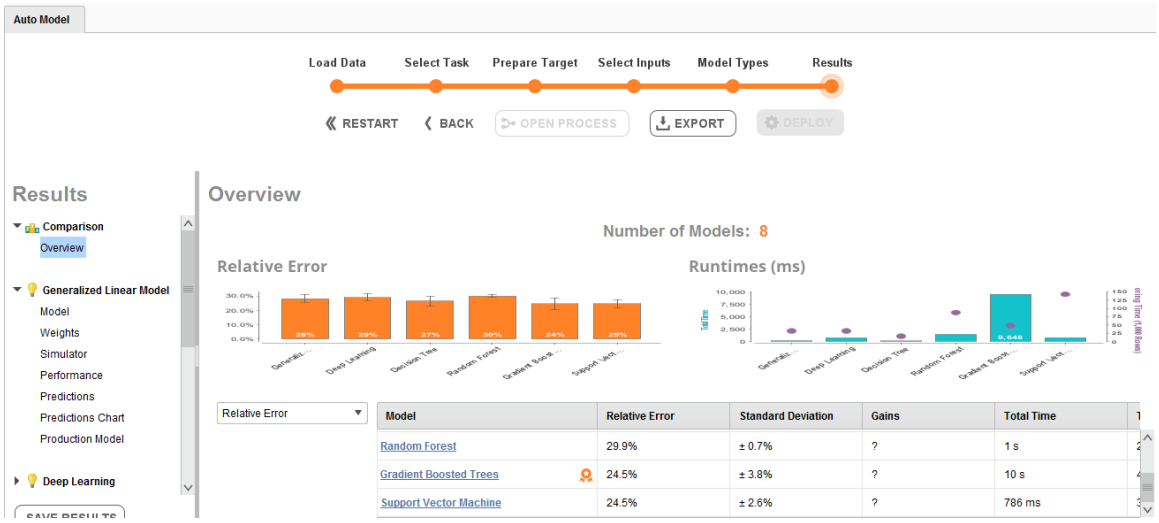


Figura 35. Modelos puntuados

La evaluación de los modelos se realizó en función de la puntuación generada por RapidMiner así como por los pesos de las variables y las predicciones generadas, lo cual lo podemos ver en la **Figura 36** y **Figura 37** respectivamente. Para este indicador se seleccionó el modelo Gradient Boosted Trees, el cual tiene el menor porcentaje de Error Relativo y el peso de las variables adecuado.

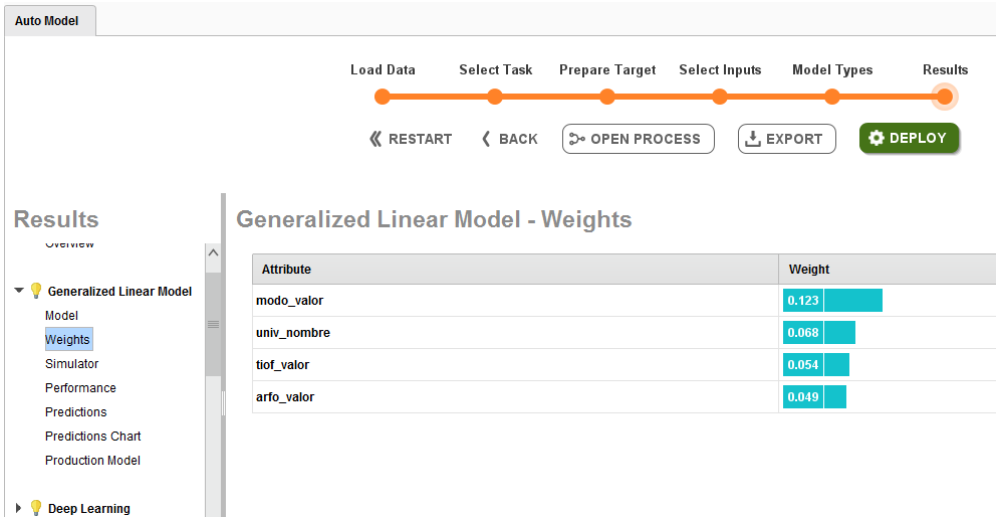


Figura 36. Pesos de las variables

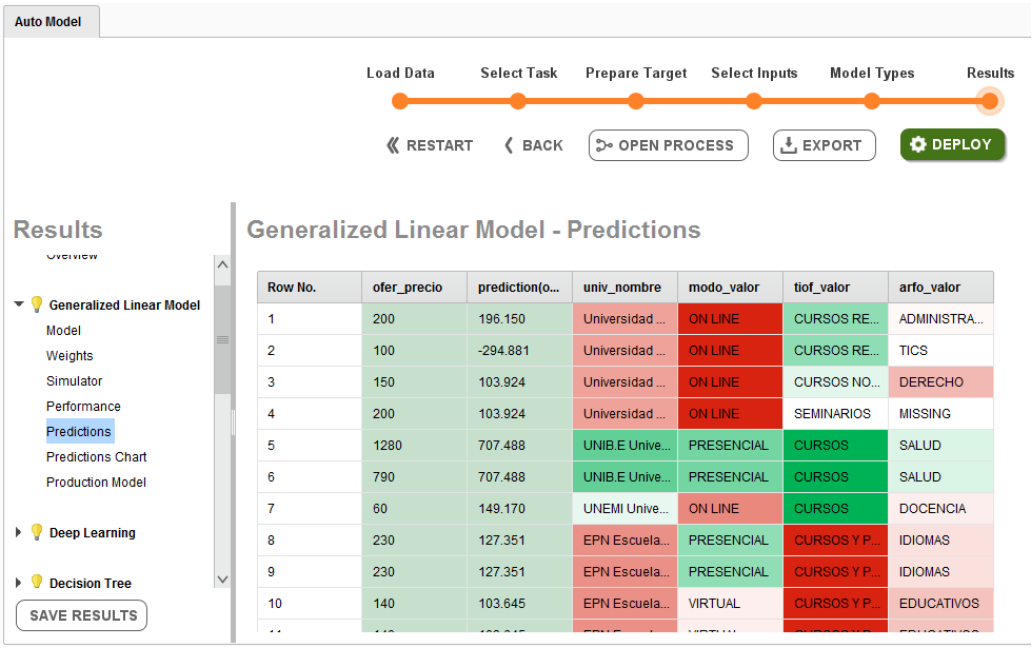


Figura 37. Resultados del entrenamiento

Para el indicador **predicción_duracion**, el algoritmo seleccionado fue **Decision_Tree** debido al porcentaje más bajo de Error Relativo y al peso de las variables como lo muestra la **Figura 38 y Figura 39** respectivamente.

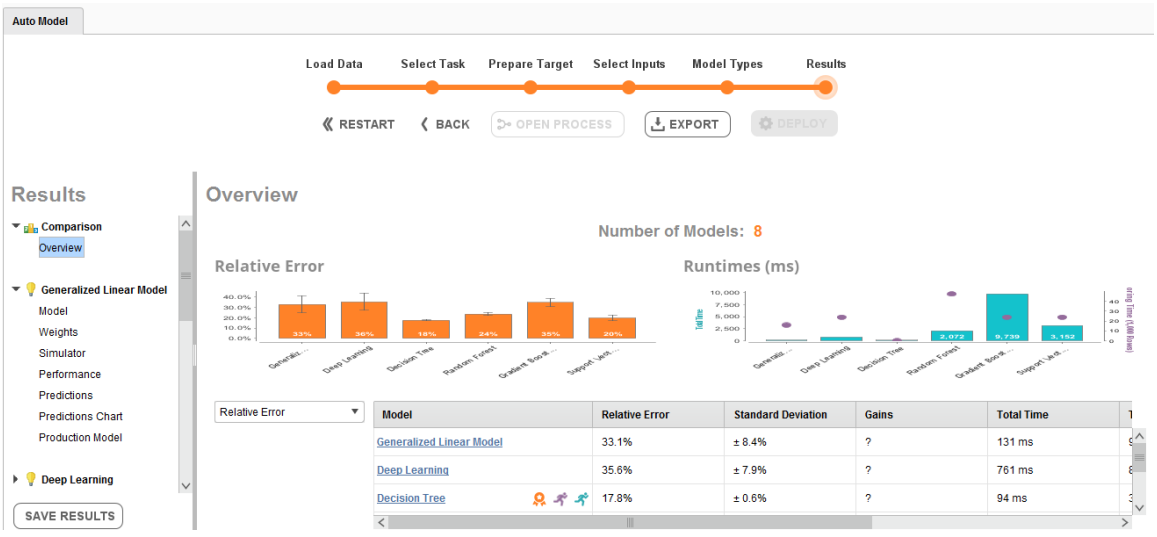


Figura 38. Selección del modelo

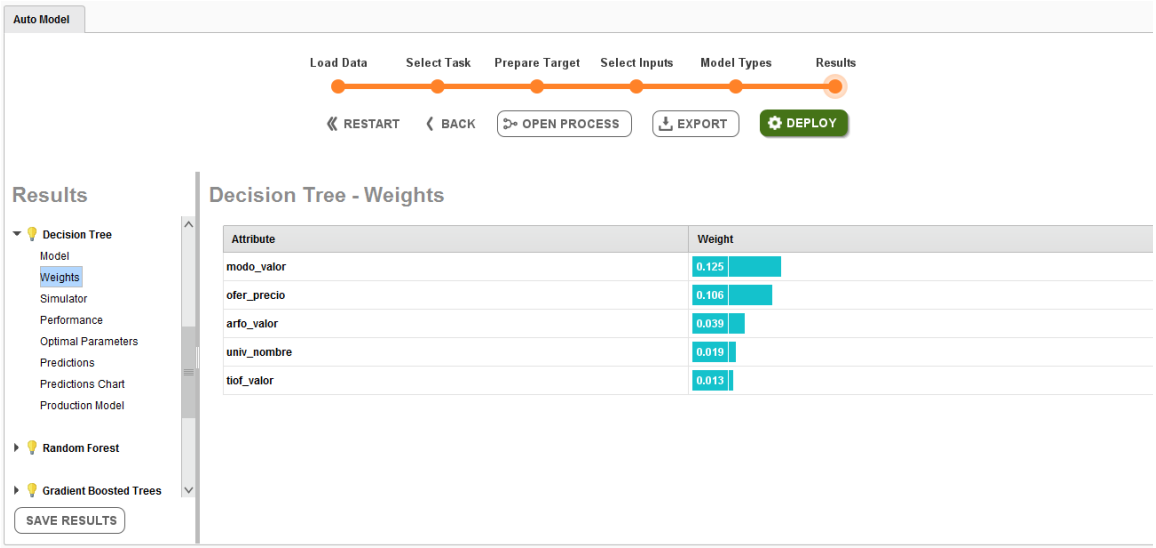
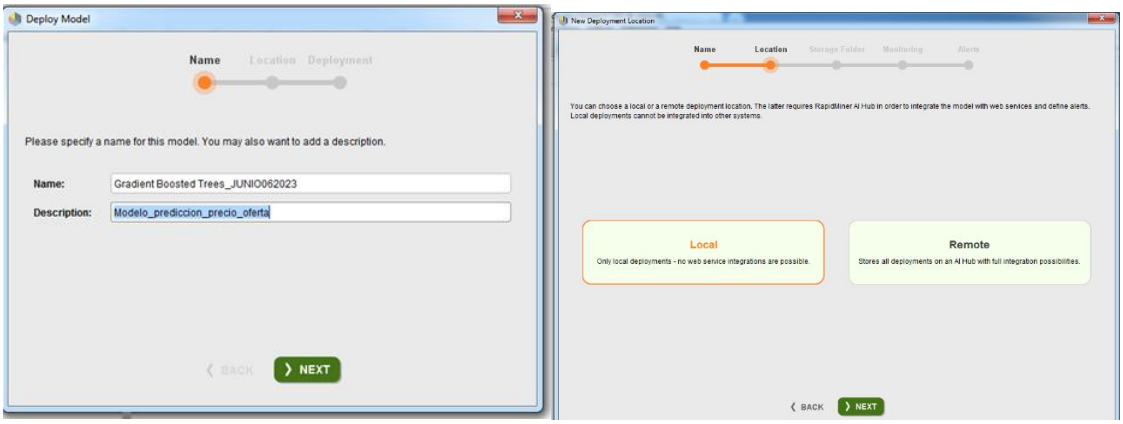


Figura 39. Pesos de las variables

Para poder utilizar los modelos generados, es necesario que se realice un proceso de DEPLOY, esto nos permitirá realizar futuras puntuaciones con nuevos datos, esta acción se resume en lo siguiente.

- Se debe asignar un nombre al modelo.
- Se debe indicar la ubicación donde guardar el modelo.
- Donde se va a desplegar
- Especificar el tipo si es regresión o clasificación.

La **Figura 40** nos da un resumen de estas acciones.



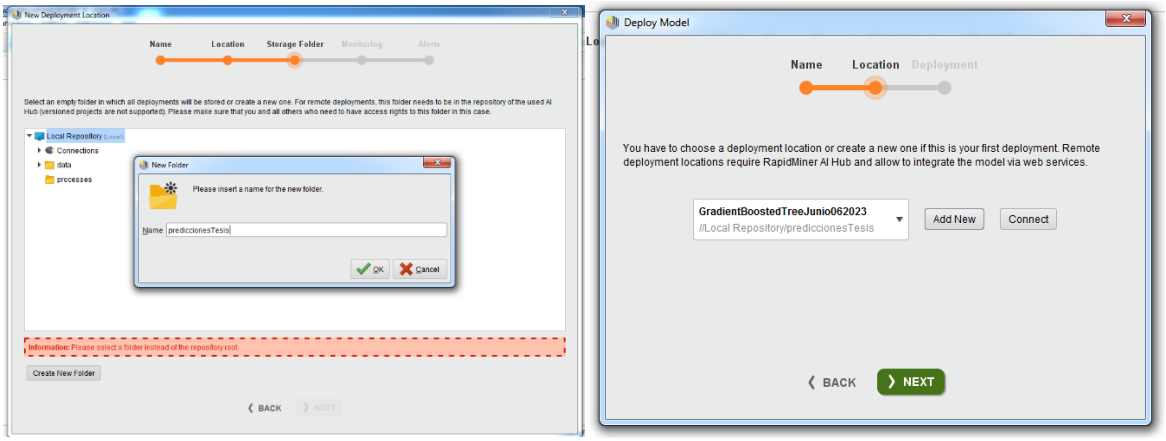
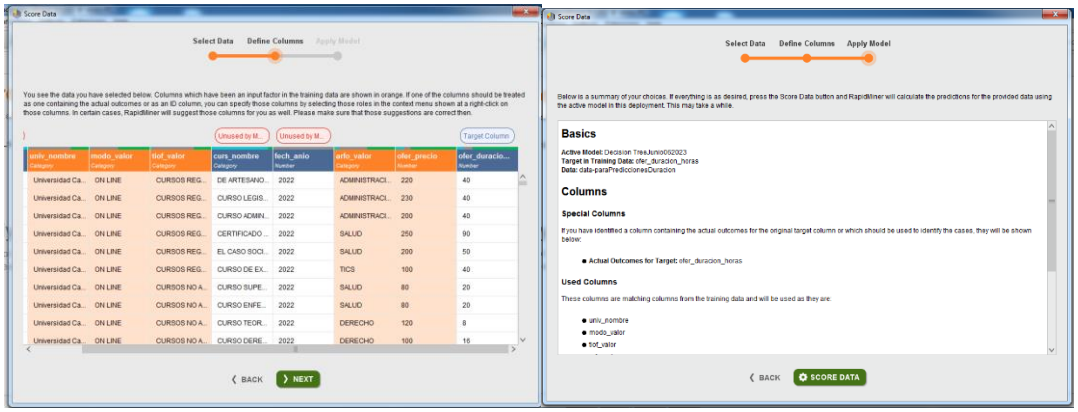


Figura 40. Procedimiento para el deploy del modelo

Finalmente, con el modelo guardado procedemos a generar los scoring para las predicciones, la **Figura 41** nos resume este proceso.



Row No.	ofer_duraci...	prediction(o...	ofer_codigo	fech_anio	univ_nombre	modo_valor	tiof_valor	arfo_valor	ofer_precio	curs_nombr...	curs_nombr...	curs_nom
310	40	50	311	2022	UPS UNIVER...	VIRTUAL	CURSO	TECNOLOGIA	90	0	0	0
311	40	50	312	2022	UPS UNIVER...	VIRTUAL	CURSO	TECNOLOGIA	220	0	0	0
312	40	50	313	2022	UPS UNIVER...	VIRTUAL	CURSO	TECNOLOGIA	200	0	0	0
313	35	50	314	2022	UPS UNIVER...	VIRTUAL	CURSO	TECNOLOGIA	200	0	0	0
314	40	50	315	2022	UPS UNIVER...	VIRTUAL	CURSO	TECNOLOGIA	100	0	0	0
315	45	50	316	2022	UPS UNIVER...	VIRTUAL	CURSO	TECNICO	180	0	0	0
316	45	50	317	2022	UPS UNIVER...	VIRTUAL	CURSO	MISSING	100	0	0	0

EXPORT

Figura 41. Generación de scoring

Estos datos se deben cargar en el DW para su correspondiente utilización, para lo cual se creó un proceso en RapidMiner, la **Figura 42** nos representa el proceso.

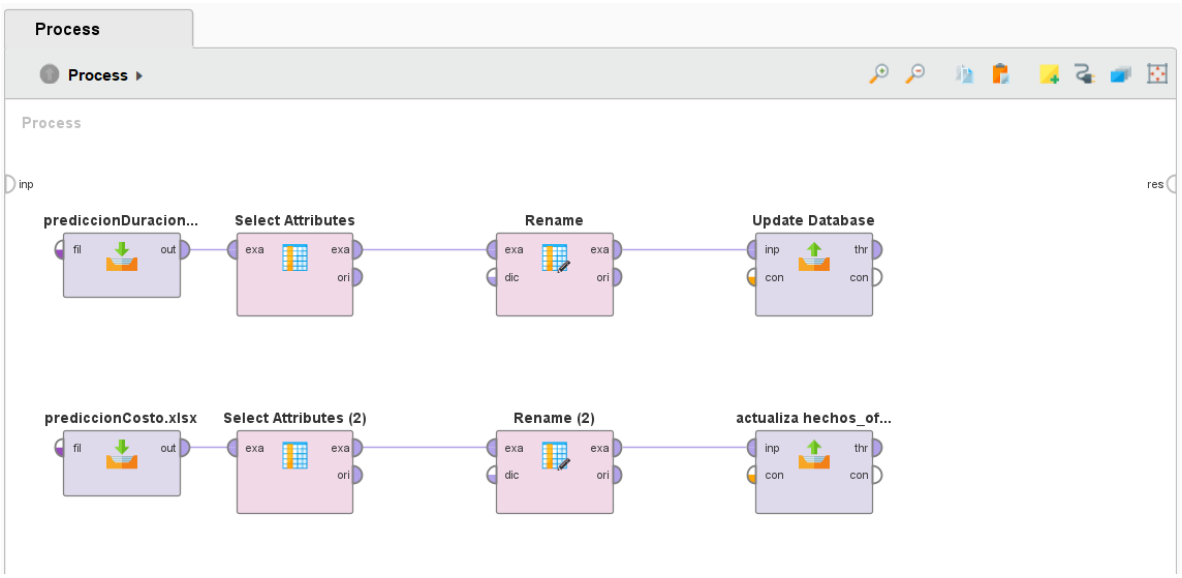


Figura 42. Proceso de actualización información indicadores

2.5. Visualización de datos.

La última etapa del proyecto es la visualización de la información, donde a través de graficas representamos información clave para la toma de decisiones por parte de los interesados.

Los elementos gráficos ayudan a un mejor entendimiento de la información que se quiere demostrar [9] sin embargo, se debe tener cuidado en la selección de los elementos visuales, pues en algunos casos en lugar de agregar valor, podrían ocasionar confusión en los resultados obtenidos.

La herramienta de software seleccionada para la visualización fue PowerBI, debido a su facilidad de uso y al proporcionar una versión de escritorio gratuita con todas las características de una versión pagada se pudo plasmar el DashBoard con los indicadores definidos.

La base de datos sobre la que se creó el DW fue Postgres [16] y la conexión con PowerBI se indica en la **Figura 43.**

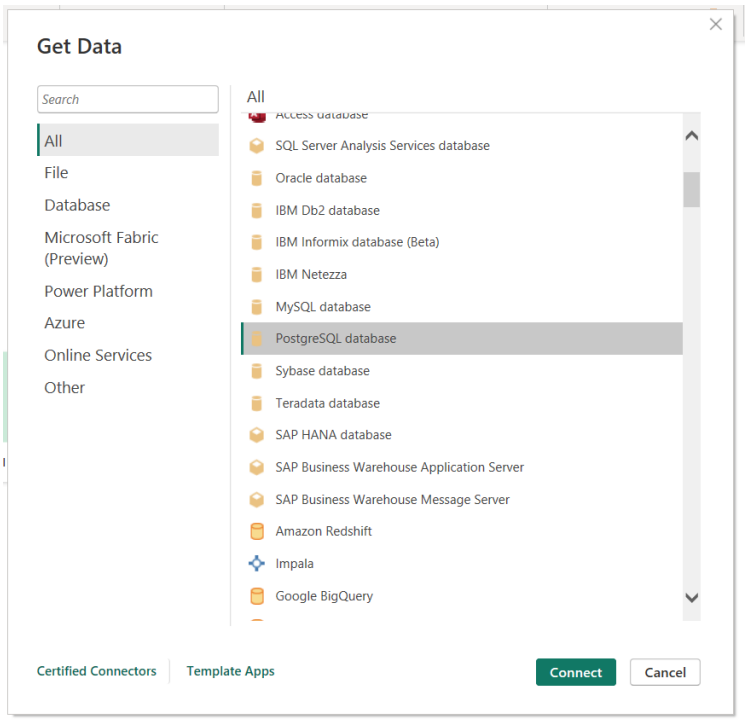


Figura 43. Conexión a la fuente de datos del DW

Una vez establecida la conexión ya tenemos a nuestra disposición la lista de tablas de dimensiones y hechos para poder elaborar las gráficas necesarias, la lista se representa en la **Figura 44.**

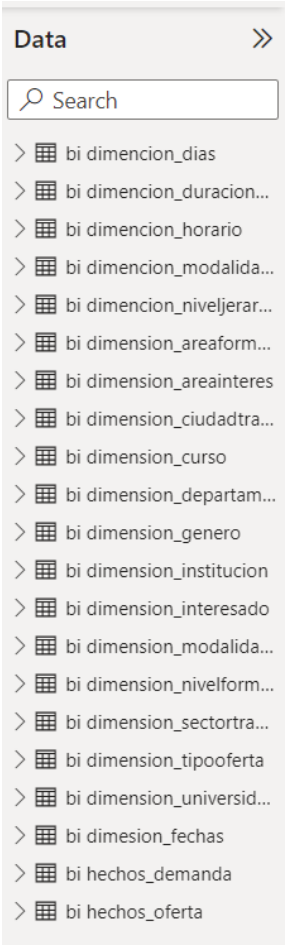


Figura 44. Tablas de dimensiones y hechos.

Los elementos gráficos fueron seleccionados en base al tipo de datos que deseamos presentar [9]. La **Tabla 25**, resume la gráfica y que tipo de información muestra.

GRAFICA	TIPO DE INFORMACION
Histogramas	Permiten visualizar como se distribuyen los datos.
Diagrama de Pasteles	Para representar la porción de datos de un todo.
Cards	Para representar información numérica de los indicadores.
Scatterplot	Para visualizar como se asocian de dos variables en un intervalo.

3. RESULTADOS.

La **Figura 45,** hace referencia al DashBoard para los indicadores clave de la Oferta.

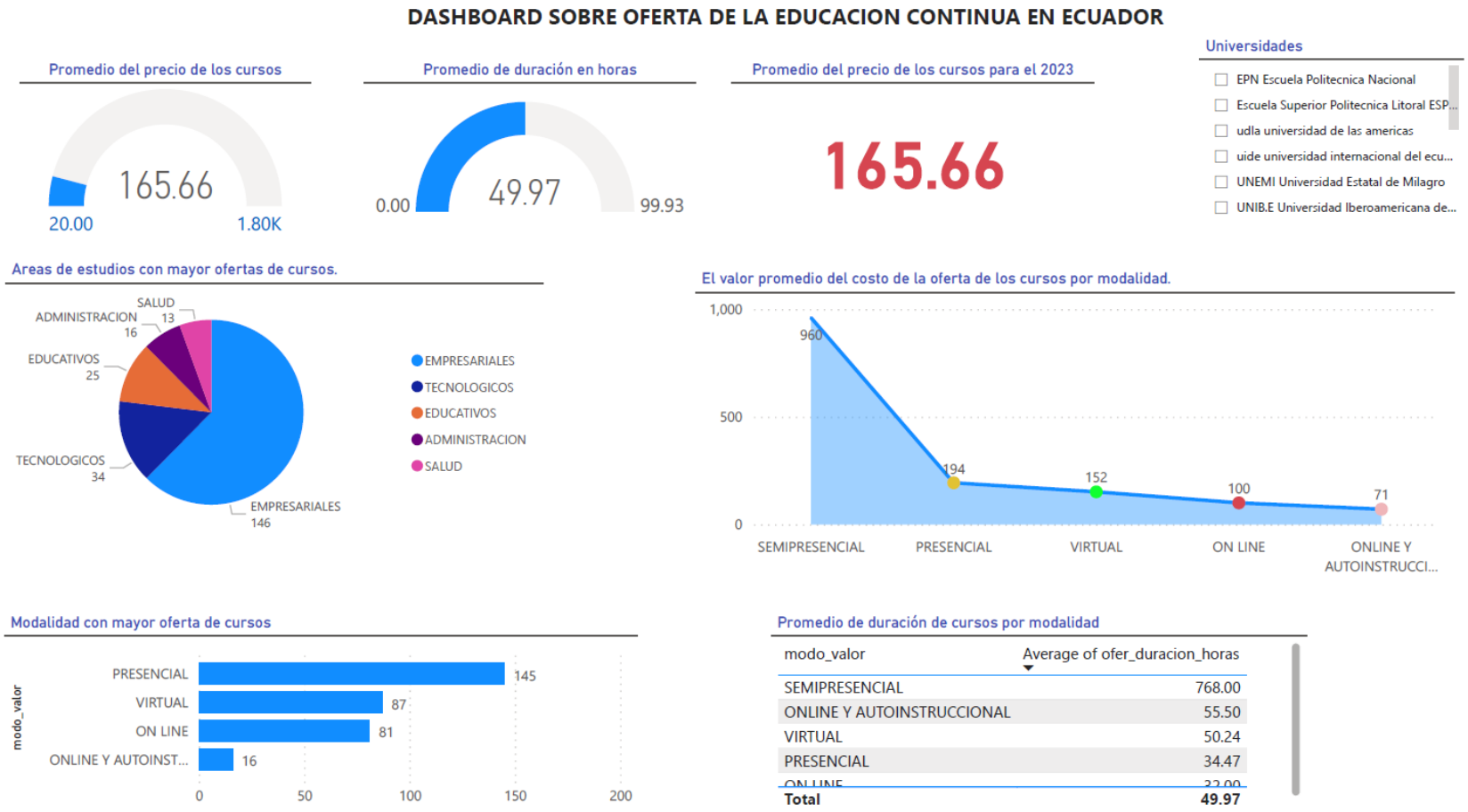


Figura 45. DashBoard para la oferta.

La **Figura 46,** hace referencia al DashBoard para los indicadores clave de la Demanda.

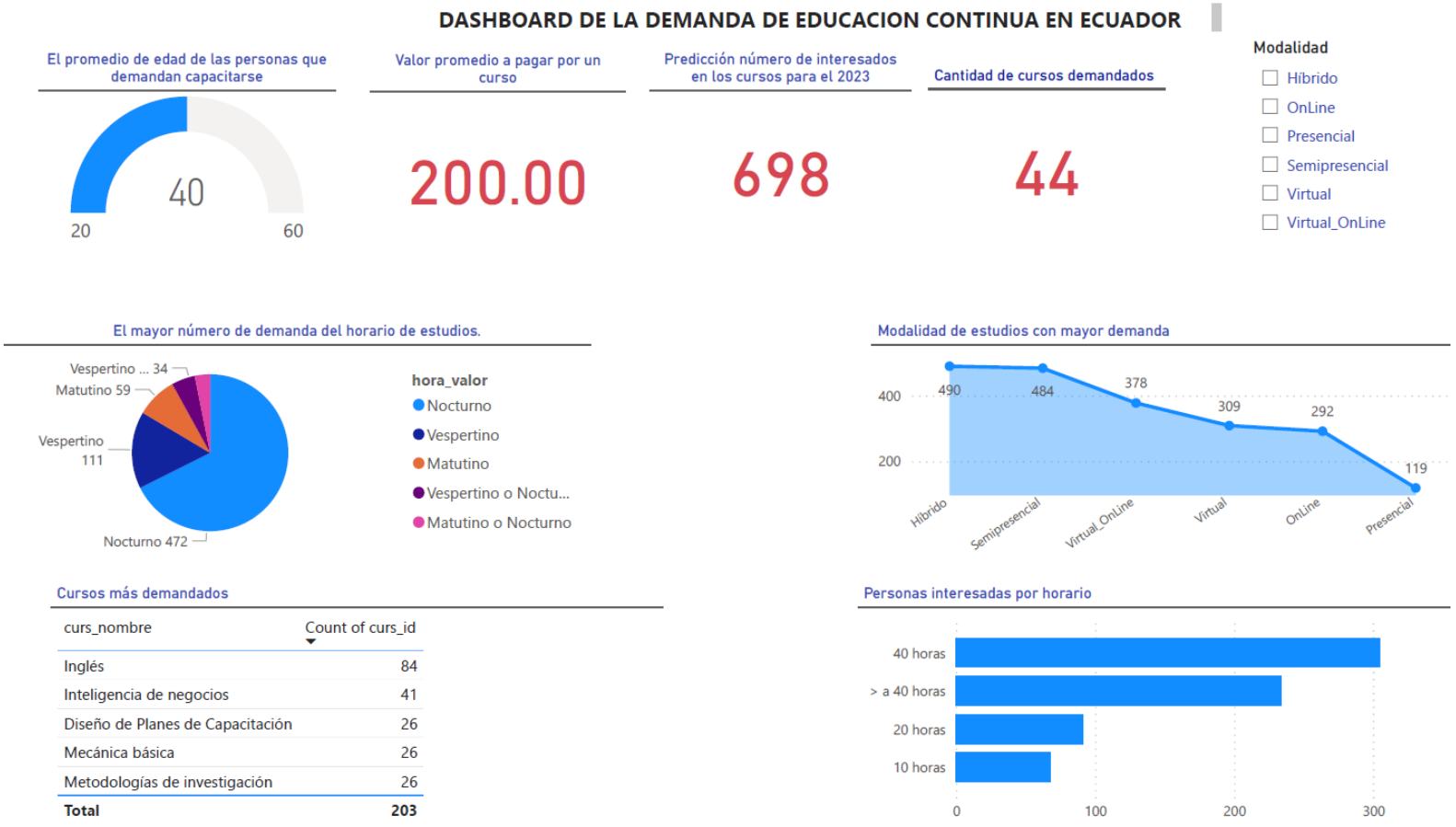


Figura 46. DashBoard para la demanda

En el DashBoard que hace referencia a la Oferta, podemos observar que el promedio de la duración en horas es de 50h, mientras que la duración máxima es de 100 horas.

Las modalidades de estudio con mayor oferta de cursos son Presencial seguido por Virtual y Online respectivamente, podemos deducir que antes de la pandemia del COVID-19 las ofertas de capacitación le daban mayor importancia a la presencialidad, con un incremento en modo Virtual y Online. Si realizamos una comparación con las modalidades de estudio que demanda en la actualidad las personas vemos que aparece la modalidad Hibrida seguido por la modalidad semipresencial que aun despierta el interés por parte de las personas; mientras que la modalidad presencial pasa a ocupar el último puesto, esto nos indica que poco a poco las personas necesitan acceder a cursos en línea.

La modalidad semipresencial tiene un promedio de precio alto con relación a las demás modalidades esto tiene sentido pues el promedio de duración de un curso en dicha modalidad también su duración es elevada.

Las personas tienen preferencia por horarios de capacitación nocturnos y vespertinos, lo que nos indica que son personas oficinistas que desean acceder a un curso luego de su jornada laboral.

Las áreas de estudio con mayor oferta de capacitación son las empresariales y tecnológicos, mientras que los cursos con mayor demanda lo lideran inglés, seguido por Inteligencia de negocios, tal como indica la **Figura 47.**

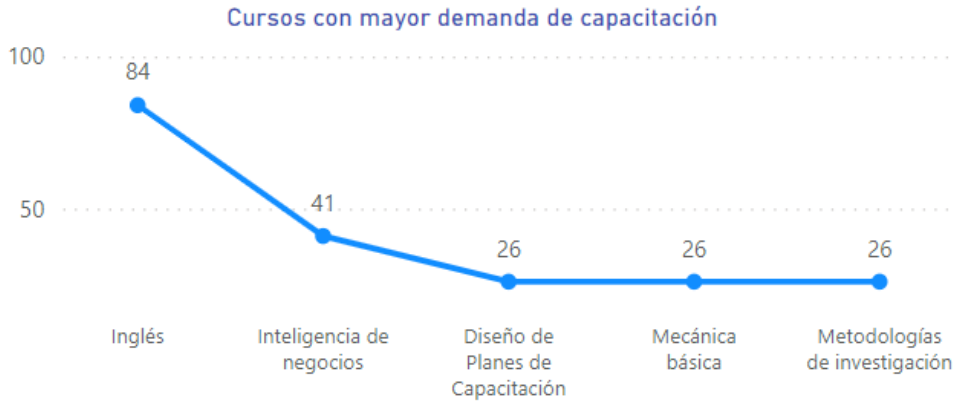


Figura 47. Cursos con mayor demanda de capacitación.

La edad de las personas que desean acceder a un curso de educación continua está entre los 20 y 60 años y su promedio de edad es de 40 años, este indicador nos puede servir para poder enfocar de una mejor forma la publicidad, su tipo y contenido.

El valor promedio que estaría dispuesto a pagar una persona por un curso seria de 200 dólares, habría que considerar si ese valor va en función de su duración.

Para el año 2023 se estima que 698 personas estarían interesadas en seguir algún curso ofertado, mientras que 44 son los cursos que las personas están demandando.

3.1. Evaluación de usabilidad.

Se aplico la prueba de usabilidad SUS a 6 personas con un mediano conocimiento de tecnologías de la información, la evaluación fue desarrollada con un formulario de Google Forms [18].

La Escala de Usabilidad del Sistema (SUS) es el cuestionario estandarizado más utilizado para la evaluación de usabilidad percibida [19], mediante un cuestionario estandarizado compuesto por 10 preguntas asociado a 5 posibles respuestas para cada pregunta (Totalmente en desacuerdo, en desacuerdo, neutro, de acuerdo, totalmente de acuerdo).

3.2. Resultados de la evaluación.

La **Figura 48**, nos muestra los resultados de la evaluación de cada pregunta.

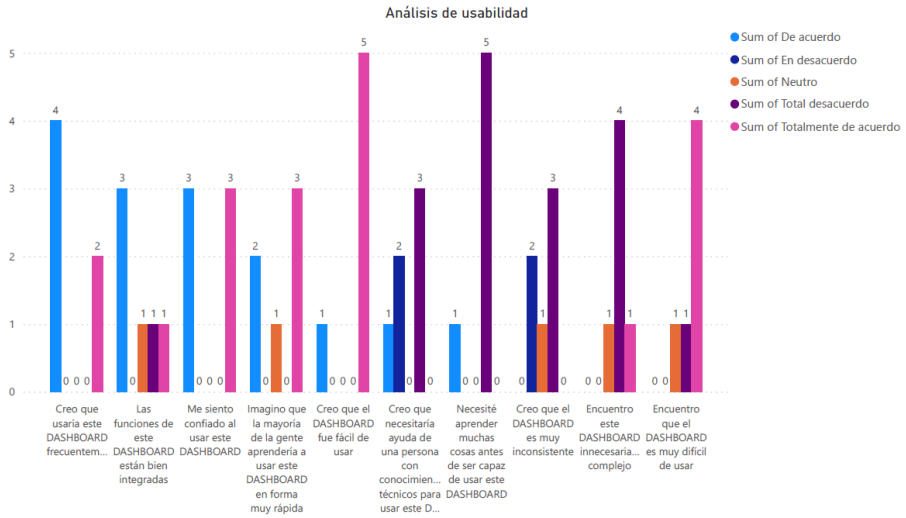


Figura 48. Resultados evaluación SUS

El resultado de la evaluación SUS es **76.25**

4. CONCLUSIONES Y RECOMENDACIONES.

4.1. Conclusiones.

- Para el objetivo general: Realizar un análisis de la oferta y demanda de la educación continua en Ecuador, mediante indicadores estáticos y dinámicos y con el diseño e implementación de un BI.
    - Conclusión 1: Se logro diseñar los DashBoards para la oferta y demanda donde se muestra los valores de los indicadores estáticos y dinámicos.
- Para el objetivo: Obtener la información de demanda y oferta en educación continua.
    - Conclusión 2: Se pudo tener información más precisa de lo que las instituciones educativas ofertan, así como cuál es la demanda de capacitación que las personas buscan, esto lo plasmamos en los DashBoards (sección 3).
- Para el objetivo: Diseñar el DW considerando el proceso ETL extracción, transformación y carga.
    - Conclusión 3: Con la ayuda de RapidMiner se logró crear y ejecutar los procesos ETL que permitieron la carga de la información en el DW (sección 2.4).
- Para el objetivo: Diseñar los tableros de información en base a los requerimientos del negocio.
    - Conclusión 4: Se crearon dos DashBoards (sección 3) que muestran la información de los indicadores que responden a las preguntas definidas (sección 2.1.1).
- Para el objetivo: Definir indicadores dinámicos y estáticos asociados a la oferta y demanda.
    - Conclusión 5: La generación de modelos de machine learning mediante la herramienta RapidMiner, nos permitió elaborar los indicadores dinámicos en

base a la evaluación de varios modelos, donde se consideró criterios de peso de variables y porcentaje de error relativo proporcionado por RapidMiner (sección 2.4.1)

4.1.1. Evaluación de usabilidad.

Al realizar la evaluación SUS, los resultados nos indican que la visualización es aceptable pero no está en el rango de excelente por lo que sería necesario volver a revisar su funcionalidad.

4.2. Recomendaciones.

- Contar con un acceso más directo a la oferta de capacitación continua de las instituciones educativas permitiría acelerar el tiempo de desarrollo del proyecto de DW.
- Se debería contar con métricas referentes a la promoción de los cursos, esto ayudaría a contar con información más amplia a la hora de diseñar un algoritmo predictivo para la oferta.
- El poder contar con un historial de la oferta de capacitación con precios y duración permitirían generar algoritmos del tipo Forecasting para pronósticos basados en series de tiempo.
- La encuesta que mide la demanda de capacitación se debería aplicar a un mayor número de personas para de esta manera contar con más datos a la hora de entrenar los algoritmos predictivos.
- Se debería crear una función o clase Python que permita consolidar las acciones de limpieza y calidad de datos en una sola unidad.
- Para los modelos predictivos se podría utilizar una herramienta OpenSource como lo es Python.

- Se podría buscar alternativas de herramientas para visualización de datos OpenSource en lugar de ocupar PowerBI, pues para un ambiente empresarial se necesita una cuenta comercial pagada.

REFERENCIAS BIBLIOGRAFICAS.

[1] Watson, H. J. (2009). Tutorial: Business Intelligence – Past, Present, and Future. Communications of the Association for Information Systems, 25, pp 6. https://doi.org/10.17705/1CAIS.02539.

[2] NOSQL databases. http://nosql-database.org/. Accedido febrero 09, 2023

[3] European Knowledge Center for Information Technology (Ed.). (2021, 28 marzo). Integración de datos (Data Integration). Consultado el 4 de marzo de 2023, TIC Portal. https://www.ticportal.es/glosario-tic/integracion-datos

[4] Rizzi, S., Abelló, A., Lechtenbörger, J., & Trujillo, J. (2006). Research in data warehouse modeling and design. Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP - DOLAP '06. doi:10.1145/1183512.1183515

[5] Skyrius, R. (2021). Business Intelligence. Progress in IS. doi:10.1007/978-3-030-67032-0

[6] Silva Peñafiel, G. E., Zapata Yánez, V. M., Morales Guamán, K. P., & Toaquiza Padilla, L. M. (2019). Análisis de metodologías para desarrollar Data Warehouse aplicado a la toma de decisiones. Ciencia Digital, 3(3.4.), 397-418. https://doi.org/10.33262/cienciadigital.v3i3.4.922

[7] DATA WAREHOUSING: Investigación y Sistematización de Conceptos (pp. 87–88). (2010). Córdoba, Argentina: Ing. Bernabeu Ricardo Dario. Córdoba, Argentina: Ing. Bernabeu Ricardo Dario.

[8] Universidades y escuelas politécnicas. (2022). Recuperado 14 de noviembre de 2022, de CES Consejo de Educación Superior website: https://www.ces.gob.ec/?page_id=326

[9] Midway, S. R. (2020). Principles of Effective Data Visualization. Patterns, 1(9), 100141. doi:10.1016/j.patter.2020.100141

[10] J. Galindo-Losada, N. Arcos, M. Carrión-Toro and M. Santórum, "SIMPA -design: A System of Indicators for Monitoring the Learning Process," 2023 IEEE World Engineering Education Conference (EDUNINE), Bogota, Colombia, 2023, pp. 1-6, doi: 10.1109/EDUNINE57531.2023.10102823.

[11] G. van Rossum, An Introduction to Python. Bristol: Network Theory Limited, 2003.

[12] "RapidMiner | Amplify the Impact of Your People, Expertise & Data". RapidMiner. https://rapidminer.com/ (accedido el 12 de junio de 2023).

[13] P. S. Diouf, A. Boly and S. Ndiaye, "Variety of data in the ETL processes in the cloud: State of the art," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 1-5, doi: 10.1109/ICIRD.2018.8376308.

[14] Noa Roy-Hubara, Arnon Sturm, Peretz Shoval,Designing NoSQL databases based on multiple requirement views,Data & Knowledge Engineering,Volume 145,2023,102149, ISSN 0169-023X, https://doi.org/10.1016/j.datak.2023.102149. (https://www.sciencedirect.com/science/article/pii/S0169023X23000095).

[15] Dalla Valle, L., & Kenett, R. (2018). Social media big data integration: A new approach based on calibration. Expert Systems with Applications, 111, 76–90. doi:10.1016/j.eswa.2017.12.044.

[16] "PostgreSQL". PostgreSQL. https://www.postgresql.org/ (accedido el 29 de junio de 2023).

[17] "Power BI". Power BI. https://app.powerbi.com/home?experience=power-bi (accedido el 29 de junio de 2023).

[18] "Google Forms: Sign-in". Sign in - Google Accounts. https://docs.google.com/forms/u/0/ (accedido el 29 de junio de 2023).

[19] Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. International Journal of Human–Computer Interaction, 34(7), 577–590. doi:10.1080/10447318.2018.1455307.

[20] "Google Forms: Sign-in". Sign in - Google Accounts. https://docs.google.com/forms/u/0/ (accedido el 3 de junio de 2023).

ANEXOS.

Anexo 1. Código Python para crear archivos csv con información de las columnas para ser consumidos con los procesos ETL.

Lectura del archivo que contiene la información **Figura 49**.

```
In [1]: import numpy as np
        import pandas as pd

        Universidades

In [7]: df=pd.read_excel('costos Oferta de Educación continua.xlsx',sheet_name='oferta')
        df.head()
```

Out[7]:

	Universidad	Tipos de Oferta	Areas de Formación
0	udla universidad de las americas	Maestrias	Administracion y negocios
1	NaN	NaN	Educacion
2	NaN	NaN	Salud
3	NaN	Diplomados	Administracion y negocios
4	NaN	NaN	Tecnologia e informacion

```
In [22]: df.shape
```

Out[22]: (89, 3)

```
In [29]: dfu=pd.DataFrame(df['Universidad'].value_counts(dropna=False)).reset_index()
         dfu.rename(columns={"index":'Universidad',"Universidad":'Cantidad'},inplace=True)
         dfu.head()
```

Figura 49. Lectura del archivo xls

Creación del csv con información de los tipos de oferta, **Figura 50**.

TIPO OFERTAS

```
In [8]: dfto=pd.DataFrame(df['Tipos de Oferta'].value_counts(dropna=False)).reset_index()
        dfto.rename(columns={"index":'Tipos de Oferta',"Tipos de Oferta":'Cantidad'},inplace=True)
        dfto['Tipos de Oferta'] = dfto['Tipos de Oferta'].str.upper()
        dfto.head()
```

Out[8]:

	Tipos de Oferta	Cantidad
0	NaN	62
1	CURSOS	3
2	MAESTRIAS	1
3	MAESTRIA	1
4	CURSOS PRESENCIAL	1

```
In [9]: dfto.to_csv('TiposOfertas.csv')
```

Figura 50. Creación csv de tipos oferta.

Creación del csv con información del área de formación, **Figura 51**.

AREA DE FORMACION

```
In [ ]: df=pd.read_excel('costos Oferta de Educación continua.xlsx',sheet_name='resumen')
        df.head()
```

```
In [32]: #dfaf=pd.DataFrame(df['Areas de Formación'].value_counts(dropna=False)).reset_index()
         dfaf=pd.DataFrame(df['Areas de Formación'])
         #dfaf.rename(columns={"index":'Areas de Formación',"Areas de Formación"},inplace=True)
         dfaf['Areas de Formación'] = dfaf['Areas de Formación'].str.strip()
         dfaf['Areas de Formación'] = dfaf['Areas de Formación'].str.upper()
         dfaf = dfaf.drop_duplicates()
         dfaf.head()
```

Out[32]:

	Areas de Formación
0	ADMINISTRACION Y NEGOCIOS
1	EDUCACION
2	SALUD
9	TECNOLOGIA E INFORMACION
25	CERTIFICACIONES

```
In [21]: dfaf.to_csv('AreasFormacion.csv')
```

Figura 51. Creación csv de las áreas de formación.

Tratamiento de los días de capacitación demandados, **Figura 52.**

DIAS

```
In [39]: df['dias_tratados']=df['¿Que días son mas cómodos para usted tomar este tipo de programas de formación de educación continua
         df['dias_tratados'].value_counts()
```

Out[39]:

Sábado;Domingo	16
Viernes;Sábado	15
Sábado	14
Lunes;Martes;Miercoles;Jueves;Viernes	8
Lunes;Miercoles;Viernes	6
Lunes;Martes;Miercoles;Jueves	5
Martes;Jueves;Sábado	3
Viernes	3
Miercoles	3
Jueves;Viernes;Sábado	3
Martes	2
Viernes;Sábado;Domingo	2
Lunes;Miercoles	2
Lunes;Martes;Miercoles	2

```
In [40]: # Create a function to classify each day
         dias_init=df["dias_tratados"].value_counts().index.tolist()
         def classify_day(day):
             if day in ['Sábado', 'Domingo']:
                 return 1 #Fin de semana
             else:
                 return 0 #Entre Semana

         # Create the original Series with different days
         series=pd.Series(dias_init)
         # Utiliza el metodo classify_day para separar días de fin de semana y entre semana
         sub_classification = series.apply(lambda x: [classify_day(day) for day in x.split(';')])
```

```
In [41]: cont1=0 #Fin de Semana
         cont0=0 #Entre Semana
         semana=[]
         for i in sub_classification:
             cont1=0;cont0=0
             if len(i)>1: #si el registro tiene más de 1 día
                 for j in i: #recorre todos los días que están asignados por 0 o 1
                     if j==1:
                         cont1=cont1+1 #si es 1, cuenta como día de fin de semana
                     if j==0:
                         cont0=cont0+1 #si es 0, cuenta como día de fin de semana
                 if (cont1>=1): #si hay días de fin de semana
                     if(cont0>=1): #evaluamos si también contiene días entre semana
                         semana.append('TodaSemana')
                     else: #sino tiene entonces solo es fin de semana
                         semana.append('FinSemana')
                 else: #evaluamos si no contiene días de fines de semana
                     if(cont0>=1):#evaluamos si solo tiene días de entre semana
                         semana.append('EntreSemana')
             else:#si el registro solo tiene un día entre semana, vemos si es de fin o entre semana
                 if i==[1]:
                     semana.append('FinSemana')
                 if i==[0]:
                     semana.append('EntreSemana')
```

Figura 52. Tratamiento de los días de capacitación