

ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

SISTEMA DE MINERÍA DE DATOS DEL PORTAL WEB SETEC PARA ANÁLISIS DE OFERTA Y DEMANDA AL DEFINIR INDICADORES ESTÁTICOS Y DINÁMICOS

PROCESO PARA LA OBTENCIÓN DE TÍTULO DE INGENIERO CIENCIAS DE LA COMPUTACIÓN

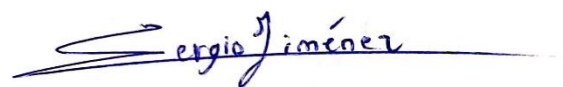
SERGIO ANDRÉS JIMÉNEZ REINO

DIRECTOR: PHD. JULIÁN GALINDO

DMQ, marzo 2023

CERTIFICACIONES

Yo, Sergio Andrés Jiménez Reino declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A handwritten signature in blue ink that reads "Sergio Jiménez". The signature is stylized with a large, sweeping initial "S" and a long horizontal line extending to the right.

Sergio Jiménez

Certifico que el presente trabajo de integración curricular fue desarrollado por Sergio Jiménez, bajo mi supervisión.

Julián Galindo
DIRECTOR

Certificamos que revisamos el presente trabajo de integración curricular.

NOMBRE_REVISOR1
REVISOR1 DEL TRABAJO DE
INTEGRACIÓN CURRICULAR

NOMBRE_REVISOR2
REVISOR2 DEL TRABAJO DE
INTEGRACIÓN CURRICULAR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

Sergio Jiménez

PhD. Julián Galindo

DEDICATORIA

El presente documento está dirigido a mi familia quienes me han apoyado incondicionalmente en los todos los aspectos de mi vida académica y profesional.

AGRADECIMIENTO

A mi familia quienes son lo más importante para mí:

A mis padres Sergio Jiménez y Bibiana Reino por todo lo que han hecho por mí, me han enseñado y me han hecho la persona que soy hoy en día.

A mis hermanos Juan José y María de Lourdes de quienes he aprendido y me han inspirado en ser mejor cada día.

A profesores quienes me enseñaron y apoyaron en distintos momentos de la carrera universitaria:

Al profesor Roberto Andrade, Denys Flores, Andrés Merino, Sang Wo, Henry Paz y Julián Galindo por sus enseñanzas y recomendaciones.

A mis amigos y compañeros con quienes compartí mi vida universitaria:

Adrián Laje quien ha sido mi compañero y amigo durante el inicio y fin de la carrera universitaria incluso prepo.

Tannya Rosero quien me apoyo y animo durante el desarrollo de este proyecto.

Daliana, Luis, Bladimir, Danny y Anthony, compañeros y amigos con quienes he trabajado y estudiado en todo el transcurso de la carrera en ciencias de la computación.

Y mis amigos Alejandra, Yadira, Johan, Rasu, Kevin, Boris, Henry, Kevin, Lauro, Wilman, Jeremy, David, Josselyn, Toa, Elisabeth, Yajaira, Gabriel, David, Wilson, Ignacio y Cristian.

ÍNDICE DE CONTENIDO

CERTIFICACIONES.....	I
DECLARACIÓN DE AUTORÍA.....	II
DEDICATORIA.....	III
AGRADECIMIENTO.....	IV
ÍNDICE DE CONTENIDO.....	V
RESUMEN	VIII
ABSTRACT	IX
1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO.....	1
1.1 Objetivo general.....	1
1.2 Objetivos específicos	1
1.3 Alcance	2
1.4 Marco teórico	3
1.4.1 Minería de Datos.....	4
1.4.2 Inteligencia de negocios:.....	4
1.4.3 CRISP-DM	4
1.4.4 Microsoft PowerBI	4
1.4.5 RapidMiner	4
1.4.6 Python	5
1.4.7 Selenium.....	5
1.4.8 Entornos de desarrollo	5
1.4.9 SQL	5
1.4.10 MySQL Workbech:.....	6
1.4.11 Web Scraping	6
1.4.12 Datawarehouse:.....	6
1.4.13 KPIs	6

1.4.14	Criterios SMART	6
2	METODOLOGÍA	7
2.1.1	Fase 1: Entendimiento del negocio.....	8
2.1.2	Fase 2: Entendimiento de Datos.....	9
2.1.3	Fase 3: Preparación de datos.....	10
2.1.4	Fase 4: Modelamiento	10
2.1.5	Fase 5: Evaluación	11
2.1.6	Fase 6: Despliegue.....	11
2.1.7	Fase 7: Visualización.....	12
3	DESARROLLO E IMPLEMENTACIÓN.....	13
3.1	Entendimiento del negocio:	14
3.1.1	Análisis y entendimiento de los módulos del portal web SETEC.....	14
3.1.2	Análisis y entendimiento del mercado laboral nacional.....	15
3.1.3	Diseño de objetivos de negocio e indicadores estáticos y dinámicos.....	16
3.1.4	Producción del plan de proyecto.....	17
3.2	Entendimiento de datos:.....	19
3.2.1	Análisis y entendimiento de los datos en los módulos del portal web SETEC 20	
3.2.2	Extracción y colección de datos.....	23
3.2.3	Descripción de datos, Análisis exploratorio y verificación de calidad de datos 27	
3.3	Preparación de datos:	29
3.4	Modelamiento y Evaluación:	33
3.5	Despliegue:	42
3.6	Visualización:	42
3.7	Evaluación de usabilidad:.....	44
3.7.1	Resultados de Nielsen.....	44
3.7.2	Resultados de SUS	45

4	ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS	46
5	CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS.....	47
5.1	Conclusiones.....	47
5.1.1	Comprensión del negocio:	47
5.1.2	Comprensión de Datos:	47
5.1.3	Preparación de datos:.....	47
5.1.4	Modelamiento:	48
5.1.5	Evaluación:	48
5.1.6	Despliegue:	48
5.1.7	Visualización:.....	48
5.1.8	Evaluación de usabilidad:	47
5.2	Recomendaciones.....	48
5.3	Trabajos futuros	49
6	REFERENCIAS BIBLIOGRÁFICAS	50
7	ANEXOS.....	52

RESUMEN

El proyecto consistió en el desarrollo de un sistema de minería de datos del portal web de la Secretaría Nacional de Cualificaciones y Capacitación Profesional (SETEC) para el análisis de la demanda de cursos y/o perfiles de profesionales cualificados a nivel nacional. El proyecto fue desarrollado usando metodología de minería de datos CRISP-DM con las fases de entendimiento del negocio para el análisis de los módulos del portal web SETEC, el entendimiento de datos que conforman cada módulo del portal web, preparación de datos, extracción de datos usando técnicas de raspado web, limpieza de datos, transformación de tipo de variables y colección de estos para la creación de un Datawarehouse. Posteriormente se realizó el análisis de la demanda de estos cursos y perfiles en base a la definición de indicadores estáticos en base a datos reales e indicadores dinámicos desarrollados a través de distintos modelos de aprendizaje automático. El resultado de final estas fases desplegaron en la herramienta Microsoft PowerBI por medio de distintos gráficos y componentes visuales que muestran los indicadores además y su variación con las distintas dimensiones de datos. La usabilidad de aplicación visual fue evaluada en base a SUS (System Usability Scale) y las 10 heurísticas de Nielsen de usabilidad, donde 40 personas entre expertos en áreas de datos y conocimientos moderados en análisis de datos y tecnologías de la información puntuaron la aplicación concluyendo en que se trata de una aplicación con muy pocos problemas estéticos y un rango de usabilidad bueno.

PALABRAS CLAVE: Web Scraping, CRISP-DM, SETEC, Data Analysis, Datawarehouse, PowerBI, R, R Studio, SQL, RapidMiner, Python, Usability, SUS, Nielsen, ETL, EDA.

ABSTRACT

The project consisted of developing a data mining system for the National Secretariat of Qualifications and Professional Training (SETEC) web portal to analyze the demand for courses and/or profiles of qualified professionals at a national level. The project was developed using CRISP-DM data mining methodology with the business understanding phase for the analysis of the SETEC web portal modules, understanding of data that make up each module of the web portal, data preparation, data extraction using web scraping techniques, data cleaning, transformation of variable types, and collection of these to create a data warehouse. Subsequently, the demand for these courses and profiles was analyzed based on the definition of static indicators based on real data and dynamic indicators developed through various machine learning models. The result of these phases was displayed in the Microsoft PowerBI tool through various graphics and visual components that show the indicators and their variation with different data dimensions. The usability of the visual application was evaluated based on the System Usability Scale (SUS) and Nielsen's 10 usability heuristics, where 40 people, including experts in data areas and those with moderate knowledge of data analysis and information technologies, scored the application. They concluded that it is an application with very few aesthetic problems and a good usability range.

KEYWORDS: Web Scraping, CRISP-DM, SETEC, Data Analysis, Datawarehouse, PowerBI, R, R Studio, SQL, RapidMiner, Python, Usability, SUS, Nielsen, ETL, EDA.

1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

El presente trabajo consiste en el desarrollo e implementación de un sistema de minería de datos de la plataforma web SETEC para el análisis de la demanda de cursos y perfiles de cualificación mediante indicadores estáticos y dinámicos utilizando la metodología de CRISP-DM más fases de visualización y evaluación de usabilidad.

1.1 Objetivo general

Desarrollar un sistema de minería de datos para el análisis de la demanda de cursos y perfiles del portal web de la SETEC utilizando indicadores estáticos y dinámicos.

1.2 Objetivos específicos

Los objetivos específicos se detallan en base a la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) la cual es utilizada para la construcción de sistemas de minería de datos con fases típicas de proyecto, las tareas involucradas en cada fase y una explicación de las relaciones entre estas tareas [1]. Estas fases serán:

1. **Entendimiento del negocio:** Análisis y entendimiento de los módulos del portal web SETEC, diseño de objetivos
2. **Entendimiento de datos:** Análisis y entendimiento de los datos en los módulos del portal web SETEC
3. **Preparación de datos:** Extracción, selección, limpieza, transformación, construcción e integración de datos del portal web
4. **Modelamiento:** Selección de técnicas de modelización, generación de un diseño de pruebas, construcción y aceptación de modelos para el análisis de la demanda en base a los datos preparados.
5. **Evaluación:** Evaluación de los modelos desarrollados en base a los objetivos planteados y revisión de actividades, decisiones de las anteriores fases y determinación de siguientes pasos a realizar.
6. **Despliegue:** Diseño de plan de despliegue de los modelos, monitorización y mantenimiento del sistema de minería de datos desarrollado

El sistema tendrá dos fases adicionales que son:

7. **Visualización:** visualización de los resultados a través de tableros

8. **Evaluación de usabilidad:** análisis de la usabilidad del sistema de minería de datos desarrollado en base a los principios de Nielsen y SUS.

1.3 Alcance

La metodología CRISP-DM se la utilizará para el diseño, construcción y evaluación de un sistema de minería de datos en base a 6 fases más dos fases adicionales de visualización y evaluación de usabilidad que se detallan a continuación para la creación del Sistema de minería de datos de SETEC.

1. Entendimiento del negocio:

- 1.1. Análisis y entendimiento de los módulos del portal web SETEC

- 1.2. Análisis y entendimiento del mercado laboral nacional

- 1.3. Diseño de objetivos de negocio e indicadores estáticos y dinámicos

- 1.4. Producción del plan de proyecto

2. Entendimiento de datos:

- 2.1. Análisis y entendimiento de los datos en los módulos del portal web SETEC

- 2.2. Extracción y colección de datos

- 2.3. Descripción de datos

- 2.4. Análisis exploratorio de datos

- 2.5. Verificación de la calidad de datos

3. Preparación de datos:

- 3.1. Selección de datos

- 3.2. Limpieza de datos

- 3.3. Transformación de datos

- 3.4. Formateo de datos

- 3.5. Construcción de nuevos datos

- 3.6. Integración de datos

- 3.7. Diseño del Datawarehouse

- 3.8. Almacenamiento de datos

4. Modelamiento:

4.1. Selección de técnicas de modelamiento

4.2. Generación de diseño de pruebas

4.3. Construcción de modelos

4.4. Evaluación de modelos

5. Evaluación:

5.1. Evaluación del resultado de los modelos desarrollados en base a los objetivos planteados

5.2. Carga de información predictiva en el Datawarehouse

5.3. Revisión de las actividades y decisiones tomadas en las anteriores fases

5.4. Determinación de los siguientes pasos

6. Despliegue:

6.1. Diseño de plan de despliegue de modelos

6.2. Plan de monitorización y mantenimiento

6.3. Despliegue de los modelos de en herramientas de visualización de datos

El sistema tendrá dos fases adicionales a las de CRISP-DM, que son:

7. Visualización:

7.1. Selección de gráficos para visualización de datos y resultados

7.2. Creación de gráficos estáticos y dinámicos

7.3. Creación de tableros de gráficos explicativos de los datos y resultados

8. Evaluación de usabilidad:

8.1. Evaluación de la usabilidad por medio de las métricas de usabilidad de Nielsen.

8.2. Evaluación de la usabilidad por medio de SUS.

1.4 Marco teórico

1.4.1 Minería de Datos

Es una disciplina que utiliza técnicas de matemáticas, estadísticas y ciencias de la computación para analizar grandes conjuntos de datos y obtener conocimientos valiosos. Utiliza herramientas de análisis de datos, como el aprendizaje automático y el procesamiento de lenguaje natural, para descubrir patrones y tendencias, y hacer predicciones basadas en la información. [2]

1.4.2 Inteligencia de negocios:

La inteligencia de negocios es un conjunto de técnicas, herramientas y sistemas que se utilizan para recopilar, integrar, analizar y presentar información empresarial con el fin de mejorar la toma de decisiones y el rendimiento de la organización. [3]

1.4.3 CRISP-DM

CRISP-DM es un modelo de proceso para el análisis y el desarrollo de proyectos de inteligencia de negocios y minería de datos. CRISP-DM es el acrónimo de Cross-Industry Standard Process for Data Mining. [1] Es ampliamente utilizado en el proceso de análisis de datos y la toma de decisiones basadas en datos al ser un modelo flexible y práctico que incluye seis fases:

1. Entendimiento del negocio.
2. Entendimiento de datos
3. Preparación de datos
4. Modelamiento
5. Evaluación
6. Implementación

1.4.4 Microsoft PowerBI

PowerBI es una plataforma de inteligencia empresarial (BI) que se utiliza para conectar, modelar y visualizar datos y crear informes personalizados y gráficos con los KPI clave. Además, permite obtener respuestas a preguntas comerciales utilizando inteligencia artificial para brindar resultados rápidos y precisos. [4]

1.4.5 RapidMiner

Plataforma de ciencia de datos que ayuda a acelerar los proyectos de ciencia de datos y mejorar la competitividad empresarial permitiendo a los usuarios de diferentes roles, como expertos en ciencia de datos, expertos en dominios, líderes y personal de TI, trabajar juntos

en proyectos de ciencia de datos y obtener información valiosa de los datos de un negocio. [5] La versión utilizada en el presente proyecto fue la académica que es de uso limitado y con una licencia de un año.

1.4.6 Python

Python es un lenguaje de programación de alto nivel e interpretado desarrollado en la década de 1990 por Guido Van Rossum. Se utiliza ampliamente en el desarrollo de aplicaciones de software, análisis de datos y aprendizaje automático gracias a su gran cantidad de bibliotecas y marcos de desarrollo disponibles que facilitan el proceso de desarrollo. [6]

1.4.7 Selenium

Es un conjunto de herramientas y APIs utilizado para la automatización de pruebas en sitios web y aplicaciones móviles. Incluye varias herramientas, como WebDriver, que utiliza las APIs de automatización del navegador para controlarlo y ejecutar pruebas y tareas automatizadas como la extracción de datos en procesos de ciencia de datos. Selenium cuenta con librerías creadas para diferentes entornos y lenguajes de programación, como Python. [7]

1.4.8 Entornos de desarrollo

Jupyter Lab

Es un proyecto de código abierto y sin fines de lucro que nació del proyecto IPython en 2014. Proporciona una plataforma interactiva para la ciencia de datos y la computación científica en múltiples lenguajes de programación. [8]

Visual Studio Code

Visual Studio Code (VS Code) es un editor de código fuente y un entorno de desarrollo integrado (IDE, por sus siglas en inglés) desarrollado por Microsoft. Es una aplicación de escritorio gratuita y de código abierto que se utiliza para escribir, depurar y ejecutar código en una amplia variedad de lenguajes de programación. [9]

1.4.9 SQL

SQL (Structured Query Language) es un lenguaje de programación utilizado para interactuar con bases de datos relacionales. SQL permite crear, modificar y eliminar bases de datos, así como recuperar y manipular datos almacenados en ellas. [10]

1.4.10 MySQL Workbench:

Es una herramienta que permite a arquitectos de bases de datos, desarrolladores y administradores de bases de datos realizar tareas visualmente. Ofrece funciones completas para modelar datos, desarrollar consultas SQL y administrar servidores: incluyendo la configuración del servidor, la gestión de usuarios y la realización de copias de seguridad. [11]

1.4.11 Web Scraping

Técnica de extracción de datos que se utiliza para recopilar información de sitios web de forma automatizada. Esta técnica implica el uso de software o herramientas para extraer datos de páginas web, analizarlos y transformarlos en un formato que pueda ser utilizado para diferentes fines, como el análisis de datos o la creación de nuevas aplicaciones. [12]

1.4.12 Datawarehouse:

Un Datawarehouse o almacenamiento de datos consiste en un repositorio central de información que permite un mejor análisis en la toma de decisiones. Los datos que lo componen son de sistemas transaccionales, bases de datos relacionales, bases de datos no relaciones, documentos, sitios web, entre otros. [13]

1.4.13 KPIs

KPI es el acrónimo de Key Performance Indicator que se traduce al español como Indicador Clave de Desempeño. Un KPI es una métrica o medida utilizada para evaluar el rendimiento o el progreso de una empresa o proyecto en relación con objetivos establecidos. [14]

1.4.1 Tipos de indicador

Estático (S): se refiere a descriptivo análisis, que es el examen de datos o contenido, generalmente realizado manualmente y sin operaciones de actualización o estimación de valores futuros, para responder a la pregunta “¿Qué sucedió?” (o ¿Qué está pasando?)” [15]

Dinámico (D): explorar el análisis predictivo ya sea a través del tiempo como una forma de análisis avanzado que examina datos o contenido para responder a la pregunta, “¿Qué es probable que suceda?” [15]

1.4.14 Criterios SMART

Propuestos por George T. Doran para la definición de indicadores clave de desempeño vienen del Specific, Measurable, Achievable, Realistic y Time-bound. [16] Los criterios SMART son un acrónimo que representa los siguientes aspectos:

1. Específico (Specific): Un KPI debe ser claro y detallado, sin ambigüedades, para que todos los involucrados sepan exactamente qué se está midiendo.
2. Medible (Measurable): El indicador debe ser cuantificable y se debe poder medir de manera objetiva y precisa.
3. Alcanzable (Achievable): El KPI debe ser realista y posible de alcanzar en función de las capacidades, recursos y limitaciones de la organización.
4. Realista (Realistic): El KPI debe ser relevante y estar alineado con los objetivos y estrategias de la organización.
5. Limitado en el tiempo (Time-bound): El KPI debe tener un plazo definido para su cumplimiento, ya sea a corto, mediano o largo plazo.

2 METODOLOGÍA

CRISP-DM + Visualización + Evaluación de Usabilidad:

Metodología de minería de datos cuyas siglas significan **C**ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining. [1] Esta cuenta con 6 fases que se interrelacionan entre sí en el ciclo de vida de minería de datos, además se incluyeron fases de Visualización y Evaluación de Usabilidad como se observa en la **Figura 2.1**. Cada fase cuenta con subprocesos como se detalla a continuación:

que se llevarán a cabo, los recursos necesarios, posibles riesgos asociados, los plazos, los objetivos y los criterios de éxito. [1]

2.1.2 Fase 2: Entendimiento de Datos

Implica observar más de cerca los datos disponibles para la minería y comprender su calidad y estructura. El objetivo de esta fase es identificar cualquier problema o limitación en los datos que puedan afectar la calidad de los resultados del proyecto. [1] Esta implica:

Recopilación de datos iniciales: Los datos pueden venir de una variedad de recursos como lo son:

- Datos existentes: Esto abarca una gran diversidad de información, como por ejemplo datos de transacciones, información obtenida de encuestas, registros en la web, y otros tipos de datos similares.
- Datos comprados: Datos complementarios como datos demográficos
- Datos adicionales: Datos externos a los que ya se cuenta como realizar encuestas, análisis de mercado, o seguimiento de almacenes de datos ya existentes.

Este proceso deberá permitir identificar los atributos más prometedores y los que son irrelevantes, evaluar si hay suficientes datos para hacer predicciones precisas, determinar si hay demasiados atributos para el método de modelado elegido, analizar la fusión de varias fuentes de datos y considerar cómo se manejan los valores faltantes en cada una de ellas. [1]

Descripción de datos: se enfoca en determinar la calidad y cantidad de datos recolectados al considerar el tamaño y los campos de los conjuntos de datos, así como los diferentes formatos de valores, para evitar problemas durante el modelado posterior.

Exploración de datos: implica el uso de tablas, gráficos y otras herramientas de visualización para analizar los datos y abordar los objetivos de minería de datos establecidos además de formulación de hipótesis y formular las posibles transformaciones de datos en la fase de preparación.

Verificación de calidad de datos: Se verifica si existen errores de codificación, valores faltantes, errores de datos, errores de medición, inconsistencias en la codificación y metadatos incorrectos. Los errores comunes son: los valores faltantes, errores tipográficos, esquemas de medición incorrectos, unidades de medida no estándar, inconsistencias de valores y discrepancias en los metadatos. [1]

2.1.3 Fase 3: Preparación de datos

Fase crucial que puede ocupar entre el 50% y el 70% del tiempo y esfuerzo de un proyecto de minería de datos. Se debe desarrollar correctamente las fases de comprensión del negocio y comprensión de los datos para minimizar el tiempo de esta fase. [1] Los procesos más comunes de preparación de datos son:

Selección de datos: Se realiza la selección de elementos o filas incluir en el análisis, y la selección de atributos o columnas utilizar en el análisis, como características específicas de los datos.

Limpieza de datos: Implica observar los problemas con los datos y establecer las estrategias ya sea para corregirlos o eliminarlos. Estos procesos involucran el excluir incoherencias, aplicar codificación correcta, eliminación, reemplazo de datos sea manual o automatizado.

Construcción de nuevos datos: Se crean nuevos datos ya sean calculados o derivados de columnas o filas

Integración de datos: Esta implica el fusionar múltiples tablas en una sola ya sea porque tienen columnas similares o se busca aumentar el número de atributos (columnas) a las tablas

Formateo de datos: Se establece si los datos requieren un formato correcto u orden en concreto previo a la fase de modelamiento

2.1.4 Fase 4: Modelamiento

El modelado es una fase iterativa de adquisición de datos que implica ejecutar múltiples modelos y ajustar parámetros para obtener resultados satisfactorios. [1] La preparación de datos puede requerir pasos adicionales antes de cada iteración de modelado. Los procesos de esta fase más comunes son:

Selección de técnicas de modelado: se consideran los tipos de datos disponibles, los objetivos de minería de datos y los requisitos específicos de modelado, como el tamaño o tipo de datos necesarios y la facilidad de presentación de resultados.

Generación de un diseño de prueba: consta de describir los criterios de "bondad" de un modelo y definir los datos para poner a prueba estos criterios como el error medio absoluto o error medio cuadrado para problemas de regresión. O también la exactitud y presión en los problemas de clasificación.

Construcción de modelos: Se experimenta con varios modelos, se toma notas sobre la configuración y los datos utilizados para cada uno. Al final del proceso, se deberá tener: la configuración de parámetros, los modelos producidos y las descripciones de los resultados del modelo, incluyendo problemas de rendimiento y datos encontrados durante la ejecución del modelo.

Configuración de parámetros: Los parámetros que tienen los modelos varían en cantidad y funciones. Estos se deberán ajustar con base a los resultados mostrados de cada iteración de construcción y ejecución de los modelos.

Aprobación de modelos: Con una lista de modelos desarrollados se debe escoger cuáles serán los más precisos o efectivos que pasarían a la fase de evaluación.

Los resultados que se obtengan definirán si se debe seguir a la siguiente fase o retornar a alguna de las anteriores

2.1.5 Fase 5: Evaluación

Se comparan los resultados obtenidos con los criterios de éxito comercial definidos anteriormente. Si los resultados cumplen con los criterios, se procede a la implementación. Si no es así, se deben tomar medidas correctivas para mejorar el modelo o los datos. [1] Se debe documentar los hallazgos y las conclusiones extraídas de los modelos y del proceso de minería de datos que tendrán que ser abordadas para futuros procesos cómo también para determinar si es posible pasar a la fase de despliegue o se tiene que retornar a una fase anterior.

2.1.6 Fase 6: Despliegue

Consiste en la integración formal del modelo en los sistemas de información existentes o usar los resultados para informar las decisiones comerciales. Esta fase debe tener una planificación y seguimiento de los resultados, así como la documentación y la revisión del proyecto [1].

La planificación debe establecer varios aspectos cómo: a qué sistema se implementará los modelos, quienes lo utilizarán, verificar si los modelos se adaptan al sistema y si son precisos, cómo se mantendrán y actualizarán los modelos y hallazgos a lo largo del tiempo o qué seguimientos de hallazgos se deberá hacer

La documentación y revisión debería explicar: descripción detallada del problema original, el proceso que se llevó a cabo para la minería de datos, si existieron gastos en los procesos, qué cambios se hizo al respecto de la planificación inicial, resumen de resultados de resultados y recomendación para trabajos futuros de minería de datos

2.1.7 Fase 7: Visualización

La visualización de datos a través de dashboards (tableros) se tiene realizar con fundamentos que garantice su comprensión hacia el público a quien va a estar dirigido.

Los datos presentados en la gráfica tendrán que ser procesables interpretables, transparentes y de fácil acceso. Estos datos se deberán adaptar al tipo de dashboard a crear y a los indicadores clave de rendimiento, manteniendo una estructura coherente ya sea por agrupación, flujo o mapa de relaciones. [17]

Su diseño deberá de preferencia presentar los indicadores de rendimiento en la parte superior de la visualización, seguida por el contexto de datos y los detalles respectivos que complementen la explicación. Junto con el diseño se debe tomar en cuenta el balance de la paleta de colores, el tamaño de la fuente de texto, la cantidad de gráficos y el uso de tablas de resumen. A su vez esto dashboards tiene que cumplir principios básicos de funcionalidad cómo el profundizar en la explicación de datos, filtros, comparación, alertas y exportación/impresión. [17]

2.1.8 Fase 8: Evaluación de Usabilidad

Son sesiones de evaluación en las que un investigador solicita a un participante que realice tareas utilizando una o más interfaces de usuario específicas. Durante la sesión, se recopila información del comportamiento del participante y comentarios cada tarea que realiza. [18]

El objetivo de las pruebas de usabilidad es evaluar la facilidad y eficiencia con la que los usuarios pueden interactuar con la interfaz de usuario o sistema, identificar problemas y mejorar la experiencia del usuario. [18]

Dos de las pruebas estandarizadas más usadas son:

Heurísticas de Nielsen: Consiste en 10 principios generales para el diseño de interacción, los cuales se conocen como "heurísticas". Esta denominación se debe a que son reglas generales que no constituyen pautas de uso específicas. [19]

Las escalas de severidad de Nielsen van de 0 a 4 en nivel de severidad y N/A si no se aplica esa heurística como se describe en la **Tabla 2.1**.

Tabla 2.1. Niveles de severidad para calificación de heurísticas. [19]

Severidad	Descripción	Significado
0	No es problema	No es considerado en su totalidad como un problema de usabilidad

1	Problema apenas estético	No necesita ser modificado, a menos que haya tiempo disponible para hacerlo
2	Problema menor de usabilidad	La solución de ese problema deberá tener baja prioridad
3	Problema mayor de usabilidad	Es importante resolverlo por tanto deberá tener alta prioridad
4	Catástrofe de usabilidad	Se requiere corregirlo de prisa o volver a hacerlo por completo
N/A	No aplica	En caso de que no se considere que exista la heurística

Escala de usabilidad del sistema: es una herramienta sencilla compuesta por diez elementos que permite obtener una idea general de las evaluaciones subjetivas de la usabilidad de un sistema. [20] Este sistema se puntúa entre valores de 1 (muy descuerdo) y 5 (muy de acuerdo). Su resultado se calcula usando la siguiente formula:

$X = \text{Suma de los puntos de todas las preguntas impares} - 5$

$Y = 25 - \text{Suma de los puntos de todas las preguntas pares}$

$\text{Puntaje SUS} = (X + Y) \times 2.5$

La puntuación total es de 100 y cada una de las preguntas tiene un peso de 10 puntos. El rango de aceptabilidad, la escala de grado y los adjetivos calificativos con base en el puntaje se fijan de acuerdo con la **Figura 2.2**

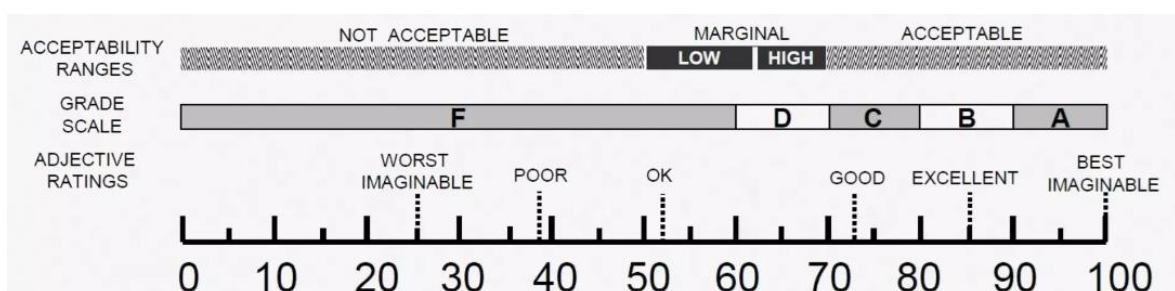


Figura 2.2. Clasificación de calificaciones de puntajes SUS [21]

3 DESARROLLO E IMPLEMENTACIÓN

El desarrollo del proyecto contó procesos de investigación y aplicación en donde se estuvo involucrado la recolección de información, codificación y ejecución scripts, tratamiento de datos automatizados y manuales, evaluaciones, diseño y despliegue. Esto se detalla a

continuación junto con el **ANEXO I** y el enlace para ver los códigos y datos desarrollados se encuentra en el **ANEXO IX**

3.1 Entendimiento del negocio:

Esta fase se caracterizó por el análisis del portal web SETEC acerca de los cursos y perfiles de cualificación registrados de personas capacitadas o certificadas por Operadores de Capacitación (**OC**), Organismos Evaluadores de la Conformidad (**OEC**) o capacitadores independientes (**CI**). El planteamiento de los objetivos a desarrollar entorno al análisis de la demanda de cursos y perfiles utilizando KPIs estáticos y dinámicos siguiendo el plan de proyecto planteado. El proceso se resume en **Figura 3.1**.

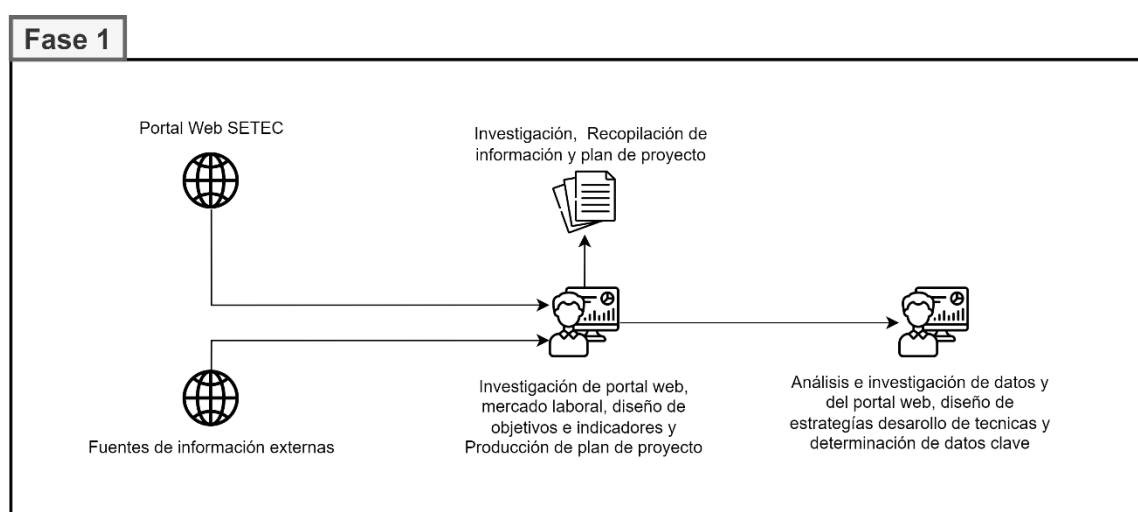


Figura 3.1. Diagrama de la Fase 1 – Entendimiento del Negocio

3.1.1 Análisis y entendimiento de los módulos del portal web SETEC

El portal web SETEC cuenta con 7 módulos con información acerca de cursos, perfiles, organizaciones, capacitadores y personas capacitadas y/o certificadas en distintas áreas y especialidades profesionales a nivel nacional. Los módulos que alberga son:

Catálogo Nacional de Cualificaciones – CNC

Este módulo cuenta con la información de todos los perfiles de cualificación a nivel nacional de Operadores de Capacitación y Evaluadores de la Conformidad, así como los perfiles inhabilitados actualmente.

Operadores de Capacitación – OC

Corresponde a la información de los Operadores de Capacitación cuyo estado puede ser habilitado, suspendido, finalizado su vigencia de calificación y cancelados.

Personas Capacitadas por OC – PCOC

Es el módulo de personas quienes han sido capacitadas en cursos y/o perfiles por Operadores de Capacitación.

Organismos Evaluadores de la Conformidad – OEC

Módulo con la información de Organismos Evaluadores de la Conformidad encargados de emitir certificaciones de perfiles a nivel nacional. Cuenta con la información de OEC habilitados, suspendidos, finalizados su vigencia de reconocimiento y cancelados.

Personas Certificadas por OEC – PCOEC

Módulo que cuenta con la información de las Personas Capacitadas por Organismos Evaluadores de la Conformidad.

Capacitadores Independientes – CI

Este módulo cuenta con la información de capacitores independientes quienes imparten cursos de capacitación a profesionales a nivel nacional.

Personas Capacitadas por CI – PCCI

Son las Personas Capacitadas por Capacitadores Independientes quienes se han certificado en los cursos que han tomado.

Un análisis más detallado se lo puede encontrar en el **ANEXO I**

3.1.2 Análisis y entendimiento del mercado laboral nacional

En Ecuador se registró a nivel nacional, en enero 2022, la tasa de empleo adecuado fue de 33,1%; para el área urbana de 41,4%; mientras que, en el área rural fue de 17,3%. [22] Esta tasa de empleo está conformada por profesionales quienes cuentan con certificaciones de cualificación de SETEC, el SECAP (Servicio Ecuatoriano de Capacitación Profesional). Estas entidades validan las cualidades profesionales en distintos sectores laborales como son (Gestión Documental, Formación de Formadores, Maquillaje, Prevención de Riesgos Laborales, Apicultor, entre otros). [23]

La probabilidad de conseguir un empleo adecuado aumenta con estas certificaciones oficiales de parte de estas instituciones gubernamentales. Estas certificaciones están disponibles para personas mayores de 16 años ecuatorianos, que cumplan los prerequisites establecidos para cada perfil de otros sectores, así como en la normativa legal vigente para el servicio de certificación de personas por competencias laborales. [23]

Ministerio del Trabajo a través de la Subsecretaría de Cualificaciones Profesionales entregó las resoluciones de Calificación como OC y el Reconocimiento como OEC a más de 60 institutos y empresas que cumplieron con los requisitos establecidos para este efecto. [24]

Estos nuevos OEC y OC se suman a los 181 Organismos Evaluadores de la Conformidad Reconocidos y 424 Operadores de Capacitación Calificados con los que cuenta actualmente el Sistema Nacional de Cualificaciones Profesionales. [24] De acuerdo con el PLAN NACIONAL DE EDUCACIÓN Y FORMACIÓN TÉCNICA Y PROFESIONAL emitido por el ministerio de educación y ministerio del trabajo. [25]

3.1.3 Diseño de objetivos de negocio e indicadores estáticos y dinámicos

Se deberá crear un sistema de minería de datos enfocado al análisis de la demanda de cursos y/o perfiles profesionales cualificados por SETEC utilizando KPIs estáticos y dinámicos. Estos deberán ser relevantes en el estudio y estar en desarrollados bajos los **criterios SMART** para poder ser presentados a través de panales analíticos (tableros o dashboards).

Siguiendo el análisis desarrollado en la sección 3.1.1 los KPIs desarrollados son de los módulos de OC, OEC y CI a los cuales se les llamará tipo de razón social. A su vez se tiene que la información de los módulos PCOC, PCOEC y PCCI acerca de los cursos y/o perfiles que han sido registrados de cada persona natural (profesional) que los ha tomado.

Los KPIs se desarrollaron en torno a los dos enfoques presentes en cada módulo:

Enfoque de Cursos/Perfiles:

- Cursos: Cuentan con un nombre de curso, área, especialidad, carga horaria (número de horas de capacitación), modalidad
- Perfiles: Cuentan con un nombre de perfil, familia y sector

Puesto a que los OC cuentan tanto con capacitaciones de cursos y certificaciones de perfiles se optó por combinar los datos de cursos y perfiles en una sola tabla discriminándolos por una nueva columna llamada **tipo**.

Y para completar las columnas modalidad y carga horaria se asignó todos los perfiles con modalidad desconocida.

Se desarrolló un estudio de mercado de 10 razones sociales con el objetivo de encontrar tanto el valor de las certificaciones que ofertan cómo la carga horaria de

las mismas. Sin embargo, la diferencia entre cursos de capacitación y perfiles de certificación no estaba clara por lo que se dispuso una carga de 6 horas que es el tiempo estimado que toma el rendir un examen de certificación de un perfil, mientras que, para el costo se determinó en 250\$.

Del mismo modo se hizo un estudio de mercado con el número de seguidores de las redes sociales de Facebook e Instagram cómo el número de convocados el cual se lo contrasto con el número de capacitados y certificados reales para obtener el porcentaje de asistencia a cursos y perfiles respectivamente.

Enfoque de Razón Social.

Una razón social puede ser un OC, OEC o CI. Los OCs al igual que los CIs realizan capacitaciones en cursos mientras que los OECs certifican en perfiles.

Hay razones sociales que pueden ser OC y OEC al mismo tiempo ya que capacitan y certifican en cursos y perfiles respectivamente.

En base a estas dos variables de determino los siguientes KPIs cómo se muestra en la siguiente **Tabla 3.1:**

Tabla 3.1. KPIs por enfoque

Por Curso/Perfil:	Por Razón Social:
<ul style="list-style-type: none"> • Número de personas capacitadas o certificadas • Total de horas de capacitación o certificación • Total de ganancias • Porcentaje de asistencia 	<ul style="list-style-type: none"> • Total de cursos • Volumen de personas capacitadas o certificadas • Volumen de horas de capacitación o certificación • Volumen de ganancias • Volumen Porcentaje de asistencia

3.1.4 Producción del plan de proyecto

Para el desarrollo de este proyecto se deberá seguir el siguiente plan de 16 semanas en donde se deberá cubrir y avanzar en cada fase de desarrollo de la metodología CRISP-DM y las dos fases extra de visualización y análisis de usabilidad como se detalla en la **Tabla 3.2**

Tabla 3.2. Plan de proyecto

Semana referencial / Etapas	Tareas específicas	Resultado esperado (sí aplica)
1	Investigación de la literatura relacionada Metodología Herramientas Conceptos teóricos y técnicos	<ul style="list-style-type: none"> • Recopilación de fuentes investigación • Desarrollo del marco teórico
2	Entendimiento del negocio Análisis y entendimiento de los módulos del portal web SETEC Análisis y entendimiento del mercado laboral nacional Diseño de objetivos de negocio e indicadores estáticos y dinámicos Producción del plan de proyecto	<ul style="list-style-type: none"> • Redacción del análisis de los módulos que conforman el portal web SETEC
3	Entendimiento de los datos Análisis y entendimiento de los datos en los módulos del portal web SETEC Extracción y colección de datos Descripción de datos Análisis exploratorio de datos Verificación de la calidad de datos	<ul style="list-style-type: none"> • Redacción del análisis de los datos que manejan los módulos del portal web SETEC • Datos recolectados iniciales • Informe de datos recolectados iniciales
4	Preparación de los datos Selección de datos Limpieza de datos Transformación de datos Formateo de datos Construcción de nuevos datos Integración de datos Diseño del Datawarehouse Almacenamiento de datos	<ul style="list-style-type: none"> • Conjuntos de datos preparados almacenados • Datawarehouse sin KPIs dinámicos
5	Modelamiento Selección de técnicas de modelamiento Generación de diseño de pruebas Construcción de modelos	<ul style="list-style-type: none"> • Modelos evaluados y aprobados

	Evaluación de modelos	<ul style="list-style-type: none"> Redacción de modelos seleccionados, evaluados y aprobados
6	Evaluación Evaluación del resultado de los modelos desarrollados en base a los objetivos planteados Carga de información predictiva en el Datawarehouse Revisión de las actividades y decisiones tomadas en las anteriores fases Determinación de los siguientes pasos	<ul style="list-style-type: none"> Información predictiva almacenada en el Datawarehouse Redacción de evaluación de modelos
7	Despliegue Diseño de plan de despliegue de modelos Plan de monitorización y mantenimiento Despliegue de los modelos de en herramientas de visualización de datos	<ul style="list-style-type: none"> Modelos desplegados Redacción de despliegue de modelos
8	Visualización	<ul style="list-style-type: none"> Gráficos de datos y resultados
9	Evaluación de usabilidad	<ul style="list-style-type: none"> Informe de evaluación de métricas de Nielsen Informe de evaluación de SUS
10	Documentación y correcciones	Trabajo de Integración Curricular
11	Documentación y correcciones	Trabajo de Integración Curricular
12	Documentación	Trabajo de Integración Curricular

3.2 Entendimiento de datos:

El entendimiento de datos comprende el análisis y entendimiento de los datos en los módulos del portal web SETEC. En esta fase se desarrolló un análisis previo tanto de los datos existentes en el portal web como de otras fuentes de información externar relacionadas con el negocio para determinar los datos a extraer y las estrategias a desarrollar para su extracción, colección y análisis de datos. Cómo se muestra en la **Figura 3.2**

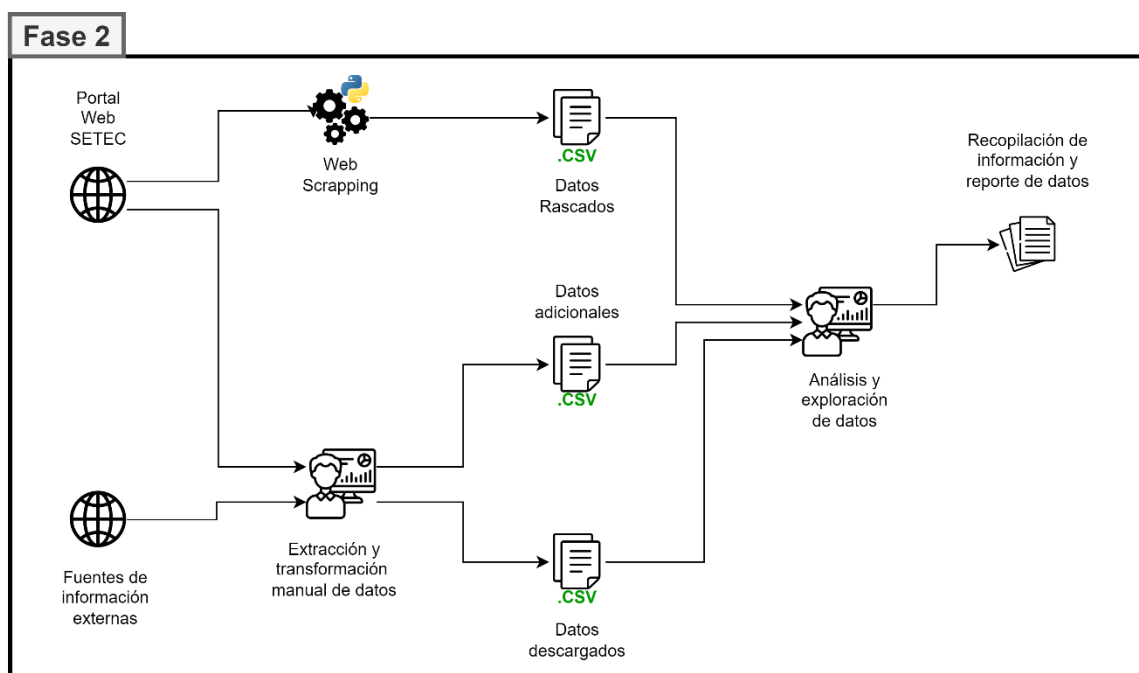


Figura 3.2. Diagrama de la fase 2 – Entendimiento de Datos

A continuación, se detalla la investigación previa a la aplicación técnica de extracción, recolección y análisis de datos:

3.2.1 Análisis y entendimiento de los datos en los módulos del portal web SETEC

Datos existentes

Cómo se analizó en la sección 3.1.1 y el **ANEXO II**, los datos del portal web están distribuidos por módulos y submódulos. Los Submódulos pueden variar de acuerdo con el módulo y la cantidad de datos disponibles también lo hará. En cuanto a la estructura de página se tiene que cuenta con submódulos de los cuales se podrá extraer información a través de web scraping al contener una tabla de datos como se observa en la **Figura 3.3**. A su vez estas tablas cuentan con un botón de detalle donde se encuentran datos relacionados con las razones sociales como se ve en la **Figura 3.4**. Los submódulos de

descarga promocionan datos listos para descargarse en formato .xlsx como se observa en la **Figura 3.5**

Operadores de Capacitación

Si estás interesado en capacitarte, en esta sección podrás encontrar a todos los Operadores de Capacitación calificados a nivel nacional. Puedes buscar la información por ubicación geográfica, oferta en capacitación continua/competencias laborales o nombre de la institución y comunicarte directamente con ellos.

*Selección Filtro: RAZÓN SOCIAL
 *Razón Social: ASOCIACION DE TRADUCTORES E INTE

Operadores de Capacitación

Ruc / Código	Razón Social	Nombre Comercial	Teléfono	Celular	Email	Número Resolución	Fecha Resolución	Estado	Cantón	Descargar	Detalle
1792136245001	ASOCIACION DE TRADUCTORES E INTERPRETES DEL ECUADOR ATIEC		00-0000000	0000000000	s.hernandez@sg	MDT-SCP-2022-0533	27-12-2022	CALIFICADO	QUITO		

Figura 3.3. Análisis de componentes del módulo de OC del portal web

Oferta de Capacitación

Capacitación Continua

Área	Especialidad	Curso	Modalidad	Carga Horaria
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	ADMINISTRACIÓN FINANCIERA	PRESENCIAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	ADMINISTRACIÓN FINANCIERA	VIRTUAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	ADMINISTRACIÓN FINANCIERA	SEMIPRESENCIAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	CONTRATACIÓN PÚBLICA PARA ENTIDADES CONTRATANTES	PRESENCIAL	80
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	CONTRATACIÓN PÚBLICA PARA ENTIDADES CONTRATANTES	VIRTUAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	CONTRATACIÓN PÚBLICA PARA ENTIDADES CONTRATANTES	SEMIPRESENCIAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	CONTRATACIÓN PÚBLICA	VIRTUAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	ADMINISTRACIÓN GENERAL	CONTRATACIÓN PÚBLICA	SEMIPRESENCIAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	GESTIÓN DEL TALENTO HUMANO	GRAFOLOGÍA	PRESENCIAL	40
ADMINISTRACIÓN Y LEGISLACIÓN	GESTIÓN DEL TALENTO HUMANO	GRAFOLOGÍA	SEMIPRESENCIAL	40

Figura 3.4. Análisis de componentes del detalle de un OC del portal web

Nombre – Provincia – Cantón

Tabla de datos

Botón de descarga

Nombre	Provincia	Canton
ARIAS SALAZAR GABRIEL ESTEBAN	AZUAY	CUENCA
ALEXANDRA ELIZABETH GARCIA REINO	AZUAY	CUENCA
UNIVERSIDAD DE CUENCA	AZUAY	CUENCA
MOROCHO JEREZ JOHNNY BYRON	AZUAY	CUENCA
SUQUISUPA ESPINOZA LUCIO HUMBERTO	AZUAY	CUENCA
DATALIGHTS CIA.LTDA.	AZUAY	CUENCA
SUQUISUPA ESPINOZA ANGEL ROGERIO	AZUAY	CUENCA
REGALADO SANGURIMA JOHANNA MARIA	AZUAY	CUENCA
GLOBAL-ESTUDIOS S.A.S.	AZUAY	CUENCA
CECAD-CENTRO DE CAPACITACIONES S.A.S. B.I.C.	AZUAY	CUENCA

Figura 3.5. Análisis de submódulo de descarga de OC

Datos adicionales

Para desarrollar la estimación de precios de los cursos ofertados en el portal web se tomó como referencia los cursos ofertados por el Centro de Educación Continua (CEC – EPN). Se tomó únicamente los datos de cursos en el sector de Tics como se muestra en la siguiente **Figura 3.6**

Tecnológicos

- Android
- Business Intelligence – Bases de datos
- Desarrollo de software
- Excel
- Fibra Óptica y Telefonía IP
- Gerencia de sistemas
- Hackeo Ético
- Informática Forense
- ITIL®
- Java
- Linux
- Office
- Power BI
- Programación
- Programación PHP
- Project
- Python
- Seguridad de la Información

Figura 3.6. Lista de cursos tecnológicos ofertados por el CEC en 2022B [26]

Esta información de costos se sumó a la investigación de mercado de empresas externas realizada en 2021 la cual cuenta con los datos de: Universidad o Institución, Ciudad, Pagina web, Servicios, Tipos de Oferta, Áreas de Formación, Nombre de los Programas, Modalidad, horarios, Tiempo de duración, Duración programa, Precio, formas de pago, certificación, plataforma educación virtual y Red de educación. Ver **ANEXO VIII**

También para determinar el porcentaje de asistencia se tuvo de referencia la cantidad de personas suscritas a páginas oficiales de Facebook e Instagram para estimar un número de personas convocadas como se observa en la **Figura 3.7**



Figura 3.7. Análisis de número de suscriptores en Facebook e Instagram de la Razón Social OC: Universidad de Cuenca

3.2.2 Extracción y colección de datos

Proceso de web scraping

De acuerdo con los análisis realizados en la sección 3.1.1 3.2.1 y el **ANEXO II** se comenzó con la extracción de datos:

El proceso de extracción se desarrolló con la técnica de *Web scraping* (Raspado Web) creando un “robot” o “araña” encargado de extraer los datos de cada módulo, submódulo y detalle. Estos robots se desarrollaron en el lenguaje **Python** con la librería **Selenium**, mientras que los datos recolectados se los almaceno en archivos **.CSV** utilizando la librería **Pandas**. El script de **web_scraper.py** está adaptado para la extracción de los módulos y submódulos este utiliza un archivo .csv llamado **data_controller_bot.csv** el cual tiene toda la información necesaria para la extracción de datos. La estructura de la tabla se muestra a continuación. Ver tabla **Tabla 3.3**

Tabla 3.3. Descripción de las columnas del archivo data_controller_bot.csv

Columna	Descripción
doc_name	Nombre del documento en el que se guardará el archivo .CSV
columns	Columnas de la tabla que se guardará en él .CSV
xpath_URL	Ruta URL del módulo a extraer información

xpath_MODULE	XPATH de Submódulo a extraer información
xpath_FILTER	XPATH de filtro de lista desplegable
xpath_OPTION	XPATH de opción de la lista desplegable
xpath_TEXTBOX_QUERY	XPATH de cuadro de texto para escribir la consulta
xpath_BOTTON_SEARCH	XPATH de botón de búsqueda
xpath_BOTTON_LAST_PAGE	XPATH de botón de última página
xpath_BOTTON_NEXT_PAGE	XPATH de botón de página siguiente
xpath_JUMP_PAGE	XPATH de botón de página en la ubicación 10 del controlador
xpath_TABLE	XPATH de la tabla a extraer información
DETAIL	Indicador de si existe o no extracción de detalles
folder	Carpeta de salida
subfolder	Subcarpeta de salida
BUTTON_DETAIL	XPATH de botón de detalle
xpath_TABLE_DETAIL	XPATH de tabla de detalle
xpath_BOTTON_DETAIL_NEXT_PAGE	XPATH de botón de página siguiente de detalle
xpath_BOTTON_EXIT	XPATH de botón de salida de detalle

Proceso de extracción y *transformación* manual de datos

En este proceso se tuvo que descargar los archivos del portal web relevantes para completar la información necesaria para el caso de negocio. Después se obtuvo la investigación de mercado de mercado de empresas externas realizada 2021. Finalmente se hizo un estudio de mercado obteniendo el número de personas convocadas estimadas tomando como referencia el número de suscriptores de las redes sociales oficiales de al menos 10 razones sociales de cada tipo (OC, OEC y CI)

Datos existentes recopilados

Los datos obtenidos fueron recolectados desde el 16 de junio de 2022 hasta el 17 de noviembre de 2022. Debido a que los datos de la página son actualizados día tras día es posible que los datos varíen de la estimación inicial del número de datos a obtener.

Se excluyo a los módulos de CNCOC, CNCOEC y CNCPI ya que representaban redundancia en cuanto a los datos de los demás módulos.

Los archivos obtenidos de la recolección de datos se observan en la **Tabla 3.4:**

Tabla 3.4. Resumen de los archivos rascados

Módulo	Submódulo	Detalle	Nombre de archivo	Tamaño del archivo
OC	OC	-	oc.csv	39 kB
OC	OCS	-	ocs.csv	92 bytes
OC	OCF	-	ocf.csv	15 kB
OC	OCC	-	occ.csv	1 kB
OC	OC	OC_CC	oc_cc.csv	445 kB
OC	OC	OC_CL	oc_cl.csv	18 kB
PCOC	PCOC	-	pcoc.csv	45.3 MB
OEC	OEC	-	oec.csv	47 kB
OEC	OECS	-	oecs.csv	367 bytes
OEC	OECF	-	oecf.csv	9 kB
OEC	OECC	-	oecc.csv	1 kB
OEC	OEC	OEC_DR	oec_dr.csv	150 kB
PCOEC	PCOEC	-	pcoec.csv	75 MB
CI	CI	-	ci.csv	421 kB
CI	CI	CI_LCA	ci_lca.csv	1.1 MB
PCCI	PCCI	-	pcci.csv	17.1 MB

Adicional a los datos extraídos a través de web scraping se descargó los archivos en formato xlsx los cuales fueron posteriormente transformados a formato .csv de los submódulos de descargas para completar los datos toda la base de datos a ser analizada. Los archivos obtenidos de acuerdo con la **Tabla 3.5** fueron:

Tabla 3.5. Lista de archivos descargados

Módulo	Nombre del archivo	Tamaño de archivo
OC	oc_dl_area_especialidad.csv	293 kB
OC	oc_dl_familia_sector_perfil.csv	38 kB
OC	oc_dl_provincia_canton.csv	31 kB
OEC	oec_dl_familia_sector_perfil.csv	378 kB
OEC	oec_dl_provincia_canton.csv	15 kB
CI	ci_dl_provincia_canton.csv	77 kB

Datos adicionales recopilados

Para el cálculo de costos y horas de cursos se utilizó datos de referencia de archivos de costos de educación continua, así como se requirieron de datos extra para la ubicación geográfica. Algunos de estos archivos fueron descargados de sitios web oficiales, otros fueron producto de investigación de mercado por parte de empresas o por investigación autoría propia. Ver **Tabla 3.6**

Tabla 3.6. Descripción de datos extra

Nombre del archivo	Descripción	Fuente
costos_oferta_educacion_continua.xlsx	Investigación de mercado de empresa externa desarrollado en 2021 que recopila información acerca de la oferta de educación continua de distintas instituciones a nivel nacional	Documento de investigación provisto por el PHD. Julián Galindo
costos_de_cursos_cec_epn.xlsx	Investigación de información de cursos de centro de educación continua de la Escuela Politécnica Nacional desarrollada en el año 2022	Documento desarrollado por autoría propia. URL: https://www.cec-epn.edu.ec/
Anexos_y_Tablas_para_entrega_Catastros_GADS.xlsx	Documento con la información de las provincias, cantones y parroquias del Ecuador	Documento descargado del sitio oficial del SRI. URL: https://www.sri.gob.ec/DocumentosAlfredoPortlet/descargar/a7ce61fa-d8e6-4b77-8999-b99b617780a2/Anexos+y+Tablas+para+entrega+Catastros+GADS.xlsx
pcoc_num_conv.csv	Investigación de número de personas suscritas a las a la página oficial de las razones sociales OC de Facebook e	Páginas de Facebook e Instagram oficiales de las razones sociales

	Instagram hasta la fecha del 8 de enero de 2023	
pcoec_num_conv.csv	Investigación de número de personas suscritas a las a la página oficial de las razones sociales OEC de Facebook e Instagram hasta la fecha del 8 de enero de 2023	Páginas de Facebook e Instagram oficiales de las razones sociales

3.2.3 Descripción de datos, Análisis exploratorio y verificación de calidad de datos

Descripción de datos existentes

Los datos recolectados se estiman en base a la cantidad de registros por página. la fórmula para la cantidad de datos estimada es de numero de registros por número de páginas del submódulo. Dependiendo del submódulo los registros por página son de entre 10 o 20 registros, salvo la última página cuya cantidad de registros puede variar entre 1 a 10 o 20 registros. En cuanto a los datos que formaban parte del detalle su número no es posible de estimar en concreto, esto pasa con los módulos OC_CC, OC_CL, OEC_DR y CI_LCA cómo se ve en la **Tabla 3.7**. La descripción específica de cada archivo se la puede observar en el **ANEXO II**

Tabla 3.7. Resumen de datos rascados

Módulos/ Submódulo	Registros	Páginas	Total registros	Archivos Finales	Filas	Columnas	Total datos
OC	20	28	560	oc.csv	542	10	5420
OCS	3	1	3	ocs.csv	1	3	3
OCF	10	26	260	ocf.csv	178	3	534
OCC	10	2	20	occ.csv	12	3	36
OC_CC	-	-	-	oc_cc.csv	5217	7	36519
OC_CL	-	-	-	oc_cl.csv	100	5	500
PCOC	20	23892	477840	pcoc.csv	251989	8	2015912
OEC	13	20	260	oec.csv	221	10	2210

OECS	2	1	2	oeecs.csv	4	3	12
OECF	10	11	110	oecf.csv	100	3	300
OECC	10	2	20	oecc.csv	14	3	42
OEC_DR	-	-	-	oec_dr.csv	1205	4	4820
PCOEC	20	19320	386400	pcoec.csv	301572	8	2412576
CI	10	271	2710	ci.csv	2970	6	17820
CI_LCA	-	-	-	ci_lca.csv	7105	6	42630
PCCI	10	19150	191500	pcci.csv	146010	5	730050

Estas tablas fueron almacenadas posteriormente en la carpeta **existing_data**, dentro de la subcarpeta **DIRTY_DATA**. En cuanto a los datos descargados los resultados fueron los siguientes que se muestran en la **Tabla 3.8**:

Tabla 3.8. Resumen de archivos descargados

Archivo	# Filas	# Columnas	Total datos	Total datos NO Nulos
ci_dl_provincia_canton.csv	1619	3	4857	4857
oc_dl_area_especialidad.csv	2905	3	8715	8715
oc_dl_familia_sector_perfil.csv	222	4	888	888
oc_dl_provincia_canton.csv	547	3	1641	1641
oec_dl_familia_sector_perfil.csv	2300	4	9200	9200
oec_dl_provincia_canton.csv	242	3	726	726

Estos fueron almacenados en la carpeta **existing_data** en la subcarpeta **DOWNLOADED_DATA**. La descripción específica de cada archivo está en el **ANEXO II**

Descripción de datos adicionales

En cuanto a los datos adicionales los resultados se resumen en la **Tabla 3.9** y **Tabla 3.10**:

Tabla 3.9. Resumen de archivos transformados manualmente

Archivo original	Transformación manual	Archivo final
costos_oferta_educacion_continua.xlsx	Selección de columnas relevantes y conversión a .CSV	costos_cursos_educacion_continua.csv

costos_de_cursos_cec_e pn.xlsx	conversión a .CSV	costos_de_cursos_cec_e pn.csv
Anexos_y_Tablas_para_e ntrega_Catastros_GADS. xlsx	Selección de columnas relevantes y conversión a .CSV	Ubicación.csv

Tabla 3.10. Resumen de datos de archivos de datos adicionales

Archivo	# Filas	# Columnas	Total datos	Total datos NO Nulos
costos_cursos_educacion_continua.csv	521	12	6252	4604
costos_de_cursos_cec_epn.csv	11	5	55	55
Ubicación.csv	224	3	672	672
pcoc_num_conv.csv	11	5	55	55
pcoc_num_conv.csv	11	8	88	88

Nota: Los datos de número de personas convocadas fue una investigación que se hizo a solo 11 y 10 razones sociales de OC y OEC respectivamente. Esta investigación se la hizo después de obtener el volumen de capacitados y certificados por razón social.

Estos archivos se guardaron en la subcarpeta **additional_data** de la carpeta **collecting_initial_data**.

3.3 Preparación de datos:

En esta fase se extrajo, seleccionó, transformó, construyó e integro los datos del portal web. Esta fase se compuso de 4 scripts en **Jupyter Lab** (ipynb) para el proceso de extracción, selección, limpieza, transformación, construcción e integración de datos. La ejecución y orden de estos se los detalla en la siguiente sección junto con el diagrama de resumen en la **Figura 3.8**:

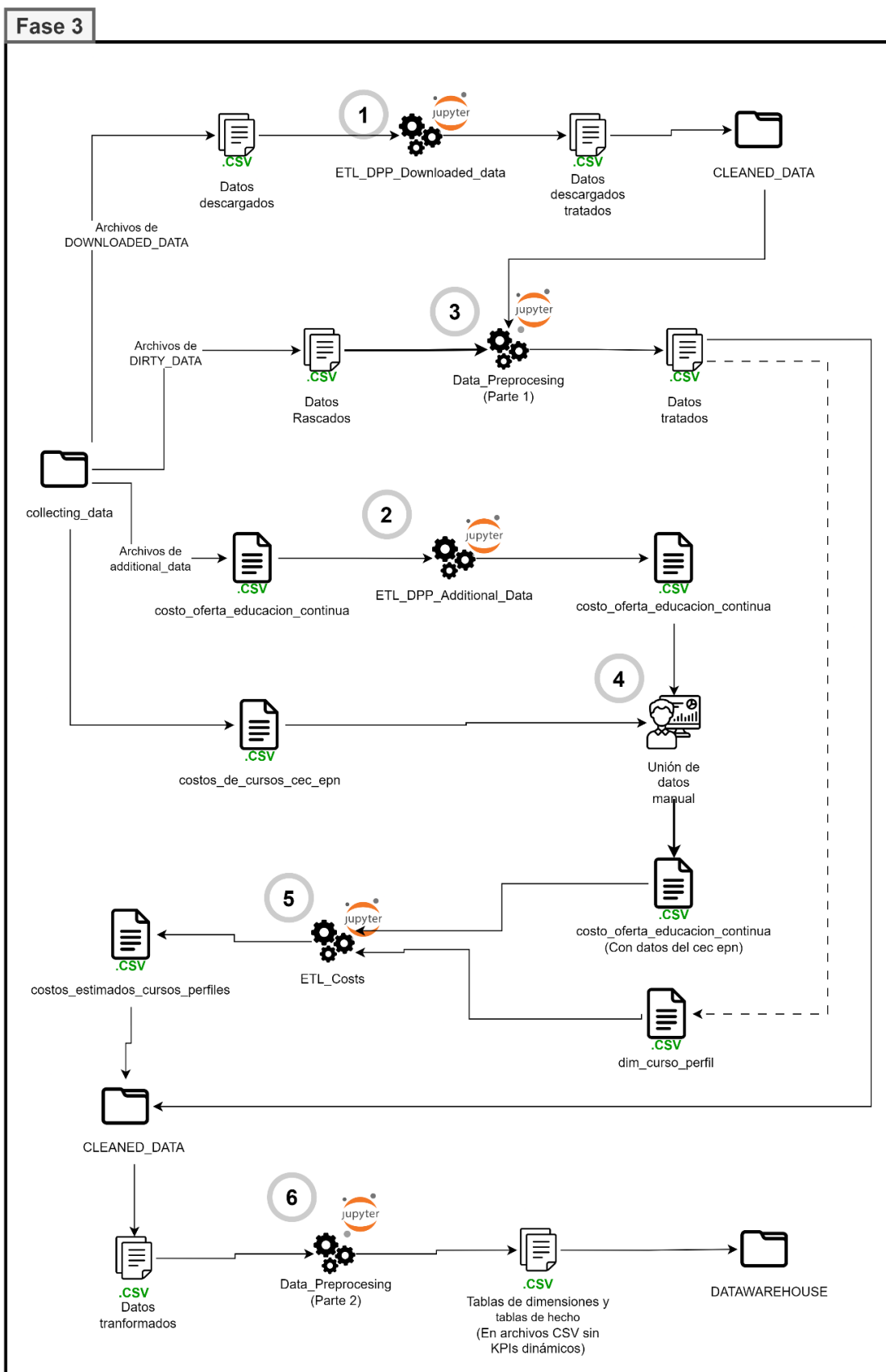


Figura 3.8. Diagrama de la Fase 3 – Preparación de Datos

1. **ETL_DPP_Downloaded_data.ipynb:**

Extrae los datos descargados del portal web SETEC ubicados en la carpeta de DOWNLOADED_DATA, estos archivos fueron transformados y eliminar valores repetidos para posteriormente se guardados en la carpeta CLEANED_DATA

2. **ETL_DPP_Additional_Data.ipynb:** Extrae los datos de los archivos costo_oferta_educacion_continua.csv ubicado en additional_data, para limpiarlos, eliminar valores repetidos, cambiar valores que se adapten al modelo de negocio de SETEC, eliminar datos nulos, dar formatos correctos y posteriormente guardados en la carpeta CLEANED_DATA para posteriores tratamientos

3. **Data_Preprocessing.ipynb (Parte 1):** Extrae los datos de rascados de la carpeta DIRTY_DATA donde trata los datos de cada uno de los archivos, elimina valores repetidos y construye nuevos datos y archivos y los guarda en CLEANED_DATA

Nota: Esta primera parte del script se ejecuta hasta la creación de las dimensiones .CSV: dim_curso_perfil, dim_fecha y dim_razon_social

4. **Unión de datos manual:** El proceso de unir los datos de la investigación de costos_de_cursos_cec_epn.csv y costo_oferta_educacion_continua.csv se lo hace de forma manual colocando los datos en las columnas correspondientes de costos_de_cursos_cec_epn.csv a costo_oferta_educacion_continua.csv. El resultado será el archivo costo_oferta_educacion_continua.csv con los datos extra de costos_de_cursos_cec_epn.csv

5. **ETL_Costs.ipynb:** Este proceso extrae los datos de los archivos costo_oferta_educacion_continua.csv (Con datos del cec epn) y del archivo dim_curso_perfil.csv creado en el proceso 3

Posteriormente se ejecuta algoritmos para evaluar el costo en dólares por hora de cada curso del archivo costo_oferta_educacion_continua.csv estos algoritmos determinan por categoría, provincia, cantón, similitud de palabras por curso, modalidad y carga horaria valores estimados de lo que debería valer la fracción de hora creando un dataframe con esta información. Luego, se compara cada característica de cada fila de dim_curso_perfil.csv con cada característica de cada fila del dataframe creado anteriormente.

El resultado será un archivo llamado costos_estimados_cursos_perfiles.csv con los identificadores (id_curso_perfil) y el valor estimado del costo por hora de ese

curso, este archivo se guardará en CLEANED_DATA para posteriores tratamientos.

6. Data_Preprocessing.ipynb (Parte 2): Con los datos de procesados de costos_estimados_cursos_perfiles.csv se continúa ejecutando el script, primero calculado el costo del curso en base al estimado de valor de hora por las horas total que tiene el curso real en el archivo dim_curso.csv. Durante el proceso de limpieza y transformación de datos se desarrolló el diseño del DATAWAREHOUSE como se muestra en la **Figura 3.9**

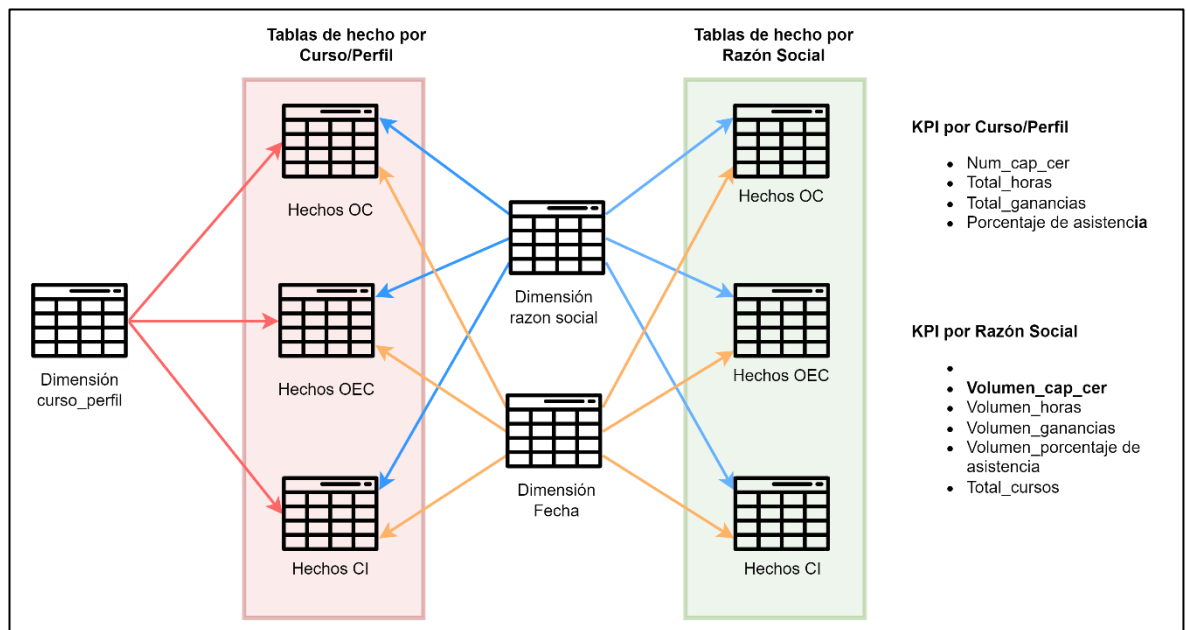


Figura 3.9. Diseño de Datawarehouse

El Datawarehouse se creó con una arquitectura con las dimensiones y tablas de hecho la cual aborda los KPIs desde un enfoque de Curso/Perfil y un enfoque de Razón Social. Cada enfoque con sus KPIs correspondientes que formarían un análisis completo de un OC, OEC o CI.

De modo que, se crean las tablas de hecho de cada tipo de razón social (OC, OEC y CI) para los respectivos enfoques. Las tablas de hechos contarán con los cálculos de numero de capacitados y/o certificados, además que se usará los archivos pcoc_num_conv.csv y pcoec_num_conv.csv para crear el KPI de porcentaje de asistencia.

Los archivos finales que se guardan en la carpeta DATAWAREHOUSE cómo se muestra en la **Tabla 3.11**:

Tabla 3.11. Lista de los archivos finales del proceso de Preparación de Datos

Nombre del archivo	Descripción
dim_curso_perfil.csv	Dimensión de los cursos y perfiles con su valor de costos
dim_fecha.csv	Fechas de capacitaciones y certificaciones registradas
dim_razon_social.csv	Razones sociales OC, OEC y CI con la información de su ubicación geográfica
fact_ci_cp.csv	Tabla de hechos de CI con enfoque en Curso/Perfil
fact_ci_rs.csv	Tabla de hechos de CI con enfoque en Razón Social
fact_oc_cp.csv	Tabla de hechos de OC con enfoque en Curso/Perfil
fact_oc_rs.csv	Tabla de hechos de OC con enfoque en Razón Social
fact_oec_cp.csv	Tabla de hechos de OEC con enfoque en Curso/Perfil
fact_oec_rs.csv	Tabla de hechos de OEC con enfoque en Razón Social

Los datos de estos archivos permitirán crear los datos para el entrenamiento de los modelos, así como para crear datos ficticios que simulen datos futuros del año 2023 para después ser evaluados por los mismos modelos. Se debe tomar en cuenta que:

- Existen muchos datos nulos o que no empataron con el nombre de algún campo al momento de hacer una agregación de datos.
- Al solo tener información del número de convocados de solo 10 razones sociales de OC y OEC el KPI porcentaje_asistencia contendrá varios valores vacíos en sus filas
- El KPI de porcentaje_asistencia no existe en las razones sociales CI ya que no se posible encontrar el número de convocados a sus cursos por medio de redes sociales oficiales

3.4 Modelamiento y Evaluación:

Las fases de modelamiento y evaluación se las trato en una sola debido a la naturaleza de la herramienta de **RapidMiner** que permite construir, entrenar y evaluar modelos usando su función de **Auto Model** a la cual se le cargo datos de entrenamiento para crear los distintos modelos para cada tabla de hechos y generar KPIs dinámicos. Una vez se hayan construido los modelos la herramienta lista y puntúa los mejores modelos para después seleccionar los mejores modelos en base a criterios del usuario quien decidirá si en base a los valores métricas arrojadas cual modelo se aprueba o no. Estos modelos se evaluarán

con datos de prueba y ficticios y su resultado se recopilará en archivos CSV tanto los datos reales y ficticios para posteriores procesos en la fase de despliegue. El resumen de estas fases se observa en la **Figura 3.10**. Los procesos de estas fases se detallan a continuación:

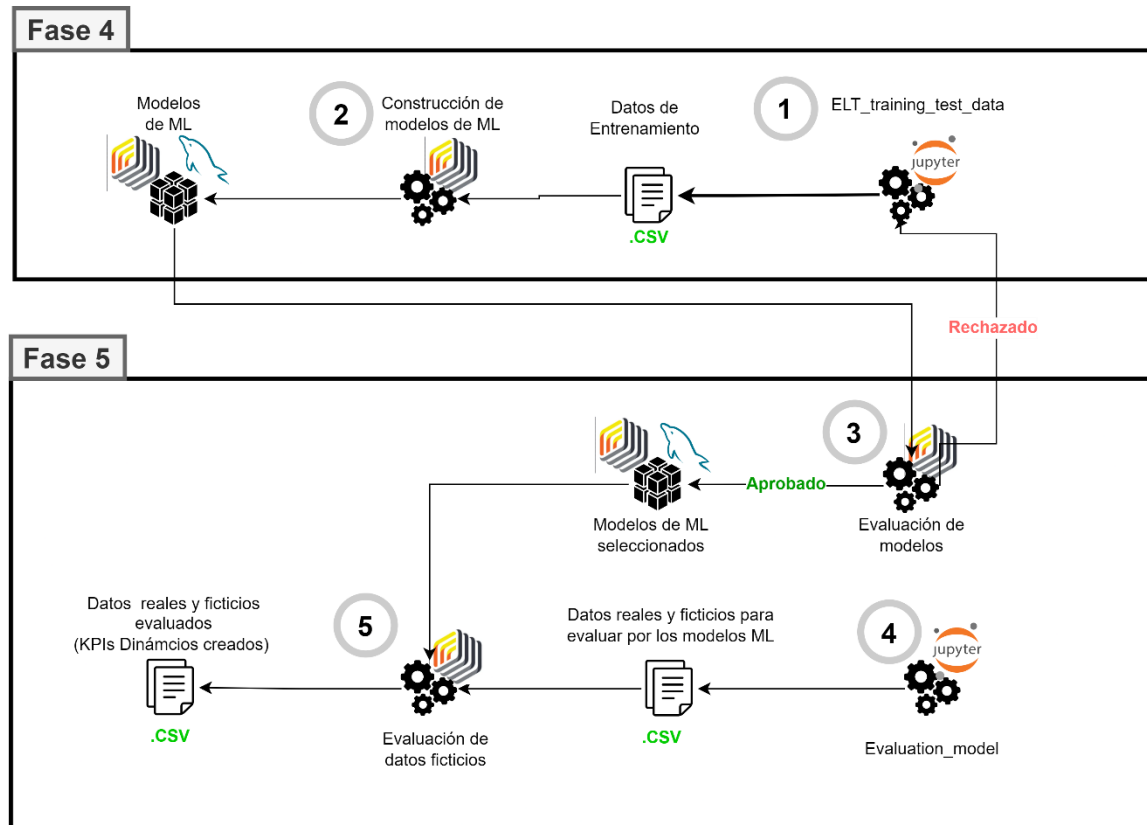


Figura 3.10. Diagrama de las fases 4 - Modelamiento y 5 Evaluación

- 1. *ELT_training_test_data.ipynb*:** Se extrae la información de los archivos .CSV de la carpeta de DATAWAREHOUSE con los que se construye datos de entrenamiento para los modelos.
- 2. *Construcción de modelos de ML*:** Dentro de la herramienta se selecciona la función de **Auto Model**. Se cargan los datos de entrenamiento como se observa en la **Figura 3.11**.

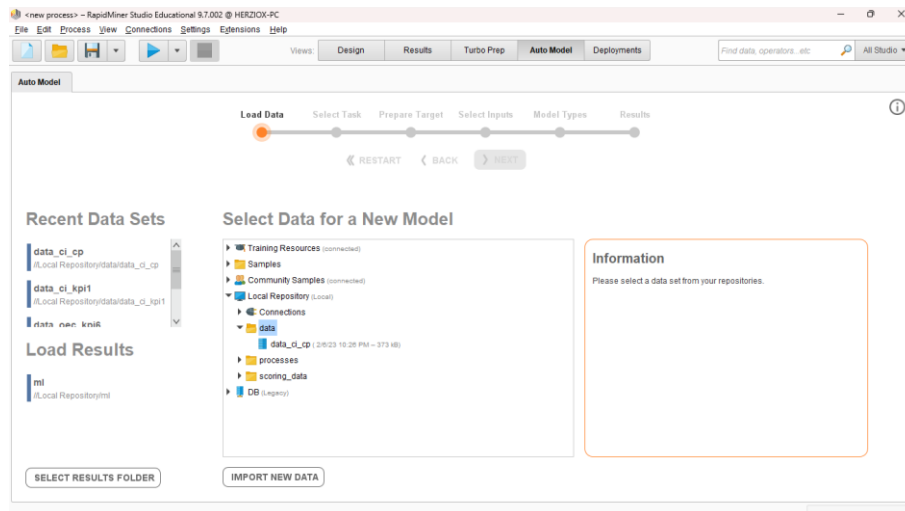


Figura 3.11. Carga de datos en RapidMiner

Después se debe seleccionar el atributo que se busca predecir. Ver **Figura 3.12**

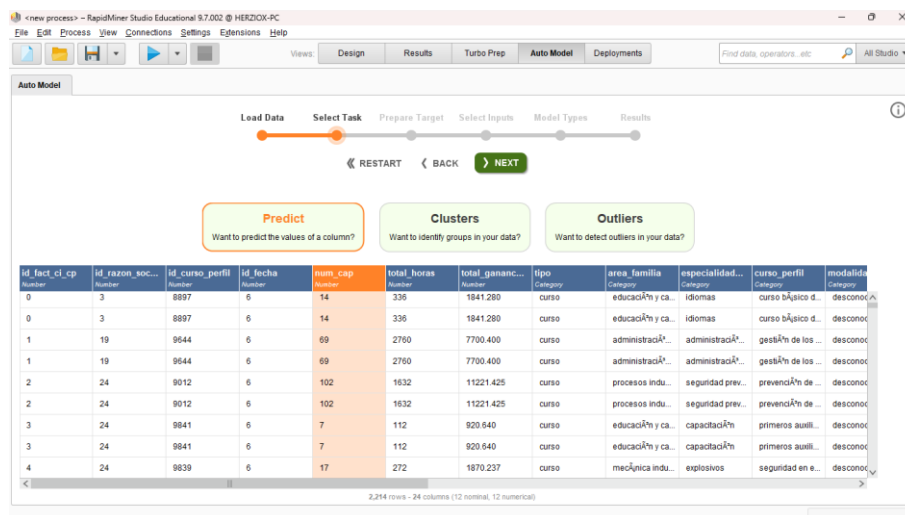


Figura 3.12. Selección de atributo a predecir

Se inicia entonces la preparación de los objetivos en caso de que sea un proceso de regresión o clasificación. Ver **Figura 3.13**

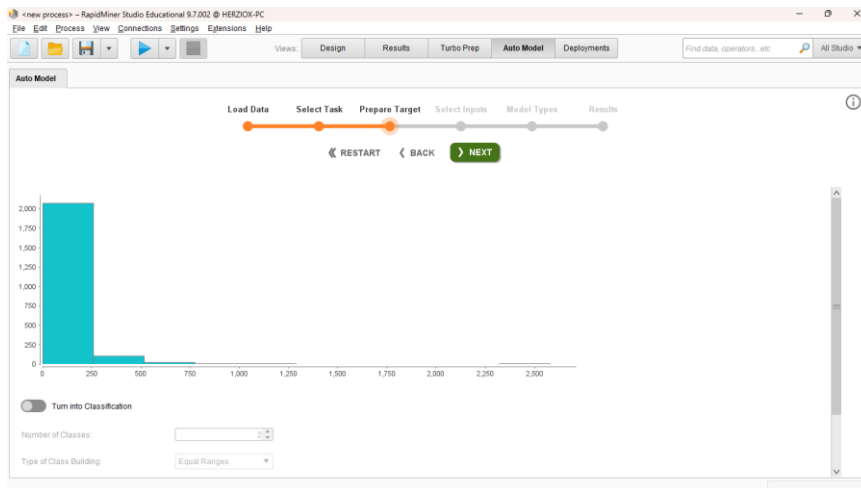


Figura 3.13. Preparación de objetivos

A continuación, se selecciona los atributos o columnas del conjunto de entrenamiento. Estos se deberán seleccionar tomando en cuenta el atributo que se busca predecir. Ver **Figura 3.14**

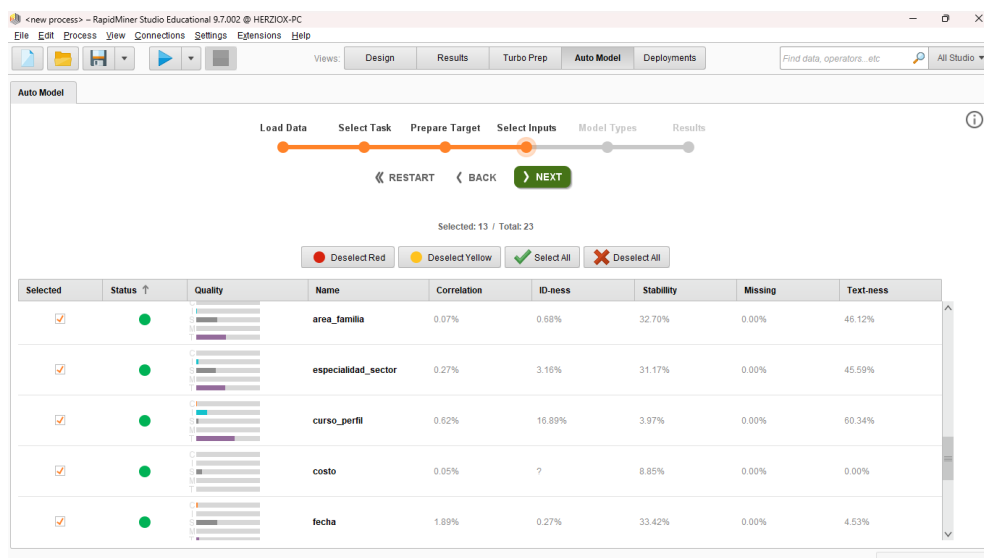


Figura 3.14. Selección de atributos para el entrenamiento

Luego se selecciona los tipos de modelos los cuales se querrá entrenar. Estos deberán ser seleccionados tomando en cuenta los resultados las Fases 2 y 3. Ver **Figura 3.15**

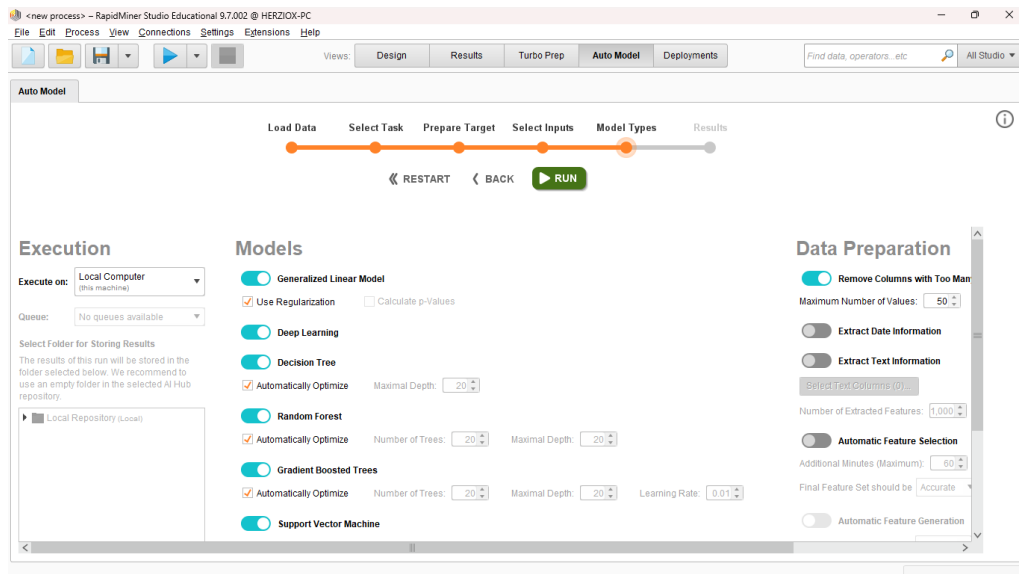


Figura 3.15. Selección de los modelos a entrenar

Después se realiza el proceso de construcción y puntuación de modelos en donde se ve en tiempo real el proceso de entrenamiento y evaluación de los modelos en base a las métricas respectivas, sean de clasificación o regresión. Como se observa en la **Figura 3.16** la evaluación de los modelos, el rendimiento en cuanto el tiempo de ejecución y las métricas de rendimiento. A su vez enlista y puntúa los modelos para su posterior selección por parte del usuario.

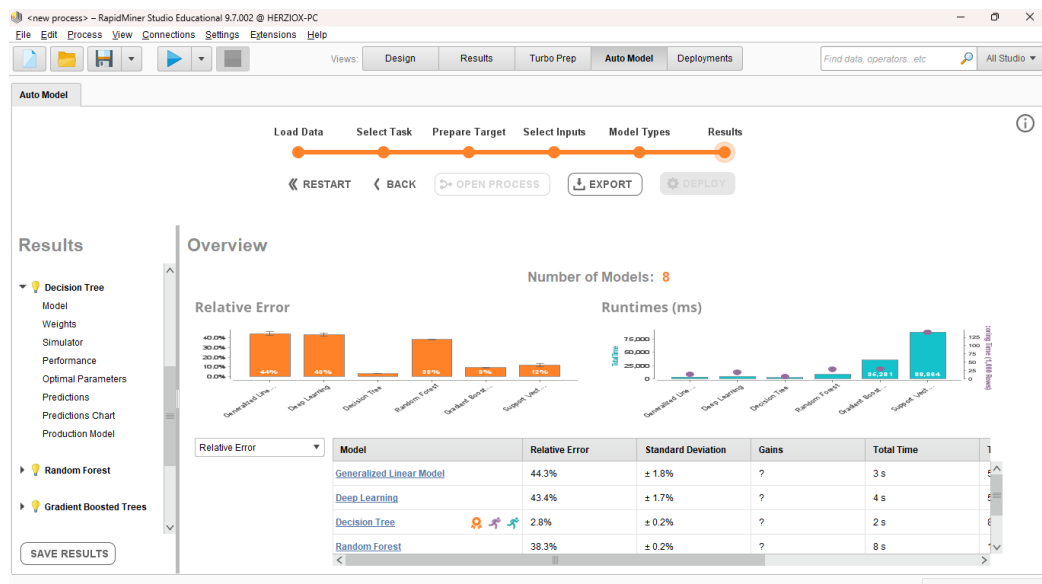


Figura 3.16. Lista de modelos obtenidos

3. **Evaluación de modelos:** La evaluación de los modelos se basó en las puntuaciones arrojadas por RapidMiner así como los resúmenes de los pesos del atributo y la predicción del modelo. Ver **Figura 3.17** y **Figura 3.18**

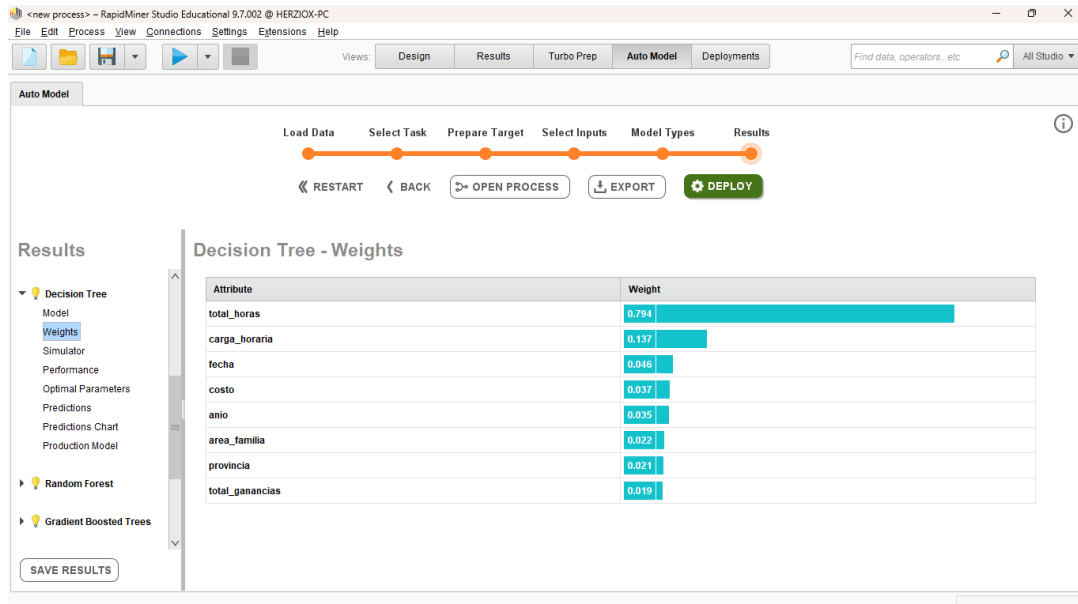


Figura 3.17. Pesos de los atributos de un modelo

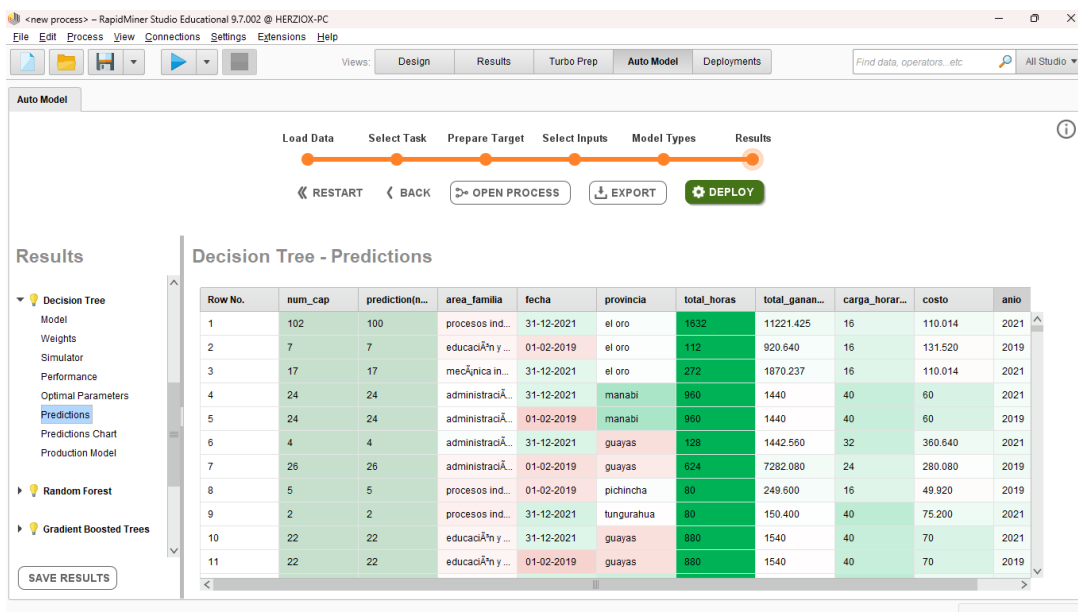


Figura 3.18. Predicciones de un modelo

Tomando en cuenta estos aspectos se procede a desplegar el modelo de producción seleccionado como se ve en la **Figura 3.19**, el cual será almacenado en RapidMiner y MySQL.

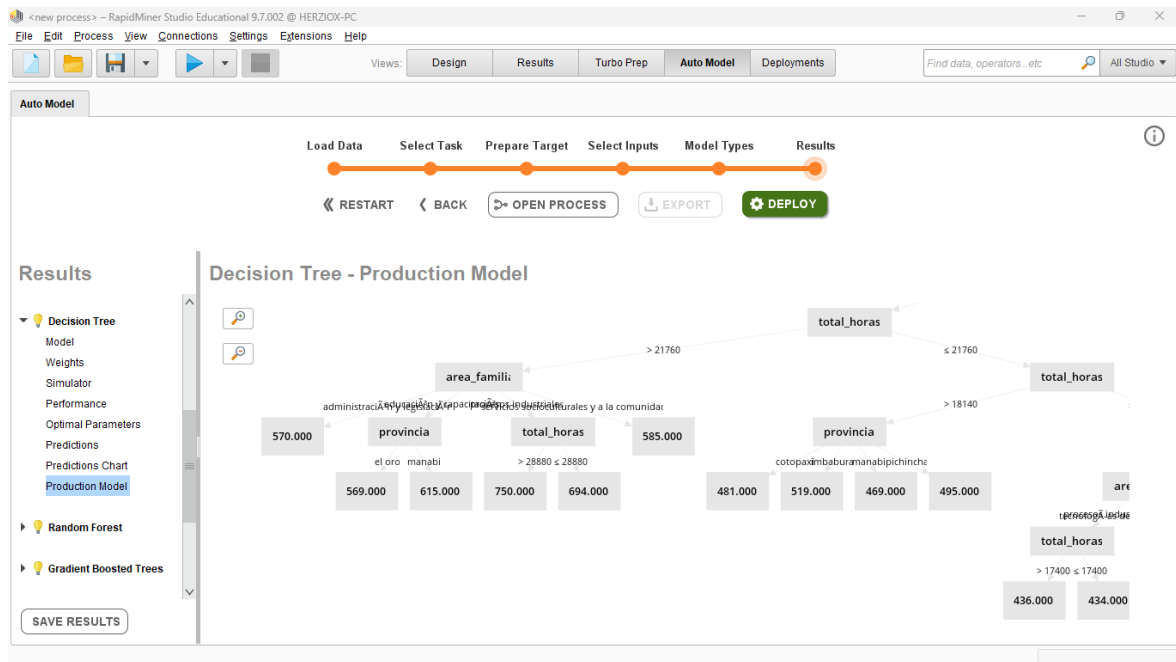


Figura 3.19. Modelo de producción

Se selecciono a todos los modelos con mejor puntuación según RapidMiner en el caso de que las puntuaciones no fueron satisfactorias se retornó la Fase 2, 3 o 4 para tratar los datos y volver a entrenar los modelos. Los modelos seleccionados para cada KPI de cada enfoque por razón social así cómo el proceso de conexión con MySQL para el respaldo de estos modelos se encuentra en el **ANEXO III**.

Para guardar un modelo:

1. se debe asignar su nombre
2. la locación en la que se debe guardar
3. en que despliegue se lo hará
4. de que tipo es (regresión o clasificación).

Los modelos se guardarán en el despliegue correspondiente para futuros procesos de puntuación de datos. Ver **Figura 3.20** y **Figura 3.21**

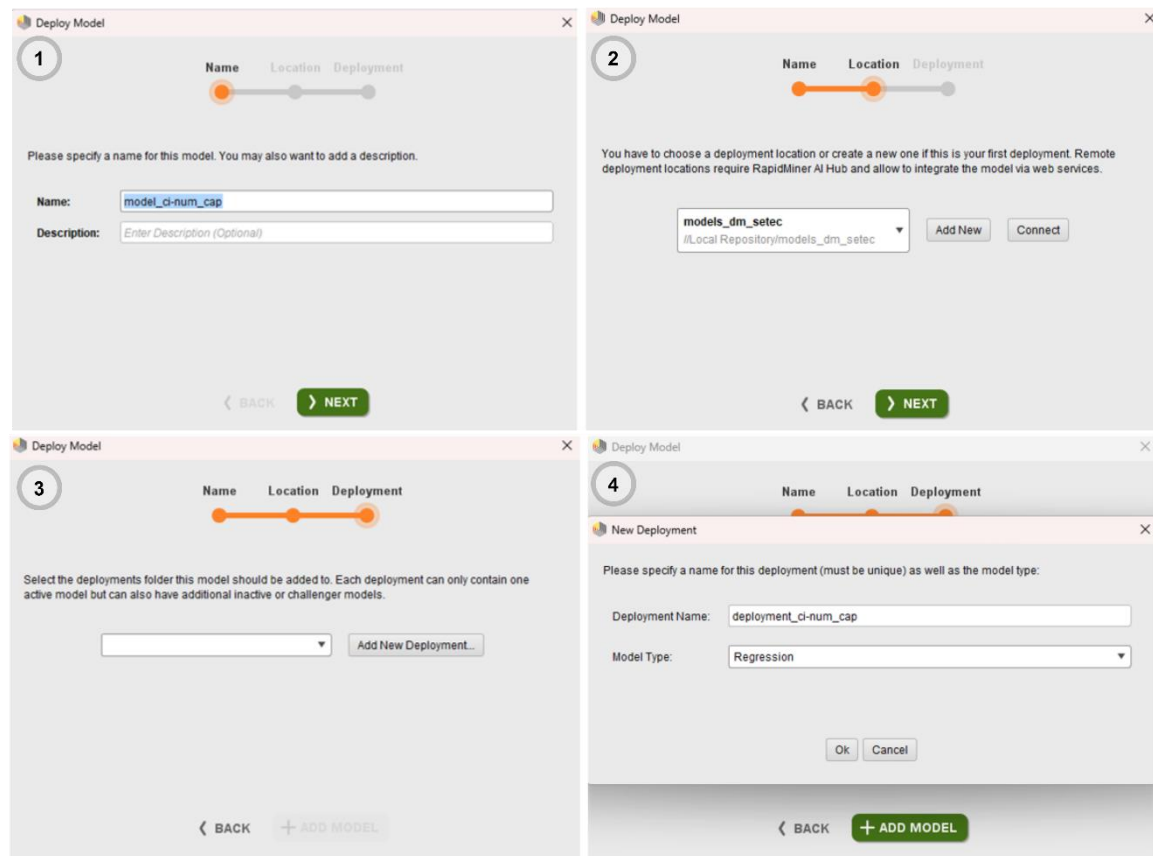


Figura 3.20. Proceso de guardado de modelo en RapidMiner y MySQL

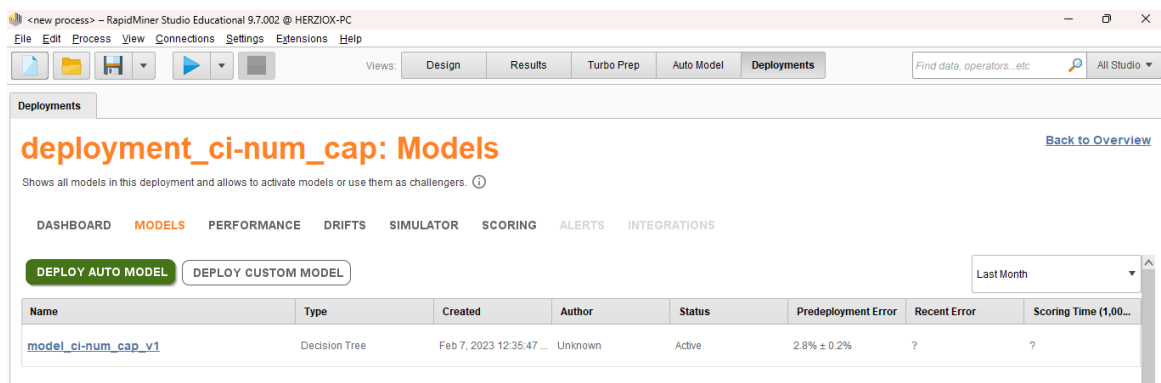


Figura 3.21. Modelo guardado y monitorizado en RapidMiner y MySQL

4. **Evaluation_model.ipynb:** Los datos .CSV en DATAWAREHOUSE son cargados y tratados para generar datos ficticios que simularán ser del año 2023 para así puntuarlos con los modelos desarrollados.
5. **Evaluación de datos simulados:** Los datos simulados de 2023 serán cargados a RapidMiner para después ser evaluados por los modelos guardados como se ve en la **Figura 3.22**.

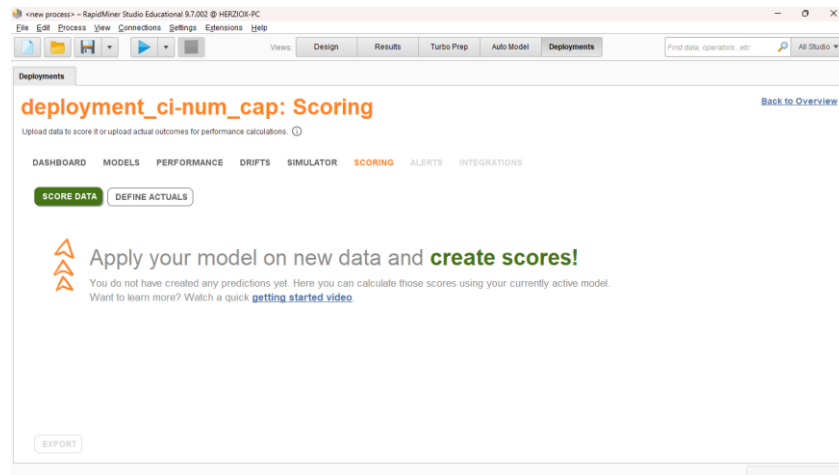


Figura 3.22. Proceso de carga de datos para puntuación de datos simulados

En este proceso se ve cuales atributos serán usados y cuál será el atributo clave a predecir. Ver **Figura 3.23**.

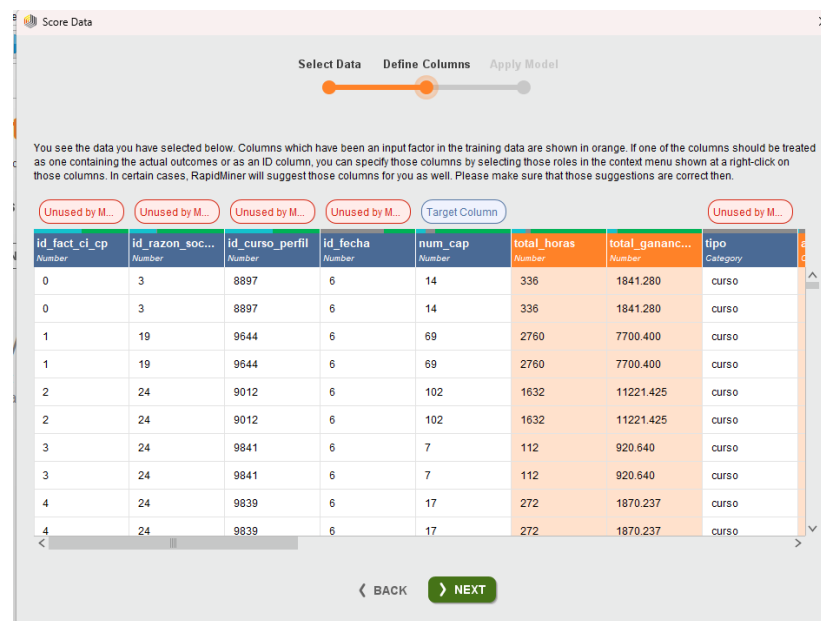


Figura 3.23. Visualización de datos a usar en la predicción.

Al final se tendrá todos los datos evaluados por el modelo. Ver **Figura 3.24**

deployment_ci-num_cap: Scoring

Upload data to score it or upload actual outcomes for performance calculations.

DASHBOARD MODELS PERFORMANCE DRIFTS SIMULATOR **SCORING** ALERTS INTEGRATIONS

SCORE DATA DEFINE ACTUALS

Error Rate: 0.00%

Row No.	num_cap	prediction(num_cap)	M_fact_ci_cp	M_razon_so...	M_curso_pe...	M_fecha	dia	mes	tipo	especialida...	curso_perfil
1	14	14	0	3	8897	6	31	12	curso	idiomas	curso básico...
2	14	14	0	3	8897	6	1	2	curso	idiomas	curso básico...
3	69	69	1	19	9644	6	31	12	curso	administrac...	gestión de lo...
4	69	69	1	19	9644	6	1	2	curso	administrac...	gestión de lo...
5	102	102	2	24	9012	6	31	12	curso	seguridad pr...	prevención d...
6	102	102	2	24	9012	6	1	2	curso	seguridad pr...	prevención d...
7	7	7	3	24	9841	6	31	12	curso	capacitación	primeros aux...
8	7	7	3	24	9841	6	1	2	curso	capacitación	primeros aux...

EXPORT

Figura 3.24. Datos predichos por el modelo

Este proceso se lo tendrá que hacer para cada modelo que predice un KPI diferente los cuales se tendrá que unir para formar un solo conjunto de datos.

3.5 Despliegue:

En esta fase únicamente se agrupo los datos ya puntuados tanto reales como simulados en un solo conjunto de datos por cada tipo de razón social (OC, OEC y CI) con sus respectivos enfoques de Curso/Perfil o Razón Social creando así el Datawarehouse con la información predictiva tanto de datos reales cómo simulados para ser cargados en una base de datos de por medio de una conexión a MySQL y usando scripts de SQL. Ver **Figura 3.25**

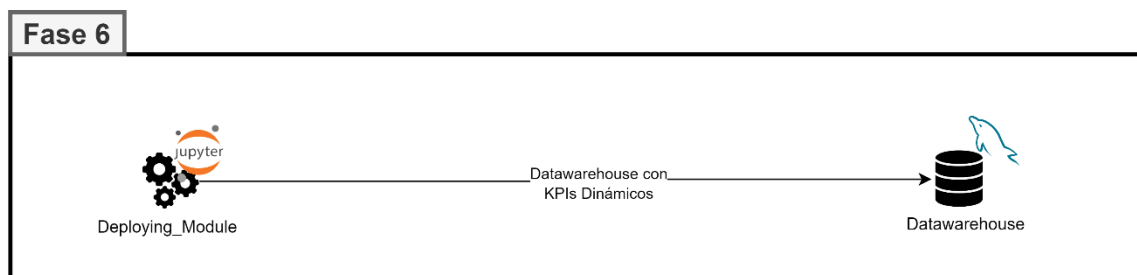


Figura 3.25. Diagrama de Fase 6 - Despliegue

El modelo físico del Datawarehouse final se la puede encontrar en el **ANEXO IV** la cual sigue el mismo diseño inicial realizado en la fase 3, ver **Figura 3.9**

3.6 Visualización:

Siguiendo principios para la creación de dashboards de la Guía para crear dashboards [17]. Se diseñó la visualización de las tablas de hechos de cada tipo de razón social con su

respectivo enfoque destacando cada KPI correspondiente tanto estático como dinámico estableciendo un contraste entre estos a través de distintas gráficas de cada componente relevante de las dimensiones del Datawarehouse. Posteriormente se redactó un manual de uso y explicación para que pueda servir a los evaluadores de usabilidad en la última fase cómo se muestra en la **Figura 3.26**.

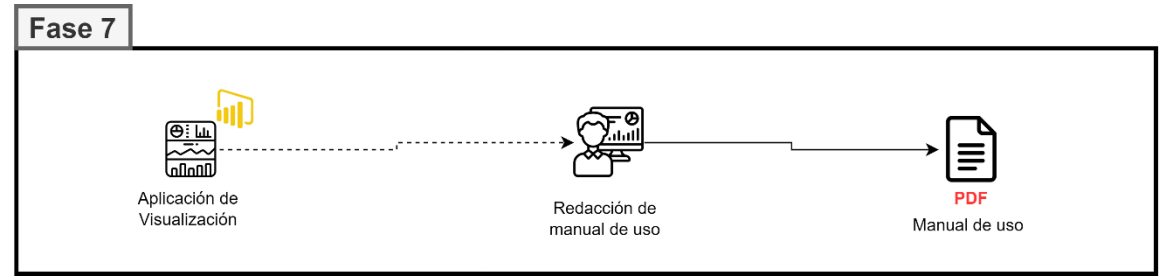


Figura 3.26. Diagrama de fase 7 - Visualización

La visualización cuenta con filtros personalizados para cada tabla de hechos, tablas de resumen en la parte inferior además de división por categorías de cada aspecto de visualización. Como se observa en la **Figura 3.27**

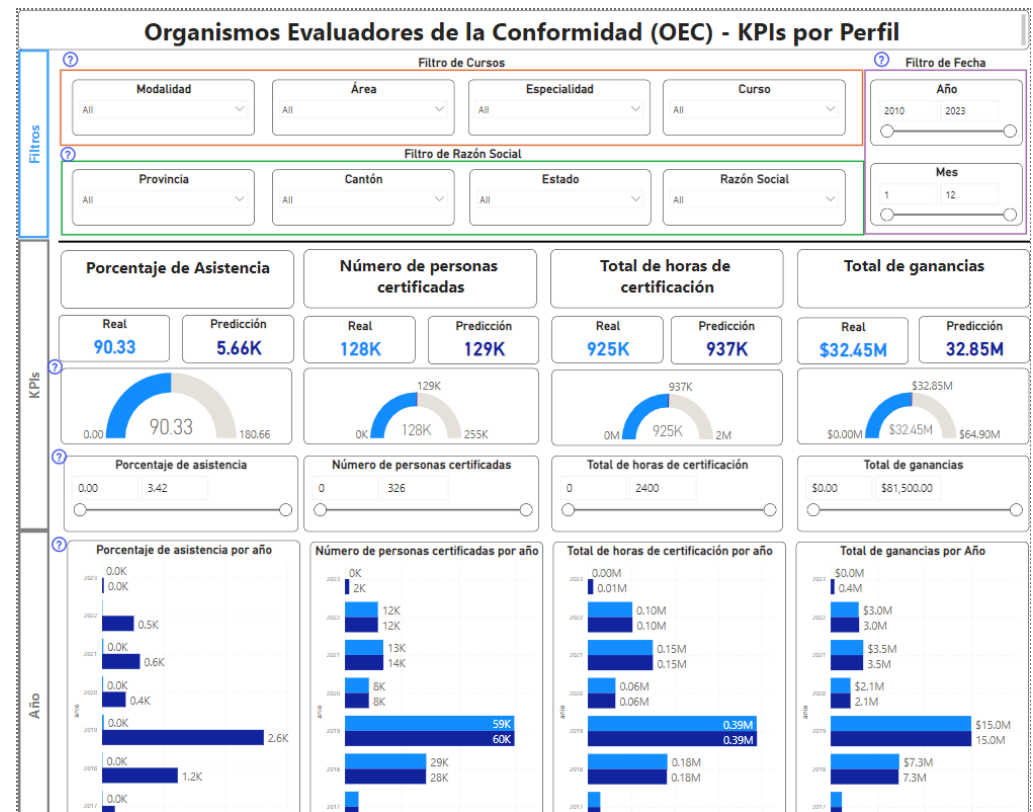


Figura 3.27. Dashboard de OEC con enfoque en perfil

El uso y explicación de esta visualización se lo encuentra en el **ANEXO V**

3.7 Evaluación de usabilidad:

Esta última fase consistió en evaluar los dashboards a través de pruebas de las evaluaciones de usabilidad SUS y heurísticas de Nielsen. La evaluación fue desarrollada por 40 personas especializadas en el área ciencia y análisis de datos o con conocimientos básicos y/o avanzados en tecnologías de la información y comunicación, y educación a quienes se les envió la aplicación, el manual de uso se encuentra en el **ANEXO VI** y el enlace de acceso a la evaluación de usabilidad que se encuentra en el **ANEXO VII**.

3.7.1 Resultados de Nielsen

El número de severidades puntuadas por cada heurística se muestra a continuación en la siguiente **Figura 3.28**

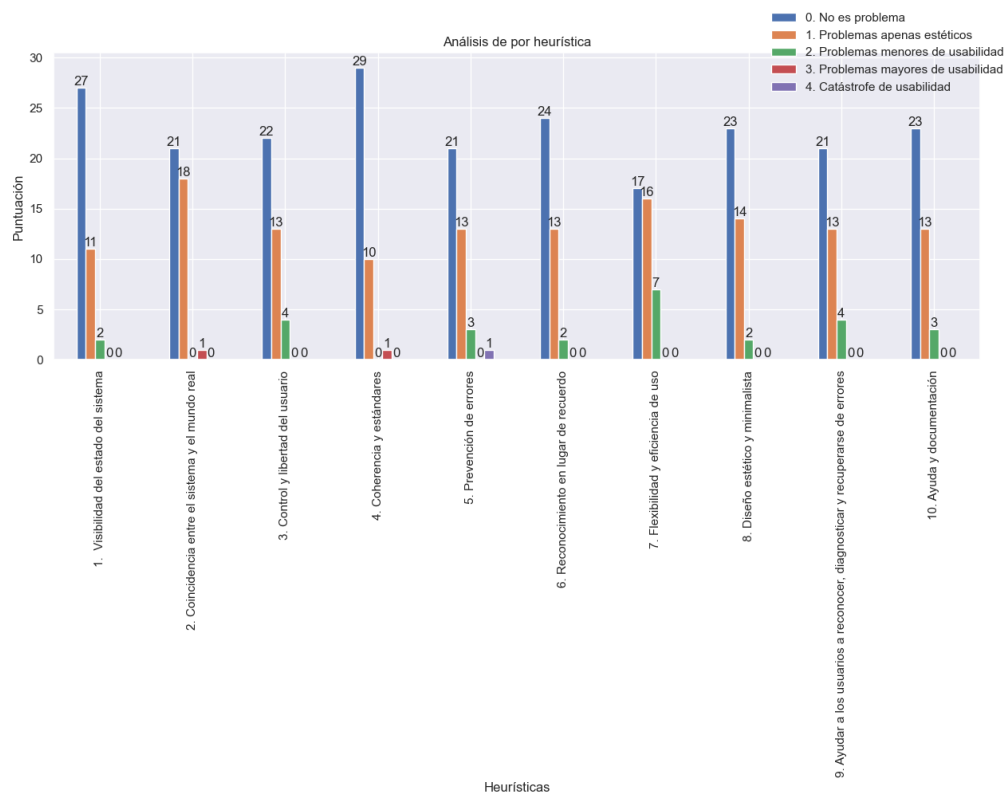


Figura 3.28. Severidad puntuada por heurística

El resumen del promedio de los resultados de cada heurística se presenta en la siguiente **Figura 3.29**.

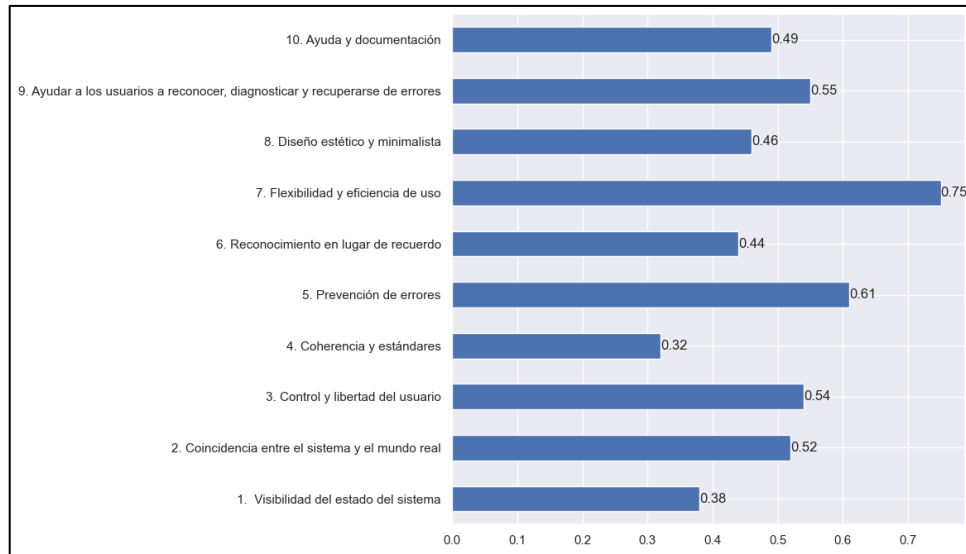


Figura 3.29. Promedio de severidad por heurística

El promedio total de Nielsen fue de: **0.50**

3.7.2 Resultados de SUS

Las calificaciones de cada pregunta se muestran en la siguiente **Figura 3.30**.

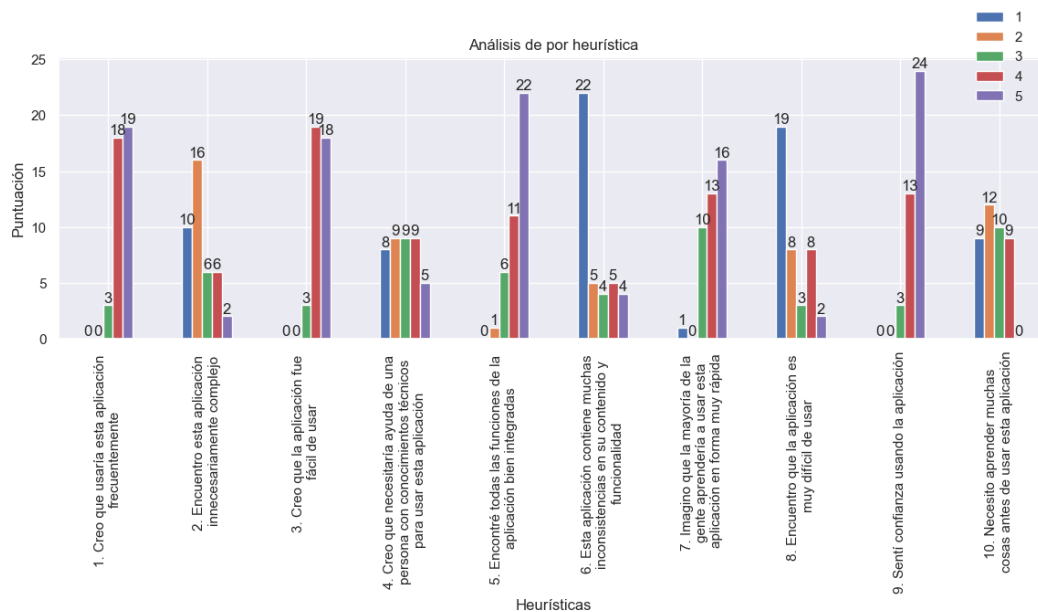


Figura 3.30. calificaciones por pregunta de SUS

El promedio total de SUS fue de: **74.5**

4 ANALISIS Y DISCUSIÓN DE LOS RESULTADOS

Terminado las 6 fases de la metodología de CRISP-DM más las fases de visualización y evaluación de usabilidad se puede comprobar los siguientes resultados:

Fase 1: La demanda de cursos y/o perfiles de cualificación profesional estará clasificada por las razones sociales que imparten estas cualificaciones. Las razones sociales pueden ser de tipo OC, OEC o CI, la demanda tendrá enfoques por razones sociales y por curso/perfil. El plan creado tuvo modificaciones en el desarrollo del proceso de minería de datos.

Fase 2: El volumen de datos obtenidos en los módulos PCCO, PCOEC y PCCI necesarios para el análisis de la demanda están entre 140 mil a más de 300 mil datos. En cuanto a la calidad esta fue variable ya que se contó con múltiples datos nulos cómo se lo comprueba en el **ANEXO II**.

Fase 3: El número de registro de las tablas de hecho se redujo debido a las agrupaciones de número de capacitados y/o certificados. Los datos fueron tratados en distintos aspectos, en caso de que se hallasen datos nulos se los completo con valores promedio, cero o un valor que significase que el dato es nulo.

Fase 4: Los modelos tuvieron métricas proporcionales en cuanto a la cantidad de datos y atributos seleccionados para su entrenamiento. Estos modelos estuvieron diseñados para cada KPI de cada enfoque.

Fase 5: Estos modelos fueron seleccionados por sus métricas de rendimiento, obteniendo un modelo por cada KPI de cada tabla de hechos por enfoque resultando en 25 modelos (7 de CI, 9 de OC y 9 de OEC)

Fase 6: Los datos predictivos fueron cargados al Datawarehouse para finalmente cargar a MySQL creando a su vez el modelo entidad relación para su futuro uso en la visualización.

Fase 7: La visualización contó filtros, indicadores estáticos y dinámicos, y gráficas para cada aspecto de las dimensiones que componen el modelo entidad relación del Datawarehouse para el análisis de la demanda.

Fase 8:

- El puntaje obtenido en Nielsen mostró un promedio menor a 1 lo que significaría que se tiene un problema muy pequeño en cuanto a la estética.
- Para obtener el puntaje total, los resultados son calculados con la siguiente formula:

El puntaje total fue de **74.5** que representa un rango de aceptabilidad **Alto un grado C** equivalente a **Bueno** Siguiendo las escalas y puntuaciones de la **Figura 2.2**.

5 CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

5.1 Conclusiones

5.1.1 Entendimiento del negocio:

- La demanda de cursos y perfiles de cualificación a nivel nacional ha ido en aumento debido a las oportunidades laborales que requieren una certificación o especialización profesional en un área en específico.
- Esta demanda se puede ser medida utilizando los datos de la Secretaría Nacional de Cualificaciones y Capacitación Profesional (SETEC) por medio de Indicadores de rendimiento como son el número de capacitados o certificados.
- El plan de proyecto puede modificarse a medida que se desarrolla y los cambios deberán justificar el resultado final de todo el proceso de minería de datos.

5.1.2 Entendimiento de Datos:

- Los datos recolectados fueron obtenidos del portal web SETEC y fuentes externas para complementar y estimar datos requeridos en el proceso de análisis de la demanda.
- Estos datos tuvieron un volumen considerable, pero distaron de la estimación antes realizada en algunos módulos esto pudo deberse al proceso de web scraping el cual tuvo complicaciones con la página como se explica en el **ANEXO II**.
- La calidad de los datos fue variable dependiendo de los módulos y estos requerían un tratamiento adecuado para cada caso.

5.1.3 Evaluación de usabilidad:

- La usabilidad respecto a Nielsen mostró un valor menor a 1 correspondiente a un problema estético que requiere pocas modificaciones. Por otro lado
- SUS mostró que la visualización es buena pero no está en el rango de excelente ya las calificaciones señalaron que los usuarios sienten que deben aprender muchas cosas o requerirían asistencia técnica para ocupar esta aplicación.

5.1.4 Preparación de datos:

- Los datos fueron tratados por medio de distintas técnicas de limpieza y transformación creando así un Datawarehouse.
- Datos como el costo o el número de convocados tuvieron que ser estimados a través de datos adicionales para complementar los datos recolectados para su posterior análisis y construcción de modelos.

5.1.5 Modelamiento:

- La construcción de modelos fue proceso rápido gracias a la función de Auto Model de RapidMiner la cual dependió de la cantidad de datos de entrenamiento para los modelos.

5.1.6 Evaluación:

- Las métricas de rendimiento de los modelos presentaron mejores resultados en cuanto a la cantidad y calidad de los datos con que fueron entrenados cada modelo seleccionando así los modelos con mejor eficiencia y rendimiento.

5.1.7 Despliegue:

- Se evaluó datos simulados del año 2023 y unirlos con los datos reales obtenidos para cargarlos en el Datawarehouse.
- El Datawarehouse con datos predictivos y sus modelos fueron guardados y monitorizados en las herramientas de MySQL y RapidMiner para ser visualizados a través de la herramienta de PowerBI

5.1.8 Visualización:

- El Datawarehouse con los datos predictivos simulados fue cargado en la herramienta de PowerBI para la creación de dashboards que mostrasen los KPIs estáticos y dinámicos de cada tipo de razón social con su respectivo enfoque y su variación respecto a los atributos de las dimensiones.

5.2 Recomendaciones

- Se debe tener un análisis y contraste de los datos de otras organizaciones de organización como el SECAP para tener un análisis mayor además de los datos de SETEC.
- Los KPIs y objetivos del proceso de minería tienen que estar bien definidos en referencia a los datos que se tenga disponibles esto para evitar iteraciones extra innecesarias en el proceso.

- Se debe tener claro las fuentes de datos y tener en cuenta cuales son los datos que se pueden recolectar además de su relevancia para los objetivos del proceso de minería.
- Si los datos recolectados no son satisfactorios se tiene que modificar el plan de proyecto, los objetivos o indicadores de ser necesario. Esto es posible gracias a la flexibilidad de la metodología.
- La preparación de datos tiene que ser realizada con todos los datos ya recolectados para realizar las transformaciones correctas, evitar procedimientos extra y construir los datos finales más rápido.
- Los modelos deben ser entrenados con una buena cantidad y calidad de datos para obtener mejores resultados en sus métricas de rendimiento y futuras predicciones.
- La evaluación tiene que ser escogiendo las métricas correctas para evitar escoger modelos de baja calidad.
- La visualización de los datos debe estar adaptada a los datos disponibles para evitar sobrecarga de información innecesaria
- La aplicación visual está centrada en análisis de datos de modo que la evaluación de usabilidad tiene que ser enviada a personas que tengan conocimientos en datos o tecnologías de la información.

5.3 Trabajos futuros

- Se debería investigar de muchas más fuentes oficiales para tener una mayor cantidad y calidad de datos.
- El script de web scraping, así como scripts de tratamiento de datos en formato ipynb que deberán integrarse y combinarse para reducir procesos repetitivos y automatizar el sistema.
- Se debería desarrollar una aplicación propia en Python u otro lenguaje de programación capaz de crear y evaluar modelos de Aprendizaje automático sin necesidad de usar RapidMiner la cual es una herramienta de pago.
- Se deberá considerar una alternativa extra para la visualización de datos además de PowerBI para compartir mucho más fácil la aplicación.
- Consolidar todos los scripts de las fases 2,3,4,5,6 y 7 en una sola la aplicación.

6 REFERENCIAS BIBLIOGRÁFICAS

- [1] IBM, «CRISP-DM Help Overview,» BM, 17 08 2021. [En línea]. Available: <https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview>. [Último acceso: 25 12 2022].
- [2] B. Beltrán, «MINERÍA DE DATOS,» de *Minería de datos*, Puebla, Benemérita Universidad Autónoma de Puebla , 2016, p. 15.
- [3] J. C. Díaz, *Introducción al Business Intelligence*, Barcelona: El Ciervo 96, S.A, 2010.
- [4] Microsoft, «What is Power BI?,» Microsoft power BI, 2022. [En línea]. Available: <https://powerbi.microsoft.com/en-us/what-is-power-bi/>. [Último acceso: 25 12 2022].
- [5] RapidMiner, «Why RapidMiner?,» 23 02 2022. [En línea]. Available: <https://rapidminer.com/why-rapidminer/>.
- [6] L. A. Bustamante Jiménez y G. J. Zapata Vera, *Desarrollo de un sistema web de apoyo a la toma de decisiones aplicando algoritmos de aprendizaje automático para la predicción de supervivencia en pacientes con hepatitis*, Guayaquil: Universidad de Guayaquil, 2022.
- [7] Selenium, «WebDriver,» 20 02 2023. [En línea]. Available: <https://www.selenium.dev/documentation/webdriver/>.
- [8] Jupyter, «About Us Project Jupyter's origins and governance,» 20 02 2023. [En línea]. Available: <https://jupyter.org/about>.
- [9] Microsoft, «Getting Started,» Visual Studio Code, 2023. [En línea]. Available: <https://code.visualstudio.com/docs>. [Último acceso: 22 02 2023].
- [10] A. Chamorro y C. Escobar, «Características generales,» de *Introducción al modelamiento de bases de datos y SLQ básico para Bibliotecarios.*, Santiago, Departamento de Gestión de Información de la Universidad Tecnológica Metropolitana, 2008, p. 22.
- [11] MySQL, «MySQL,» MySQL Workbench, 2022. [En línea]. Available: <https://www.mysql.com/products/workbench/>. [Último acceso: 26 12 2022].

- [12] E. Bonifaz, «Web Scraping,» de *WEB SCRAPING PARA ANÁLISIS DE LOS DATOS DEL PERSONAL DEL MINISTERIO DE EDUCACIÓN EN EL PERIODO 2015-2021*, Quito, ESCUELA POLITÉCNICA NACIONAL, 2023, pp. 15-16.
- [13] Amazon Web Services, «Conceptos relacionados con el almacenamiento de datos,» Amazon Web Services, 2022. [En línea]. Available: <https://aws.amazon.com/es/data-warehouse/>. [Último acceso: 06 10 2022].
- [14] H. Criollo, «IMPLEMENTACIÓN DE UN SISTEMA BUSINESS INTELLIGENCE BASADO EN KEY PERFORMANCE INDICATORS PARA LA EMPRESA DELIMARKET DE LA CIUDAD DE PÍLLARO,» de *Key Performance Indicators*, Ambato, Pontificia Universidad Católica del Ecuador , 2018, pp. 29-32.
- [15] Universidad siblo 21, «Indicadores estáticos y dinámicos,» Universidad siblo 21, Córdoba, 2020.
- [16] E. LUCARNO, ELABORACIÓN DE UN TABLERO DE CONTROL DE GESTIÓN EFECTIVA DE PROCESOS Y PERSONAS, APLICABLE A INSTITUCIONES EDUCATIVAS DE GESTIÓN PRIVADA DE CÓRDOBA, Cordoba: UNIVERSIDAD SIGLO XXI , 2021.
- [17] Juice, Inc, A Guide to Creating Dashboards, Juice, Inc, 2009.
- [18] K. Moran, «Usability Testing 101,» Nielsen Norman Group, 1 12 2019. [En línea]. Available: <https://www.nngroup.com/articles/usability-testing-101/>. [Último acceso: 22 02 2023].
- [19] J. Nielsen, «10 Usability Heuristics for User Interface Design,» Nielsen Norman Group, 24 04 1994. [En línea]. Available: nngroup.com/articles/ten-usability-heuristics/. [Último acceso: 22 02 2023].
- [20] J. Brooke, «SUS: A quick and dirty usability scale,» Redhatch Consulting Ltd., United Kingdom, 1995.
- [21] B. Aaron, K. Phil y M. James, «Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale,» *J. Usability Stud.*, vol. 4, pp. 114-123, 2009.
- [22] INEC, «Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), enero 2022,» INEC, Quito, 2022.

- [23] SECAP, «Certificación de personas por competencias laborales en otros sectores,» SECAP, 2022. [En línea]. Available: <https://www.gob.ec/secap/tramites/certificacion-personas-competencias-laborales-otros-sectores>. [Último acceso: 07 10 2022].
- [24] ecuadorenvivo, «Ministerio de Trabajo calificó a nuevos Operadores de Capacitación,» ecuadorenvivo, 23 09 2022. [En línea]. Available: <https://www.ecuadorenvivo.com/index.php/economia/item/150135-ministerio-de-trabajo-califico-a-nuevos-operadores-de-capacitacion>. [Último acceso: 07 10 2022].
- [25] Educación, Ministerio de; Trabajo, Ministerio de, Plan Nacional de Educación y Formación Técnica y Profesional, Quito: Ministerio de Educación, 2021.
- [26] CEC EPN, «Centro de Educación Continua EPN,» CEC EPN, 2022. [En línea]. Available: <https://www.cec-epn.edu.ec/>. [Último acceso: 15 11 2022].
- [27] L. Carvajal, Metodología de la Investigación Científica. Curso general y aplicado, 28 ed., Santiago de Cali: U.S.C., 2006, p. 139.

7 ANEXOS

- ANEXO I [Arquitectura de Sistema de Minería de Datos](#)
- ANEXO II [Fase 1 y 2 Entendimiento de Negocio y datos](#)
- ANEXO III [Fase 4 y 5 Modelamiento y Evaluación](#)
- ANEXO IV [Modelo Físico del Datawarehouse en MySQL](#)
- ANEXO V [Manual de uso y explicación de los tableros \(dashboards\)](#)
- ANEXO VI [Resumen de pruebas de usabilidad](#)
- ANEXO VII [Prueba de Usabilidad Nielsen y SUS](#)
- ANEXO VIII [Costos Oferta Educación continua](#)
- ANEXO IX [Data-Mining-System-SETEC](#)