**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**

**NOMBRE DEL COMPONENTE**

**INGENIERO CIENCIAS DE LA COMPUTACIÓN**

**SERGIO ANDRÉS JIMÉNEZ REINO**

**DIRECTOR: 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.

**  
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

Julián Galindo

NOMBRE\_COLABORADOR(ES)

DEDICATORIA

El presente documento dirigida a mi familia quien me ha apoyado incondicionalmente en los aspectos de mi vida académica.

AGRADECIMIENTO

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 e inspirado en ser mejor cada día.

A profesores y compañeros quienes me enseñaron, inspiraron 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.

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

A Johan que me dijo que la carrera de Sistemas se llama ahora Computación y terminé 6 años aquí.

A Jimmy que me enseñó que no es necesario entrar a clases, hacer proyectos o dar exámenes para aprobar una materia.

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

ÍNDICE DE CONTENIDO

Contents

[CERTIFICACIONES I](#_Toc127984541)

[DECLARACIÓN DE AUTORÍA III](#_Toc127984542)

[DEDICATORIA IV](#_Toc127984543)

[AGRADECIMIENTO V](#_Toc127984544)

[ÍNDICE DE CONTENIDO VI](#_Toc127984545)

[RESUMEN X](#_Toc127984546)

[ABSTRACT XI](#_Toc127984547)

[1 DESCRIPCIÓN DEL COMPONENTE DESARROLLADO 1](#_Toc127984548)

[1.1 Objetivo general 1](#_Toc127984549)

[1.2 Objetivos específicos 1](#_Toc127984550)

[1.3 Alcance 2](#_Toc127984551)

[1.4 Marco teórico 4](#_Toc127984552)

[1.4.1 Inteligencia de negocios: 4](#_Toc127984553)

[1.4.2 CRISP-DM: 4](#_Toc127984554)

[1.4.3 Ciencia de Datos: 5](#_Toc127984555)

[1.4.4 Microsoft Power BI: 5](#_Toc127984556)

[1.4.5 RapidMiner: 5](#_Toc127984557)

[1.4.6 Python: 6](#_Toc127984558)

[1.4.7 Selenium 6](#_Toc127984559)

[1.4.8 Entornos de programación 6](#_Toc127984560)

[1.4.9 Lenguaje de programación en R 7](#_Toc127984561)

[1.4.10 SQL: 7](#_Toc127984562)

[1.4.11 MySQL Workbech: 7](#_Toc127984563)

[1.4.12 DataWarehouse: 8](#_Toc127984564)

[1.4.13 KPIs: 8](#_Toc127984565)

[1.4.14 Objetivos SMART 8](#_Toc127984566)

[2 Metodología 9](#_Toc127984567)

[2.1 CRISP-DM: 9](#_Toc127984568)

[3 Desarrollo e implementación 11](#_Toc127984569)

[3.1 Entendimiento del negocio: 11](#_Toc127984570)

[3.1.1 Análisis y entendimiento de los módulos del portal web SETEC 11](#_Toc127984571)

[3.1.2 Análisis y entendimiento del mercado laboral nacional 13](#_Toc127984572)

[3.1.3 Diseño de objetivos de negocio e indicadores estáticos y dinámicos 14](#_Toc127984573)

[3.1.4 Producción del plan de proyecto 16](#_Toc127984574)

[3.2 Entendimiento de datos: 18](#_Toc127984575)

[3.2.1 Análisis y entendimiento de los datos en los módulos del portal web SETEC 18](#_Toc127984576)

[3.2.2 Extracción y colección de datos 34](#_Toc127984577)

[3.2.3 Descripción de datos 41](#_Toc127984578)

[3.2.4 Análisis exploratorio de datos 49](#_Toc127984579)

[3.2.5 Verificación de la calidad de datos 49](#_Toc127984580)

[3.3 Preparación de datos: 49](#_Toc127984581)

[3.3.1 Selección de datos 49](#_Toc127984582)

[3.3.2 Limpieza de datos 49](#_Toc127984583)

[3.3.3 Transformación de datos 49](#_Toc127984584)

[3.3.4 Formateo de datos 49](#_Toc127984585)

[3.3.5 Construcción de nuevos datos 49](#_Toc127984586)

[3.3.6 Integración de datos 49](#_Toc127984587)

[3.3.7 Diseño del Datawarehouse 49](#_Toc127984588)

[3.3.8 Creación de modelos entidad relación 49](#_Toc127984589)

[3.3.9 Almacenamiento de datos 49](#_Toc127984590)

[3.4 Modelamiento: 49](#_Toc127984591)

[3.4.1 Selección de técnicas de modelamiento 49](#_Toc127984592)

[3.4.2 Generación de diseño de pruebas 49](#_Toc127984593)

[3.4.3 Construcción de modelos 49](#_Toc127984594)

[3.4.4 Evaluación de modelos 49](#_Toc127984595)

[3.5 Evaluación: 56](#_Toc127984596)

[3.5.1 Evaluación del resultado de los modelos desarrollados en base a los objetivos planteados 56](#_Toc127984597)

[3.5.2 Carga de información predictiva en el Datawarehouse 56](#_Toc127984598)

[3.5.3 Revisión de las actividades y decisiones tomadas en las anteriores fases 56](#_Toc127984599)

[3.5.4 Determinación de los siguientes pasos 56](#_Toc127984600)

[3.6 Despliegue: 56](#_Toc127984601)

[3.6.1 Diseño de plan de despliegue de modelos 56](#_Toc127984602)

[3.6.2 Plan de monitorización y mantenimiento 56](#_Toc127984603)

[3.6.3 Despliegue de los modelos de en herramientas de visualización de datos 56](#_Toc127984604)

[3.7 Visualización: 63](#_Toc127984605)

[3.7.1 Selección de gráficos para visualización de datos y resultados 63](#_Toc127984606)

[3.7.2 Creación de gráficos estáticos y dinámicos 63](#_Toc127984607)

[3.7.3 Creación de tableros de gráficos explicativos de los datos y resultados 63](#_Toc127984608)

[3.8 Evaluación de usabilidad: 63](#_Toc127984609)

[3.8.1 Evaluación de la usabilidad por medio de las métricas de usabilidad de Nielsen. 63](#_Toc127984610)

[3.8.2 Evaluación de la usabilidad por medio de SUS. 63](#_Toc127984611)

[4 ANALISIS Y DISCUSIÓN DE LOS RESULTADOS, CONCLUSIONES Y RECOMENDACIONES 64](#_Toc127984612)

[4.1 Pruebas 64](#_Toc127984613)

[4.2 Resultados 64](#_Toc127984614)

[5 CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS 65](#_Toc127984615)

[5.1 Conclusiones 65](#_Toc127984616)

[5.2 Recomendaciones 65](#_Toc127984617)

[5.3 Trabajos futuros 65](#_Toc127984618)

[ANEXO I 69](#_Toc127984619)

[6.1.1 ASPECTOS DE FORMATO 70](#_Toc127984620)

[6.1.2 Tipografía 70](#_Toc127984621)

[6.1.3 70](#_Toc127984622)

[6.1.4 Espaciado 70](#_Toc127984623)

[6.1.5 70](#_Toc127984624)

[6.1.6 Numeración 70](#_Toc127984625)

[6.2.1 CI num\_cap 71](#_Toc127984626)

[6.2.2 CI total\_horas 71](#_Toc127984627)

[6.2.3 CI Total\_ganancias 71](#_Toc127984628)

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 para construcción de sistemas 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 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 desplegó 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 del sistema fue evaluada en base a SUS (System Usability Scale) y las 10 herurísticas de Nielsen de usabilidad, donde 30 personas entre expertos en áreas de datos y conocimientos moderados en análisis de datos y tecnologías de la información puntuaron a aplicación y se determino una aplicación con

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

ABSTRACT

(Máximo 250 palabras)

**KEYWORDS:** word1, word2, …, word6.

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.

* 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. 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 [4]. 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 activades, 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:

1. Visualización: visualización de los resultados a través de tableros
2. Evaluación de usabilidad: análisis de la usabilidad del sistema de minería de datos desarrollado en base a los principios de Nielsen y SUS.
   1. 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 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. Marco teórico

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. [1] Su uso abarca todos los aspectos de una empresa u organización, desde la planificación y el análisis financiero hasta la toma de decisiones de marketing y la gestión de operaciones, y puede ser útil para mejorar la eficiencia y la productividad, reducir costos y aumentar la satisfacción del cliente. [1]

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 un acrónimo de Cross-Industry Standard Process for Data Mining. [2] Se ha convertido en un estándar de la industria y es ampliamente en el proceso de análisis de datos y la toma de decisiones basadas en datos.

El modelo CRISP-DM incluye seis fases:

1. Comprensión del negocio: se define el problema de negocio o la oportunidad de negocio que se quiere abordar y se establecen los objetivos del proyecto.
2. Recopilación de datos: se recopilan y seleccionan los datos relevantes para el proyecto.
3. Preparación de datos: se limpian y se preparan los datos para el análisis.
4. Modelado: se utilizan técnicas de análisis de datos para construir modelos que puedan ayudar a predecir resultados o entender tendencias.
5. Evaluación: se evalúan los resultados y se determina si se han alcanzado los objetivos del proyecto.
6. Implementación: se toman medidas en base a los resultados del proyecto y se implementan cambios en el negocio.

CRISP-DM es un modelo flexible y puede ser adaptado a diferentes proyectos y entornos. Es importante tener en cuenta que el modelo no es lineal y es posible que se tenga que volver a visitar ciertas fases varias veces a lo largo del proceso de minería de datos.

Ciencia de Datos:

La ciencia 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. Se aplica en una variedad de campos, como la salud, la financiación, la publicidad y el gobierno, para mejorar la eficiencia y el rendimiento mediante la toma de decisiones informadas.

Microsoft Power BI:

Power BI 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. Es una herramienta escalable y fácil de usar para el autoservicio. [3]

RapidMiner:

Plataforma de ciencia de datos que ayuda a acelerar los proyectos de ciencia de datos y mejorar la competitividad empresarial permitiendo permite 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. [4]

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 frameworks disponibles que facilitan el proceso de desarrollo. [5]

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 controlar el navegador y ejecutar pruebas. Permite simular las interacciones de un usuario con un sitio web o aplicación móvil, lo que facilita la realización de pruebas automatizadas, incluyendo la extracción de datos en procesos de ciencia de datos. Selenium cuenta con librerías creadas para diferentes entornos y lenguajes de programación, como Python. [6]

Entornos de programación

#### 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. [7] Cuenta con una versión de escritorio y web que se utiliza para escribir y ejecutar código, así como para crear documentos interactivos que combinan código, texto, imágenes y gráficos.

#### 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.

#### Google Colab

Es un servicio de Google que proporciona un entorno de desarrollo en la nube para escribir y ejecutar código en Python y capaz de conectarse con Google Drive para el almacenamiento de archivos. Colab es especialmente útil para el análisis de datos y el aprendizaje automático, ya que proporciona acceso a GPUs y TPUs de alta potencia para acelerar el procesamiento.

SQL:

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

MySQL Workbech:

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 el servidor, incluyendo la configuración del servidor, la gestión de usuarios y la realización de copias de seguridad. [8]

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 no relacionales, bases de datos no relaciones, documentos, sitios web, entre otros. El almacenamiento de datos es fundamental para el análisis de los datos de una organización o mercado a través de la creación de informes, paneles y otras herramientas de análisis. [9]

KPIs:

KPI es el acrónimo de Key Performance Indicator (Indicador Clave de Rendimiento), 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 un proyecto en relación con un objetivo establecido. Los KPIs se utilizan a menudo para medir el éxito de una estrategia o para tomar decisiones de negocio. Estos pueden ser numéricos, de progreso y/o cambio y requieren una medición, un objetivo, una fuente y una frecuencia de medición.

Objetivos SMART

Los criterios SMART propuestos por George T. Doran para la definición de indicadores clave de desempeño vienen del Specific, Measurable, Achievable, Realistic y Time-bound. [10] 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. Debe tener una definición clara y concreta, evitando términos vagos o ambiguos que puedan generar malentendidos o interpretaciones erróneas. [10]
2. Medible (Measurable): El indicador debe ser cuantificable y se debe poder medir de manera objetiva y precisa. Es necesario contar con un sistema de medición que permita recolectar los datos necesarios para evaluar el progreso del indicador. El KPI debe ser expresado en una unidad de medida clara y definida. [10]
3. Alcanzable (Achievable): El KPI debe ser realista y posible de alcanzar en función de las capacidades, recursos y limitaciones de la organización. No debe ser ni demasiado fácil ni demasiado difícil, sino que debe ser desafiante pero alcanzable, para motivar a las personas a trabajar en pos de su logro. [10]
4. Realista (Realistic): El KPI debe ser relevante y estar alineado con los objetivos y estrategias de la organización. Debe tener una relación clara con el contexto y la situación actual de la organización, y estar alineado con los valores, la misión y la visión de la misma. Es importante que sea factible de alcanzar en un marco de tiempo realista y que sea sostenible en el tiempo. [10]
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. Esto permite establecer una fecha límite para alcanzar el objetivo, lo que ayuda a motivar a las personas a trabajar en pos de su logro. También permite evaluar el progreso de forma periódica y realizar ajustes necesarios en el camino. [10]
6. Metodología
   1. CRISP-DM:

Metodología de minería de datos cuyas siglas significan **C**ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining. Esta cuenta con 6 fases que se interrelacionan entre si en el ciclo de vida de minería de datos como se observa en la Figura 2.1

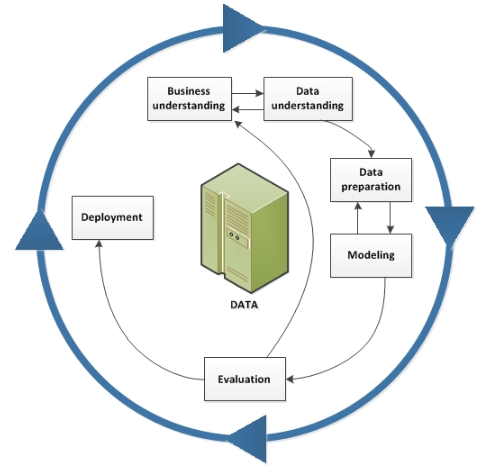


Figura . Ciclo de vida de crisp-dm [2]

Cada fase cuenta con subprocesos como se detalla a continuación:

Fase 1: Entendimiento del negocio.

Esta fase es el proceso de comprender el enfoque empresarial del problema de minería de datos para diseñar un plan de proyecto preliminar para alcanzar los objetivos principales definidos en el mismo. Esta fase tiene los siguientes procesos que son:

**Determinación de objetivos de negocio**: Establecer los objetivos comerciales para la minería de datos a través de la recopilación de información, documentación de los objetivos específicos y generales estableciendo los criterios de éxito. Los objetivos metas específicas y medibles que se deben alcanzar a través del uso de la minería de datos en un proyecto de negocios, y que deben ser acordados por las partes interesadas. [2]

**Evaluación de la situación:** ayuda a identificar los recursos disponibles, los riesgos y las oportunidades, lo que permite tomar decisiones informadas y desarrollar un plan sólido [2]

**Determinación de los objetivos de minería de datos:** se analiza las necesidades comerciales, identifica los problemas específicos que se van a resolver, identificar los datos necesarios a utilizar asignando prioridades a los objetivos en el proceso de minería. [2]

**Desarrollo de plan de proyecto:** El plan de proyecto es un documento detallado que describe el enfoque general para llevar a cabo el proyecto de minería de datos incluye las tareas específicas que se llevarán a cabo, los recursos necesarios, posibles riesgos asociados, los plazos, los objetivos y los criterios de éxito. [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. 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. [2]

**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. [2]

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. Los procesos más comunes de preparación de datos son:

**Selección de datos:** Se realiza la selección de elementos o filas incluir en el análisis, yque 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 incluyen 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 sean 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 numero 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

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. 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 experimentar 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 cuales serás 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

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. Se debe documentar los hallazgos y las conclusiones extraídas de los modelos y del proceso de minería de datos y contestar preguntas cómo:

* ¿Existen datos incompletos o sesgos en los resultados?
* ¿Hay algún modelo que no cumpla con los criterios de éxito empresarial pero que aún así pueda proporcionar información valiosa?
* ¿Se necesitan cambios en los procesos empresariales para aprovechar los hallazgos?
* ¿Se requieren nuevos recursos o habilidades para utilizar los modelos seleccionados?

Así como también se debe hacer una revisión de los procesos y fases realizadas determinando:

* ¿Qué fase contribuyó al valor del resultado final?
* ¿Cómo puede simplificar o mejorar esta actividad o actividad en particular?
* ¿Cuáles son los bugs o errores en esta etapa?
* ¿Qué se debe evitar la próxima vez?
* ¿Hay callejones sin salida, como modelos específicos que no han tenido éxito?
* ¿Hay alguna manera de predecir tales callejones sin salida para que las instrucciones funcionen de manera más eficiente?
* ¿Hay sorpresas (buenas y malas) en este punto?
* ¿Existe una manera clara de predecir la ocurrencia de tales eventos?
* ¿Hay decisiones o estrategias alternativas que se pueden utilizar en un momento dado?

Estas consideraciones tendrán que ser abordadas y documentadas para futuros procesos de minería de datos cómo también para determinar la siguiente fase: Si es posible pasar a la fase de despliegue o se tiene que retornar a una fase anterior.

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 ejecución de tareas de síntesis, como la elaboración del informe final y la revisión del proyecto:

**Plan de despliegue**: Este debe establecer varios aspectos para el despliegue:

* A sistemas se implementará los modelos
* Quienes deberán estar capacitados
* 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
* Qué seguimientos de hallazgos se deberá hacer

**Producción de informe final:** Documento para comunicar resultados a distintas personas de la organización. Este informe deberá cumplir:

* Descripción detallada del problema original
* El proceso que se llevo a cabo para la minería de datos
* Costos si es que aplica
* Consideraciones respecto a cambios de la planificación inicial
* Resumen de resultados
* Descripción general del plan de despliegue
* Recomendación para trabajos futuros de minería de datos

1. Desarrollo e implementación
   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, organismos evaluadores de la conformidad o capacitadores independientes. El planteamiento de los objetivos a desarrollar entorno a análisis de la demanda de cursos y perfiles utilizando KPIs estáticos y dinámicos siguiendo el plan de proyecto planteado

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 en distintas áreas y especialidades profesionales a nivel nacional como se muestra en la Fig. 1

Imagen que contiene Texto

Descripción generada automáticamente

Fig. menú de los módulos del portal web de SETEC

Este portal cuenta con la siguiente estructura de módulos y submódulos:

**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. Este módulo se compone de los siguientes submódulos:

* CNC para Operadores de Capacitación - CNCOC
* CNC para Organismos Evaluadores de la Conformidad - CNCOEC
* CNC Perfiles Inhabilitados - CNCPI
* Reportes
* Descargas

**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. Los submódulos que conforman este módulo son:

* Operadores de Capacitación – OC
* OC - Suspendidos – OCS
* OC - Finalizaron su vigencia de Calificación – OCF
* OC - Cancelados – OCC
* Reportes
* Descargas

**Personas Capacitadas por OC – PCOC**

Es el módulo de personas quienes han sido capacitadas en cursos y/o perfiles por Operadores de Capacitación. Los submódulos que lo conforman son:

* Búsqueda Personas Capacitadas – PCOC
* Reportes
* Descargas

**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. Los submódulos que lo conforman son:

* Organismos Evaluadores de la Conformidad – OEC
* OEC - Suspendidos – OECS
* OEC - Finalizaron su vigencia de Reconocimiento – OECF
* OEC - Cancelados – OECC
* Reportes
* Descargas

**Personas Certificadas por OEC – PCOEC**

Módulo que cuenta con la información de las Personas Capacitadas por Organismos Evaluadores de la Conformidad. Los submódulos que lo conforman son:

* Búsqueda Personas Certificadas – PCOEC
* Reportes
* Descargas

**Capacitadores Independientes – CI**

Este módulo cuenta con la información de capacitores independientes quienes imparten cursos de capacitación a profesionales a nivel nacional. Los submódulos que lo conforma son:

* Búsqueda de Capacitadores Independientes – CI
* Reportes
* Descargas

**Personas Capacitadas por CI – PCCI**

Son las Personas Capacitadas por Capacitadores Independientes quienes se han certificado en los cursos que han tomado. Los submódulos que lo conforman son:

* Búsqueda de Personas Capacitadas – PCCI
* Reportes
* Descargas

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%. [11] Esta tasa de empleo está conformada por profesionales quienes cuentan con certificaciones de cualificación de SETEC (Secretaria de Cualificaciones), 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). [12]

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 prerrequisitos establecidos para cada perfil de otros sectores, así como en la normativa legal vigente para el servicio de certificación de personas por competencias laborales. [12]

Ministerio del Trabajo a través de la Subsecretaría de Cualificaciones Profesionales entregó las resoluciones de Calificación como Operadores de Capacitación (OCC) y el Reconocimiento como Organismo de la Conformidad (OEC) a más de 60 institutos y empresas que cumplieron con los requisitos establecidos para este efecto. [13]

Estos nuevos OEC y OCC 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. [13] De acuerdo con el PLAN NACIONAL DE EDUCACIÓN Y FORMACIÓN TÉCNICA Y PROFESIONAL emitido por el ministerio de educación y ministerio del trabajo. [14]

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).

En base al análisis desarrollado en la unidad 2.1.1 los KPIs a desarrollar son de los módulos de OC, OEC y CI a las 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 que los ha tomado.

Los KPIs se desarrollaron en torno a las dos variables presentes en cada módulo:

**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 opto 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 asigno todos los perfiles con modalidad desconocida.

Se desarrolló un benchmark 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 bechmark 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.

**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:

**Por Curso/Perfil**:

* Número de personas capacitadas o certificadas
* Total de horas de capacitación o certificación
* Total de ganancias
* Porcentaje de asistencia

**Por Razón Social:**

* Total de cursos
* Volumen de personas capacitadas o certificadas
* Volumen de horas de capacitación o certificación
* Volumen de ganancias
* Volumen Porcentaje de asistencia

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:

|  |  |  |
| --- | --- | --- |
| **Semana referencial /**  **Etapas** | **Tareas específicas** | **Resultado esperado (si aplica)** |
| **1** | **Investigación de la literatura relacionada**  **Metodología**  **Herramientas**  **Conceptos teóricos y técnicos** | **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** | **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** | **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** | * **Conjuntos de datos preparados almacenados** * **Datawarehouse sin KPIs dinámicos** |
| **5** | **Modelamiento**  **Selección de técnicas de modelamiento**  **Generación de diseño de pruebas**  **Construcción de modelos**  **Evaluación de modelos** | * **Modelos evaluados y aprobados** * **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** | * **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** | * **Modelos desplegados** * **Redacción de despliegue de modelos** |
| **8** | **Visualización** | * **Gráficos de datos y resultados** |
| **9** | **Evaluación de usabilidad** | * **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 |

En la siguiente Figura 3.2 se muestra la arquitectura y fases del plan de proyecto.

Diagram, schematic

Description automatically generated

Figura .. Arquitectura y ciclo de desarrollo del sistema de minería de datos de SETEC

* 1. Entendimiento de datos:

El objetivo específico de entendimiento de datos comprende el análisis y entendimiento de los datos en los módulos del portal web SETEC.

En esta fase se desarrolla 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 0.3

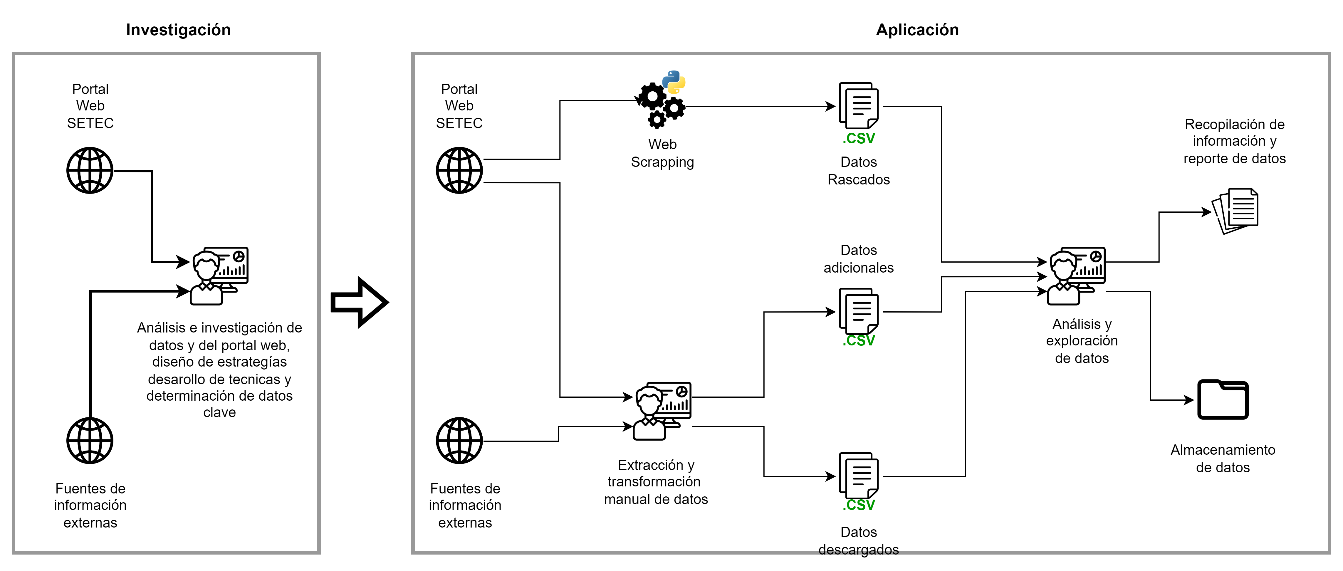


Figura .. Fase 2: Entendimiento de datos

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

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

#### Datos existentes

Cómo se analizó en la sección 3.1.1 Análisis y entendimiento de los módulos del portal web SETEC, 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 va a variar. 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 0.4. 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 0.5. Los submódulos de descarga promocionan datos listos para descargase en formato .xlsx como se observa en la Figura 0.6Timeline

Description automatically generated

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

A screenshot of a computer

Description automatically generated

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

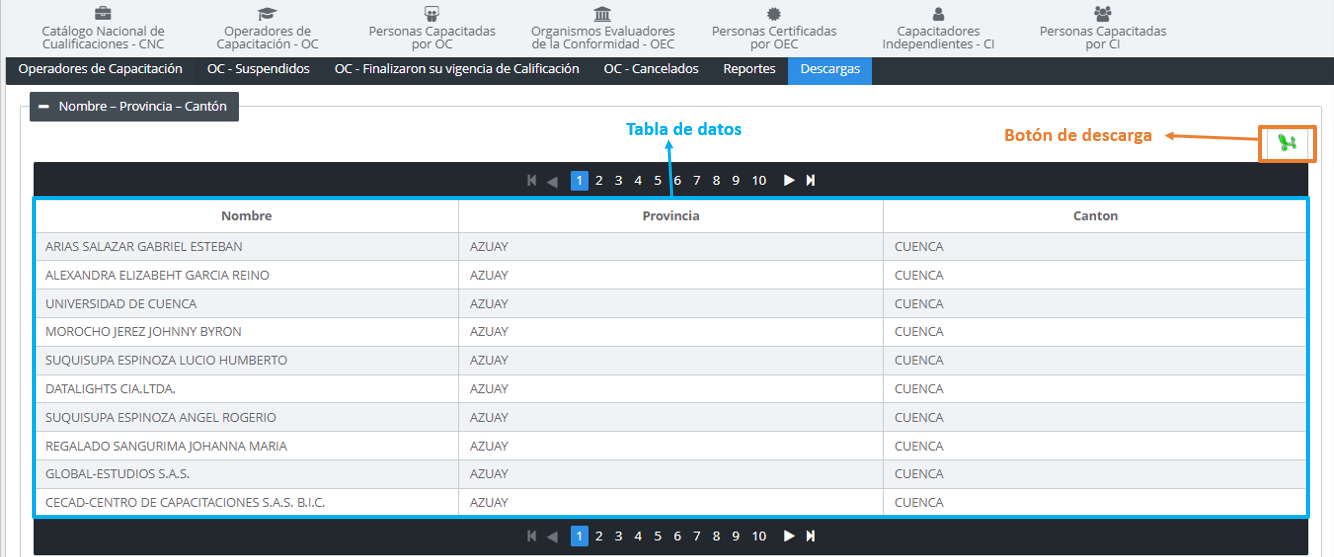


Figura .. Análisis de submódulo de descarga de OC

#### 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 0.7

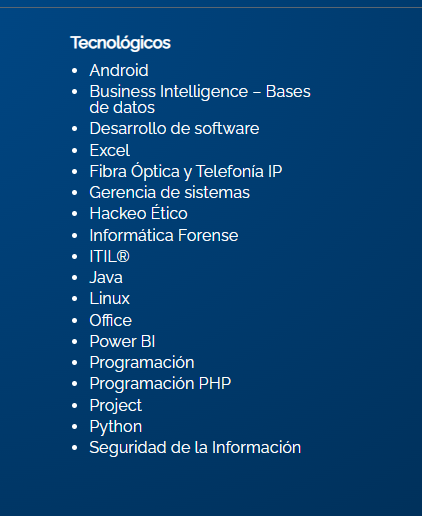


Figura .. Lista de cursos tecnológicos ofertados por el CEC en 2022B [15]

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 ,Areas de Formación,Nombre de los Programas, Modalidad, horarios, Tiempo de duración, Duración programa, Precio, formas de pago, certificación, plataforma educacion virtual y Red de educación cómo se muestra en la Figura 0.8

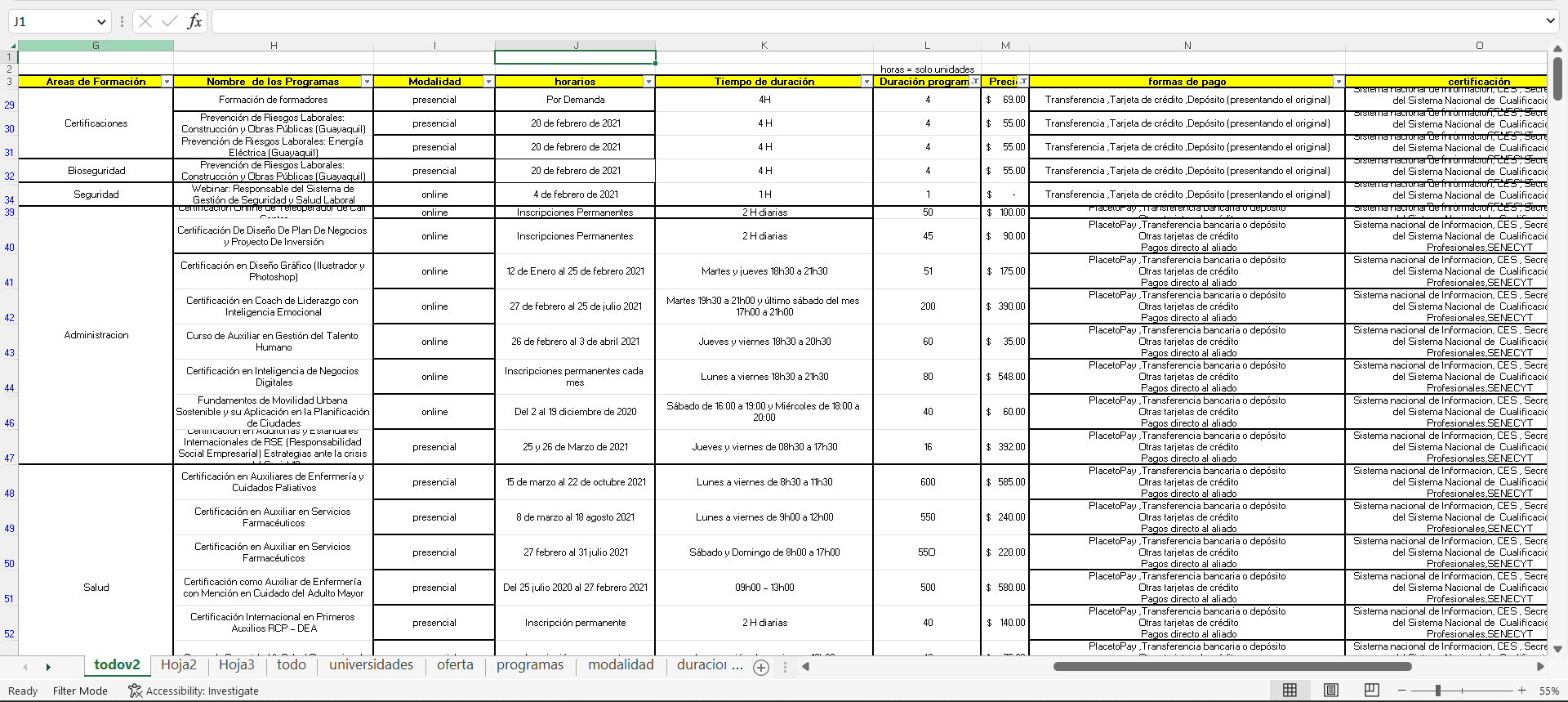


Figura .. Archivo xlsx con la investigación de mercado de empresas externas realizada 2021

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.

Graphical user interface, application

Description automatically generated

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

Extracción y colección de datos

#### Proceso de web scrapping

De acuerdo con los análisis realizados en la sección 3.2.1 Análisis y entendimiento de los datos en los módulos del portal web SETEC, los análisis y consideraciones realizadas y planteadas en el ANEXO 2 se comenzó con la extracción de datos.

El proceso de extracción se la desarrollo 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.

Tabla .. 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 | Carpteda 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 tuvo 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 suscritores de las páginas 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 fueron los siguientes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 fueron:

|  |  |  |
| --- | --- | --- |
| 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 invesgación autoría propia.

|  |  |  |
| --- | --- | --- |
| 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 continúa 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/DocumentosAlfrescoPortlet/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 Instagram hasta la fecha del 8 de enero de 2023 | Páginas de Facebook e Instagram oficiales de las razones sociales |
| 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 |

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.

Nota: 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

Tabla .. 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 | oecs.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 |

La descripción específica de cada archivo se la puede observar en el ANEXO 2

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 .. Resumen de archivos descargados

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 2

#### Descripción de datos adicionales

En cuanto a los datos adicionales los resultados se resumen en:

Tabla .. 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\_epn.xlsx | conversión a .CSV | costos\_de\_cursos\_cec\_epn.csv |
| Anexos\_y\_Tablas\_para\_entrega\_Catastros\_GADS.xlsx | Selección de columnas relevantes y conversión a .CSV | Ubicacion.csv |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |
| Ubicacion.csv | 224 | 3 | 672 | 672 |
| pcoc\_num\_conv.csv | 11 | 5 | 55 | 55 |
| pcoec\_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.**

* 1. Preparación de datos:

Esta fase responde al objetivo específico de Preparación de datos: Extracción, selección, limpieza, transformación, construcción e integración de datos del portal web.

Esta fase se compuso de 4 scripts en Jupyter notebook para el proceso de extracción, selección, limpiezas, 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 0.3:

Graphical user interface

Description automatically generated

Figura .. Diagrama de Fase 3: Prepación de datos

1. ***ETL\_DPP\_Downloaded\_data.ipynb:*** 
   1. 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\_Preprocesing.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 como en DATAWAREHOUSE.

***Nota***: Este script se ejecuta hasta la creación de las dimensiones .CSV: dim\_curso\_perfil, dim\_fecha y dim\_razon\_social

1. ***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
2. ***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 y ubicado en la carpeta de DATAWAREHOUSE. 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.
3. ***Data\_Preprocesing.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

Diagram

Description automatically generated

Figura .. Diseño de Datawarehouse

El Datawarehouse se creó con una arquitectura de tipo constelación 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:

Tabla .. 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 servir de base para crear datos ficticios que simulen datos futuros del año 2023 para después ser evaluados por los mismos modelos.

Notas:

* 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 numero 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.

* 1. 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 y entrenar 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

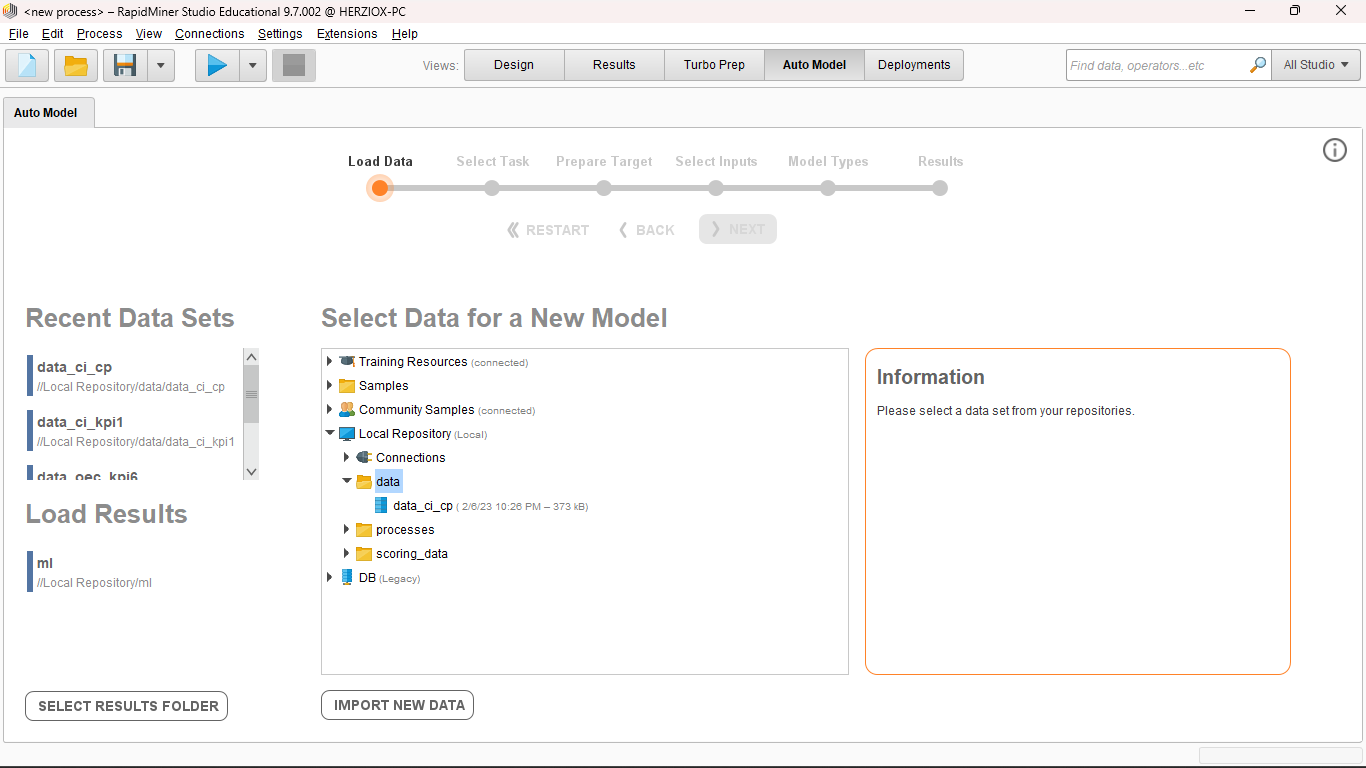
Timeline

Description automatically generated

Figura .. Fase 4 y 5 del proceso de CRISP-DM

Los procesos de estas fases se detallan a continuación:

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 selección la función de **Auto Model*.*** Se cargan los datos de entrenamiento como se observa en la FIGURA.

****

Después se debe seleccionar el atributo que se buscar predecir. Ver FIGURA

**Graphical user interface

Description automatically generated**

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

**Graphical user interface, text, application, email

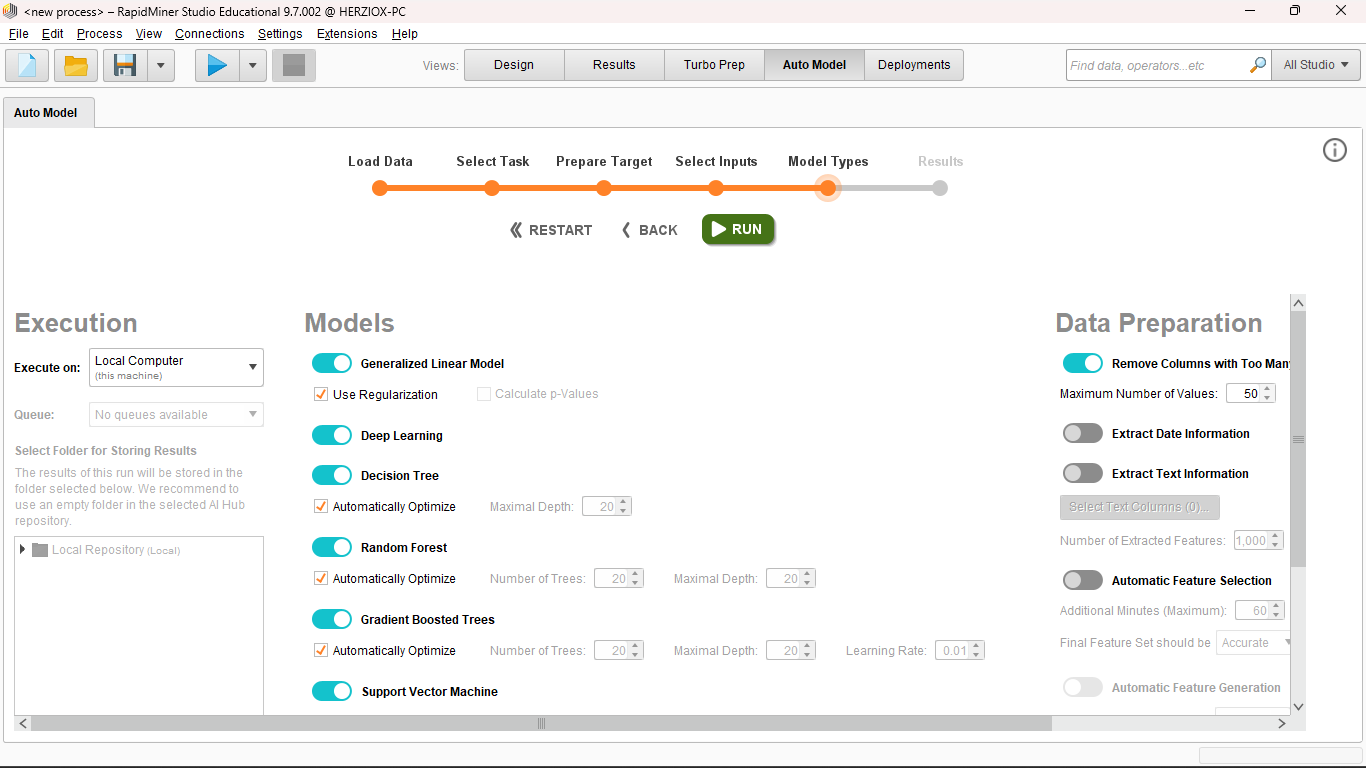
Description automatically generated**

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.

**Graphical user interface, text

Description automatically generated with medium confidence**

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

****

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. Cómo se observa en la FIGURA 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.

Graphical user interface

Description automatically generated with low confidence

1. **Evaluación de modelos**: La evaluación de los modelos se baso en las puntuaciones arrojadas por RapidMiner así cómo los resúmenes de los pesos de los atributo y la predicción del modelo.

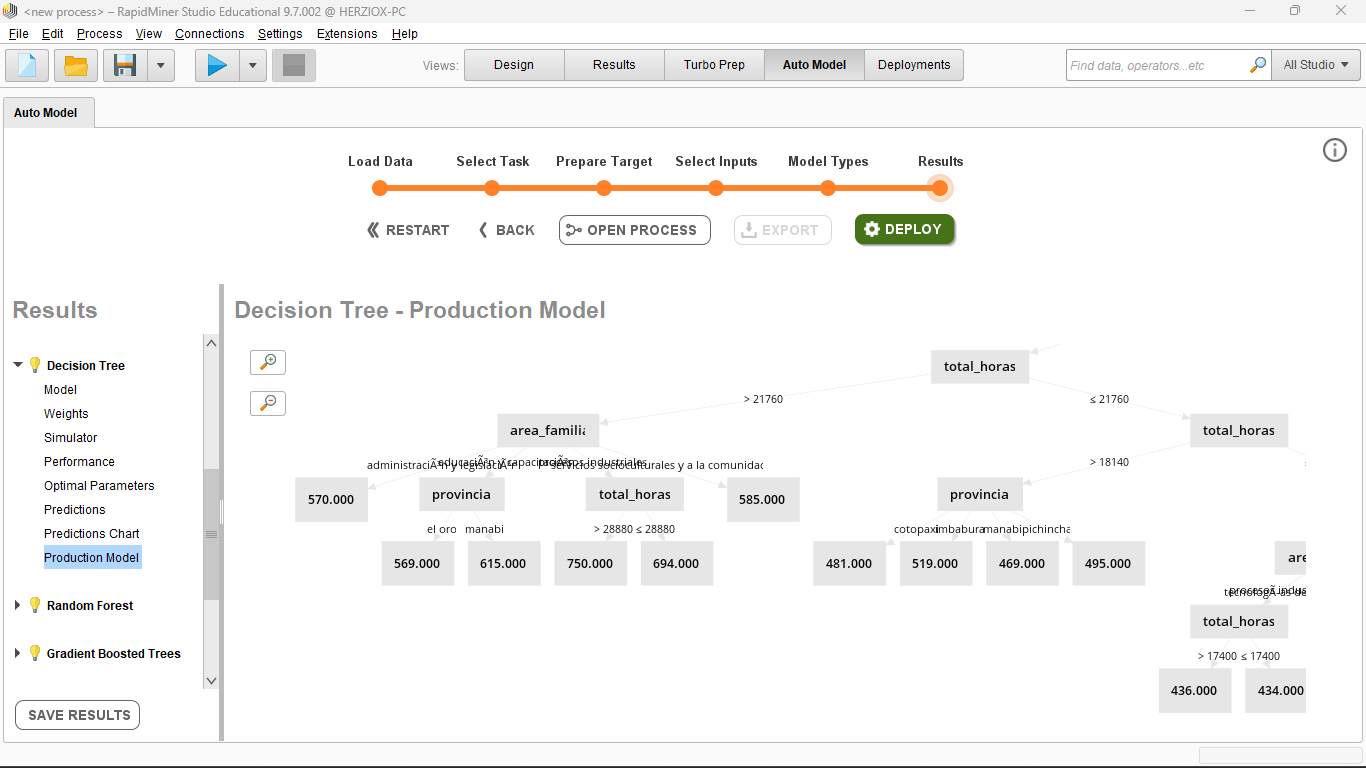
**Graphical user interface, text

Description automatically generated with medium confidence**

**Graphical user interface, application, table

Description automatically generated**

Tomando en cuenta estos aspectos se procede a desplegar el modelo de producción seleccionado el cual será almacenado en RapidMiner y MySQL.



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 encuentran en el ANEXO 4.

Al momento de guardar un modelo se debe asignar su nombre, la locación en la que se debe guardar, en que despliegue se lo hará y de que tipo es (regresión o clasificación). Los modelos se guardarán en el despliegue correspondiente para futuros procesos de puntuación de datos.

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, email

Description automatically generated

1. **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.
2. **Evaluación de datos simulados**: Los datos simulados de 2023 serán cargados a RapidMiner para después ser evaluados por los modelos guardados como se ve enla FIGURA. En este proceso se ve cuales atributos serán usados y cual será el atributo clave a predecir. Ver FIGURA. Al final se tendrá todos los datos evaluados por el modelo. Ver FIGURA

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.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

* 1. 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 MySQL. Ver FIGURA

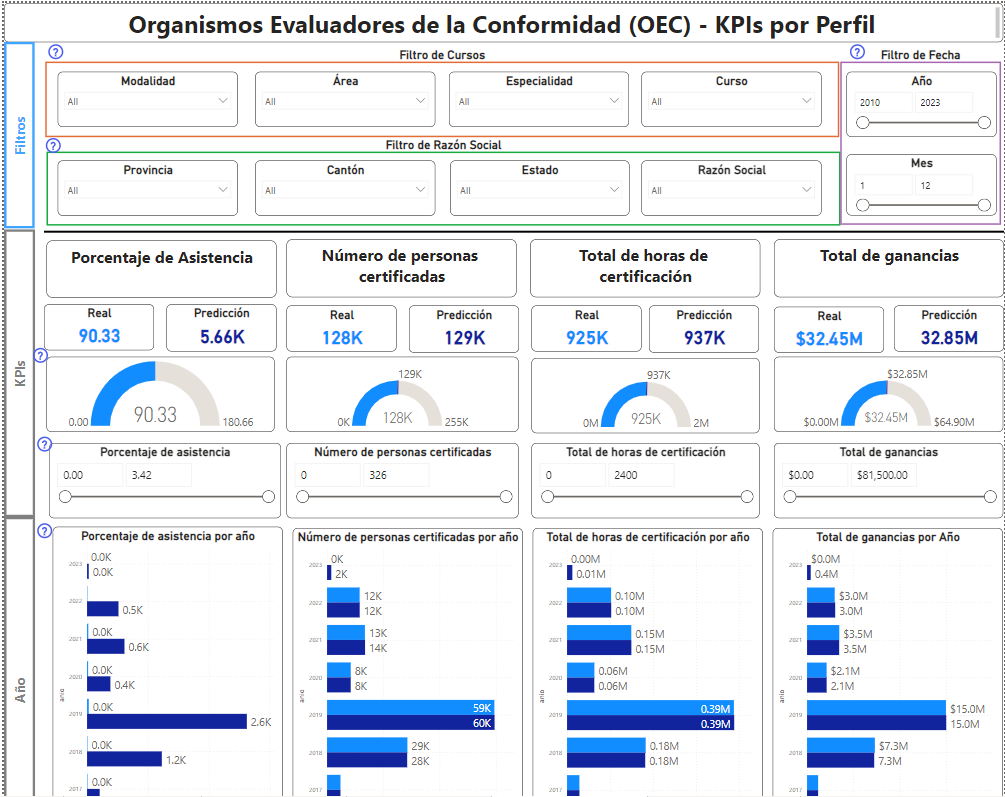
Text

Description automatically generated with medium confidence

El modelo físico del Datawarehouse final se observa en la FIGURA la cual sigue el mismo diseño del Datawarehouse inicial realizado en la fase 3, ver FIGURA

* 1. Visualización:

Siguiendo principios para la creación de dashboards del ANEXO 5 se diseñó la visualización de las tabla de hechos de cada tipo de razón social con su respectivo enfoque destacando cada KPI correspondiente tanto estático cómo dinámico estableciendo un contraste entre estos a través de distintas gráficas de cada componente relevante de las dimensiones del Datawarehouse. 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



El uso y explicación de esta visualización se lo encuentra en el ANEXO 6

* 1. 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, con conocimientos básicos y/o avanzados en tecnologías de la información y comunicación, y educación.

Resultados de Nielsen

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

|  |  |  |
| --- | --- | --- |
| 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 deprisa o volver a hacerlo por completo |
| N/A | No aplica | En caso de que no se considere que exista la heurística |

El resumen del promedio de los resultados de cada heurística se presenta en la siguiente FIGURA.

Chart, bar chart

Description automatically generated

El numero de severidades puntuadas por cada heurística se muestra a continuación en la siguiente FIGURA

Chart, box and whisker chart

Description automatically generated

Resultados de SUS

La escala de evaluación de usabilidad de un sistema consta de 10 preguntas evaluadas de 1 punto (Totalmente en desacuerdo) hasta 5 puntos (Totalmente de acuerdo).

Para obtener el puntaje total, los resultados son calculados con la siguiente formula:

* X = Suma de los puntos de todas las preguntas impares – 5
* Y = 25 – Suma de los puntos de todas las preguntas pares
* Puntaje SUS = (X + Y) x 2.5

La puntuación total es de 100 y cada una de las preguntas tiene un peso de 10 puntos.

El puntaje total fue de **74.5** que representa:

Un rango de aceptabilidad: **Alto**

Grado: **C**

Equivalente a: **Bueno**

Siguiendo las escalas y puntuaciones de la siguiente FIGURA.

A picture containing timeline

Description automatically generated

<https://mpiua.invid.udl.cat/system-usability-scale-sus-en-el-s-xxi/>

1. ANALISIS Y DISCUSIÓN DE LOS RESULTADOS

* 1. Resultados

Exponer los resultados obtenidos, utilizando para esto el apoyo de tablas, figuras, entre otros.

Un ejemplo de una tabla se presenta a continuación:

**Tabla 1.** Resultados de las pruebas realizadas

|  |  |  |
| --- | --- | --- |
| **No. Prueba** | **Resultado** | **Tiempo [s]** |
| 1 | 10 | 0.9 |
| 2 | 5 | 0.5 |

Un ejemplo de una figura se presenta a continuación:

**Figura 1.** Resultados de las pruebas realizadas

Un ejemplo de una ecuación se presenta a continuación:

**Ecuación 1.** Trinomio Cuadrado Perfecto

1. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS
   1. Conclusiones

Presenta lo novedoso del trabajo de integración curricular, así como evaluación del cumplimiento o no de lo propuesto en los objetivos. En el caso en que no se cumpla uno o varios objetivos, y no se logren los resultados esperados, se propone una posible respuesta que explique por qué sucedió esto o las falencias de la planteado.

* 1. Recomendaciones

Indicar las recomendaciones formuladas a partir del desarrollo de este trabajo de integración curricular.

* 1. Trabajos futuros

REFERENCIAS BIBLIOGRÁFICAS

|  |  |
| --- | --- |
| [1] | J. C. Díaz, Introducción al Business Intelligence, Barcelona: El Ciervo 96, S.A, 2010. |
| [2] | 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]. |
| [3] | 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]. |
| [4] | RapidMiner, «Why RapidMiner?,» 23 02 2022. [En línea]. Available: https://rapidminer.com/why-rapidminer/. |
| [5] | 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. |
| [6] | Selenium, «WebDriver,» 20 02 2023. [En línea]. Available: https://www.selenium.dev/documentation/webdriver/. |
| [7] | Jupyter, «About Us Project Jupyter’s origins and governance,» 20 02 2023. [En línea]. Available: https://jupyter.org/about. |
| [8] | MySQL, «MySQL,» MySQL Workbench, 2022. [En línea]. Available: https://www.mysql.com/products/workbench/. [Último acceso: 26 12 2022]. |
| [9] | 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]. |
| [10] | 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. |
| [11] | INEC, «Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU), enero 2022,» INEC, Quito, 2022. |
| [12] | 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]. |
| [13] | 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]. |
| [14] | Educación, Ministerio de; Trabajo, Ministerio de, Plan Nacional de Educación y Formación Técnica y Profesional, Quito: Ministerio de Educación, 2021. |
| [15] | CEC EPN, «Centro de Educación Continua EPN,» CEC EPN, 2022. [En línea]. Available: https://www.cec-epn.edu.ec/. [Último acceso: 15 11 2022]. |
| [16] | L. Carvajal, Metodología de la Investgación Científica. Curso general y aplicado, 28 ed., Santiago de Cali: U.S.C., 2006, p. 139. |

ANEXOS

En caso necesario, el documento escrito deberá incluir los anexos y secciones que incorporan información que sea relevante, pero que, por su extensión, no pueden ser incorporadas directamente en ninguna de las secciones anteriores. Normalmente, en la sección de Anexos se incluyen conjuntos de datos extensos, formatos de encuestas, entrevistas, enlaces hacia videos o programas que sean producto o formen parte del Trabajo de Integración Curricular, entre otros.

Ejemplo de Anexos se muestran a continuación:

ANEXO I. Conjunto de Datos Extensos

ANEXO II. Formato de Entrevista

ANEXO III. Enlaces

La numeración de los Anexos debe realizarse con números en formato romano.

ANEXO I

Incluir el contenido del Anexo I.

ASPECTOS DE FORMATO

Tipografía

Un resumen de la tipografía se presenta a continuación:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Texto** | **Tamaño** | **Fuente** | **Estilo** | **Ejemplo** |
| Normal | 11 | Arial | Normal; Alineado justificado. | Texto de párrafo |
| Nivel 1 | 14 | Arial | Negrita; Enumerada; Mayúscula, Con sangría; Alineado izquierda. | **1 INTRODUCCIÓN** |
| Nivel 2 | 14 | Arial | Negrita; Enumerada; Con sangría; Alineado izquierda. | **1.2 Objetivo general** |
| Nivel 3 | 12 | Arial | Negrita; Alineado izquierda. | **Subsección** |

Espaciado

Todo el documento debe tener espaciado de 1,5. Las tablas pueden usar espaciado simple.

Numeración

Un resumen de la numeración se presenta a continuación:

|  |  |  |
| --- | --- | --- |
| **Elemento** | **Estilo** | **Número** |
| Tabla | Superior al elemento, Alineado centro. | Continuo, Número arábigo. |
| Figura | Inferior al elemento, Alineado centro | Continuo, Número arábigo. |
| Ecuación | Inferior al elemento, Alineado centro. | Continuo, Número arábigo. |
| Páginas | Inferior  Centro | Continuo, Número romano hasta Abstract y en arábigo hasta el final. |
| Anexos |  | Continuo, Número romano. |

Anexos