# PROPUESTA DE UNA METODOLOGÍA BASADA EN MLOPS PARA EL SOPORTE DE LA GESTIÓN EN PROYECTOS DE CIENCIA DE DATOS EN LA EMPRESA WIZIT MIND BLOWING SOLUTIONS S.A.S.

**ANGELA ALEXANDRA ORDOÑEZ BOLAÑOS**

**PRÁCTICA PROFESIONAL**

**UNIVERSIDAD DEL CAUCA**

**FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES**

**PROGRAMA DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES**

**POPAYÁN**

**2022**

# PROPUESTA DE UNA METODOLOGÍA BASADA EN MLOPS PARA EL SOPORTE DE LA GESTIÓN EN PROYECTOS DE CIENCIA DE DATOS EN LA EMPRESA WIZIT MIND BLOWING SOLUTIONS S.A.S.


**Trabajo de Grado Práctica Profesional para optar por el título de:**

**INGENIERA EN ELECTRÓNICA Y TELECOMUNICACIONES**

**Estudiante: ANGELA ALEXANDRA ORDOÑEZ BOLAÑOS**

**Director: PhD. GUSTAVO ADOLFO RAMÍREZ GONZÁLEZ**
**Asesor: PhD. JUAN SEBASTIAN ROJAS MELENDEZ**

**UNIVERSIDAD DEL CAUCA**

**FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES**

**PROGRAMA DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES**

**POPAYÁN**

**2022**

# Dedicatorias

A mi madre Shirley J. Bolaños Molina, quien ha sido la mujer que ha construido mi fuerza. La persona que siempre ha celebrado cada uno de mis triunfos y nunca me ha dejado sola en mis caídas.
Gracias a mi madre por construir la mujer que represento hoy en día.

A mi padre Carlos Alberto Ordoñez R. por ser la definición del concepto más puro que conozco del amor. Por ser la fuente de alegría más grande en mi vida y por lograr lo imposible para ser el mejor padre del mundo.
Gracias a mi padre por existir.

A Carlos Mario Fernandez Medina por ser mi apoyo incondicional en los días más difíciles, por encontrar la manera de siempre hacerme sentir especial.
Gracias a Carlos por convertirse en el hermano que siempre quise. Y gracias a la vida por permitirme coincidir con una persona tan maravillosa.

A Dios por darme la valentía para estar lejos de casa. Por poner ángeles en mi camino a quienes llamo amigos, y por enseñarme a controlar mi mente y mis pensamientos cuando no logro salir de ellos.

Y por último, pero no menos importante, a mi por mi hard-work.

# Reconocimientos

Especial reconocimiento al PhD. Juan Sebastian Rojas por ser un excelente líder, por ser la única persona constante que me brindó su apoyo y conocimiento de principio a fin de este trabajo y sobre todo por su calidad humana y su paciencia. Porque además de ser un excelente profesional es un excelente ser humano.

# Resumen

Muchas empresas de ingeniería han decidido incursionar en el área de ciencia de datos, y de MLOps con el fin de crear, extraer y analizar grandes cantidades de datos. Este enfoque es importante debido a su carácter multidisciplinario, ya que combina principios, conceptos y prácticas del campo de la ingeniería, el aprendizaje automático, las matemáticas, entre otras. El objetivo principal de explorar este campo de trabajo es la rapidez, el alto rendimiento, la eficiencia y la eficacia que se puede obtener cuando se aplican correctamente los conceptos, las metodologías, los procedimientos y lineamientos que esta área de trabajo ofrece. Sin embargo, es importante mencionar que, aunque la definición y el concepto son claros, la información es muy poca con respecto a las metodologías y procedimientos a llevar a cabo cuando se trabaja en un proyecto de este tipo, esto es debido a que MLOps es un término muy reciente. Por ello, este documento monográfico provee un modelo de referencia basado en MLOps que ayuda al soporte de gestión de proyectos de ciencias de datos. Es importante mencionar que este proyecto está basado en una empresa Colombiana llamada WIZIT MIND BLOWING SOLUTIONS S.A.S. Sin embargo en la investigación realizada se han tomado artículos, documentos e información de diferentes países.

**Palabras clave**: Aprendizaje automático, MLOps, operaciones de aprendizaje automático, gestión de proyectos, metodología, ciencia de datos, modelo de referencia, conjunto de lineamientos.

# Abstract

Many engineering companies have decided to venture into the area of data science, and MLOps in order to create, extract and analyze large amounts of data. This approach is important due to its multidisciplinary nature, since it combines principles, concepts and practices from the field of engineering, machine learning, mathematics, among others. The main objective of exploring this field of work is the high performance, efficiency and effectiveness that can be obtained when the concepts, methodologies, procedures and guidelines that this area of work offers are correctly applied. However, it is essential to mention that although the definition and concept are clear, the information is very scarce regarding the methodologies and procedures that must be carried out in a project of this type. This is because MLOps is a very new term. Therefore, this monographic document provides a methodology based on MLOps that helps support data science project management. It is important to mention that this project is based on a Colombian company called WIZIT MIND BLOWING SOLUTIONS S.A.S. However, in the research carried out, papers, documents and information from different countries have been taken.

**Keywords:** Machine Learning, MLOps, machine learning operations, project management, methodology, data science, reference model, set of guidelines.

# Contenido

# Lista de acrónimos

| | |
|---|---|
| **ML** | Machine Learning: Aprendizaje automático |
| **MLOPS** | Machine Learning Operations: Operaciones de aprendizaje automático |
| **DDS** | Data Driven Scrum |
| **KDD** | Knowledge Discovery in Database |
| **EDA** | Exploratory Data Analysis |
| **DEVOPS** | Development Operations |
| **CI** | Continuous Integration: Integración continua |
| **CD** | Continuous Delivery: Entrega continua |
| **CT** | Continuous Training: Entrenamiento Continuo |
| **DL** | Deep Learning |

# Capítulo 1

# Introducción

## 1.1. Planteamiento del problema

Actualmente, gran parte de las actividades diarias de todo tipo de empresas Colombianas se instituyen en proyectos. Los proyectos son llevados a cabo para satisfacer las necesidades de los clientes. Por tanto, ejecutar un proyecto implica un alto nivel organizacional, por lo que existen técnicas orientadas a aumentar la probabilidad de éxito del proyecto y a alinear diferentes grupos de la organización, que han sido creados exclusivamente para apoyar el proceso de fabricación del producto o servicio.

Dicho esto, es importante mencionar que todo proyecto requiere un diseño previo para poder ser desarrollado y ejecutado. Para ello, es indispensable comprender a qué clase de proyecto la organización se está enfrentando y cuáles son los lineamientos correspondientes a seguir para la terminación efectiva y exitosa del proyecto en cuestión.

Cuando se trata de un proyecto de desarrollo software, existen diversas metodologías que son ampliamente aplicadas (Scrum, kanban, Dynamic Systems Development Method, etc) [1]. Sin embargo, en un área creciente como la ciencia de datos donde los proyectos suelen tener tanto enfoques diferentes, como tiempos de desarrollo diferentes, estas metodologías no son suficientes para cubrir los requerimientos y la complejidad que poseen los proyectos de esta área. Esta complejidad se genera debido a que la ciencia de datos es un área de trabajo multidisciplinaria, porque incluye variedad de procesos para obtener los datos y así prepararlos. Cuando estos datos son preparados y entendidos entonces se convierten en datos de utilidad [2].

Ahora bien, aunque existen metodologías como KDD (Knowledge Discovery in Database), SEMMA [3], CRISP-DM, y DDS: Data-Driven Scrum [4], que son ampliamente utilizadas para gestionar proyectos de ciencia de datos, no existe una metodología de referencia completa e integrada que tenga un proceso bien definido para la gestión y finalización exitosa de proyectos basados en ciencia de datos. Por esa razón, la mayoría de las empresas en Colombia desarrollan proyectos de

ciencia de datos de manera empírica, porque no cuentan con estrategias para implementar y producir soluciones correctamente.

Por tanto, el equipo de analítica de datos de la empresa define de manera experimental todas las etapas a seguir, sin lineamiento alguno, y lo hacen usando el método heurístico “prueba y error”. Entonces, al probar una alternativa, se corre el riesgo de que no funcione y, por ende, se debe intentar una alternativa diferente, haciendo de cada proyecto un proceso arduo, tardío y tedioso.

Para afrontar estas brechas, varios estudios han planteado y resuelto preguntas de investigación que ofrecen algunas guías para el desarrollo de esta práctica empresarial. Estos estudios proponen interesantes recomendaciones para combinar las prácticas de DevOPS y ciencia de datos para así mejorar el desarrollo de proyectos de ciencia de datos. También hay estudios con importantes recomendaciones sobre ML y estudios de cómo darle valor a las compañías basándose en el tipo de organización empresarial que es y los objetivos principales que posee.

Sin embargo, en estas contribuciones realizadas por la investigación mencionada anteriormente, ninguna investigación ha utilizado MLOps (Operaciones de aprendizaje automático) la cual ayuda al equipo de producción y a los científicos de datos a comunicarse y colaborar para que puedan administrar el ciclo de vida de producción de aprendizaje automático, eliminar el desperdicio y hacer el sistema más escalable al proporcionar automatización para gestionar de manera eficiente los proyectos de ciencia de datos [5]. Por lo tanto, es necesario investigar el uso de MLOps en proyectos de ciencia de datos para estructurar un modelo de referencia que apoye la regulación, organización y efectividad de las empresas.

Con base en lo anterior, el aporte de este documento monográfico es un modelo de referencia respaldado por MLOps para el apoyo en la gestión de proyectos de ciencia de datos. Este modelo es validado en la empresa Colombiana WIZIT MIND BLOWING SOLUTIONS la cual es una compañía que lleva 8 años dedicándose a actividades de desarrollo de sistemas informáticos (planificación, análisis, diseño, programación y pruebas)  y se encuentra ubicada en la localidad de POPAYAN, en el departamento de CAUCA.

## 1.2. Objetivos

### 1.2.1. Objetivo general

Diseñar una metodología para apoyar el proceso de gestión de proyectos de ciencia de datos al interior de la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S.

### 1.2.2. Objetivos específicos

- Identificar las prácticas actuales de la empresa respecto al desarrollo de proyectos de ciencia de datos
- Estructurar una metodología para la gestión de proyectos de ciencia de datos basado en la aproximación MLOps de Amazon al interior de la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S
- Validar la metodología propuesta a través de su aplicación en la ejecución de un proyecto piloto dentro de la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S.

## 1.3. Contenido

El presente documento consiste en 5 capítulos desarrollados de la siguiente manera:

- **Capítulo 1: Introducción:** Establece la relevancia del tema de investigación, esclarece el problema y menciona brevemente la solución de este.

- **Capítulo 2: Estado del arte:** Presenta trabajos previos de investigadores, ingenieros, y especialistas en el tema, que han propuesto metodologías similares. Además de ello, expone la investigación realizada por medio de los diferentes conceptos y prácticas que sustentan este trabajo e información que se debe tener en cuenta al momento de realizar un proyecto en el campo de machine learning, MLOps y data science.

- **Capítulo 3: Prácticas internas de la empresa:** Se identifican las prácticas internas dentro de la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S. referentes al área de analítica de la empresa para el desarrollo de proyectos de ciencia de datos.

- **Capítulo 4: Implementación del modelo de referencia propuesto:** Presenta el modelo de referencia basado en MLOPS para la gestión de proyectos de ciencia de datos, con ello se describe detalladamente cada paso del proceso realizado. Dicho modelo de referencia desarrollado sigue un

orden estructurado que permite construir un proyecto de ciencia de datos registrando cada detalle desde su inicio hasta su finalización.

- **Capítulo 5: Validaciones:** Presenta los resultados de la validación que se le realizó al modelo de referencia. Esta validación se realizó por medio de diferentes herramientas que condujeron a la verificación del buen funcionamiento del prototipo del modelo de referencia desarrollado.

- **Capítulo 6: Conclusiones:** Presenta la síntesis breve de los aportes realizados en esta práctica profesional, además de los trabajos futuros y sus consideraciones.

# Capítulo 2

# Estado del Arte

## 2.1. Conceptos y definiciones

**Machine Learning**

Machine learning es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante. Las técnicas basadas en ML (Machine learning) se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por computadora, la ingeniería de naves espaciales, las finanzas, el entretenimiento y la biología computacional hasta aplicaciones biomédicas y médicas [6]

**Algoritmo de Machine Learning**

Para definir el algoritmo de ML (Machine Learning) es importante mencionar que ML posee 3 modalidades las cuales son: aprendizaje supervisado, no supervisado y por refuerzo.
El aprendizaje supervisado consiste en deducir una función con base en datos de entrenamiento, en donde los datos de entrenamiento son un conjunto de ejemplos de entrenamiento. No obstante, el conjunto de datos de entrada se divide en dos, el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Mientras que el conjunto de datos de salida puede ser un valor numérico (Regresión) o una etiqueta de clase (Clasificación). Dicho esto, se puede afirmar que los algoritmos de aprendizaje supervisados necesitan ayuda externa. Por otra parte, en el aprendizaje no supervisado, no hay un conjunto de datos con referencia a resultados conocidos o etiquetados, ya que no hay respuestas correctas y no hay maestro. Entonces, cuando se introducen nuevos datos, el algoritmo aprende algunas características de los datos y utiliza dichas características aprendidas para descubrir y presentar la estructura de los datos. Por tanto, los métodos del aprendizaje no supervisado no pueden ser aplicados a un problema de regresión o clasificación, debido a que se desconocen los posibles valores correctos de salida. Finalmente, el aprendizaje reforzado es un área de ML que se encarga de determinar qué acciones los agentes de software deben llevar a cabo para maximizar alguna noción de recompensa acumulativa [7].

**Machine Learning operations (MLOps)**

Machine Learning Operations es la combinación de ML, DevOps e ingeniería de datos, que ayudan a implementar el sistema de ML de manera confiable y eficiente. Entonces se define MLOps como los componentes de operaciones y ML que ayudan al equipo de producción y a los científicos de datos a comunicarse y colaborar para que puedan administrar el ciclo de vida del ML de producción, eliminan el desperdicio y hacen que el sistema sea más escalable al proporcionar automatización y ayudar a obtener información valiosa y coherente de los modelos de ML [5].

**Gestión de proyectos (Project management)**

La gestión de proyectos consiste en aplicar habilidades, herramientas, conocimientos, y técnicas a diferentes actividades de un proyecto con el fin de cumplir con los requerimientos del mismo. La gestión de proyectos debe ser organizada y planificada de acuerdo a los recursos humanos y capital financiero de la organización que esté llevando a cabo el proyecto [8].

**Ciencia de datos (Data science)**

La ciencia de datos es el campo de estudio que combina las habilidades de programación y el conocimiento de áreas como las matemáticas y estadística con el fin de extraer información significativa de los datos [9].

La ciencia de datos proporciona el lenguaje y las técnicas necesarias para comprender y tratar los datos. La ciencia de datos implica el diseño, recolección, análisis e interpretación de datos numéricos, con el objetivo de extraer patrones y otra información útil. Los pasos típicos de un proyecto de ciencia de datos son [10]:

- Recopilación de datos para obtener información sobre una pregunta de investigación.
- Limpieza, resumen y visualización de los datos.
- Modelado y análisis de los datos.
- Traducir decisiones sobre el modelo en decisiones y predicciones sobre la pregunta de investigación.

**Definición de roles importantes en un equipo de trabajo para realizar proyectos basados en ciencias de datos:**

**Científico de datos [9]**

Un científico de datos es una persona que podrá llevar a cabo proyectos de ciencia de datos de principio a fin. Un científico de datos puede ayudar a almacenar grandes cantidades de datos, crear procesos de modelado predictivo y realizar análisis profundos.

Las habilidades principales de un científico de datos son las matemáticas, informática, estadística y probabilidad, análisis, programación, comunicación (Habilidades blandas).

Algunas cualidades de los científicos de datos son:

- Inquisitivo
- Profesional en ingeniería, informática, o campos afines.
- Orientado al resultado: sabe cómo crear productos de datos y visualizaciones para lograr que los datos se puedan entender.
- Conocimiento del dominio: entiende el negocio y cómo analizarlo.
- Capaz de encontrar respuestas a incógnitas conocidas.

**Ingeniero de datos [11]**

Es una persona versátil que utiliza la informática para ayudar a procesar grandes conjuntos de datos. Por lo general, se centran en la codificación, la limpieza de conjuntos de datos y la implementación de solicitudes que provienen de los científicos de datos.

Los ingenieros de datos son los encargados de encontrar tendencias en conjuntos de datos y desarrollar algoritmos para ayudar a que los datos sin procesar sean más útiles para la empresa.

Las habilidades principales de un ingeniero de datos son la programación y manejo de big data.

Las cualidades de un ingeniero de datos se pueden describir como las siguientes:

- **Trabajan en la arquitectura de datos:** Utilizan un enfoque sistemático para planificar, crear y mantener arquitecturas de datos.

- **Recopilan datos:** Antes de iniciar cualquier trabajo en la base de datos, deben obtener y almacenar datos de las fuentes adecuadas y confiables.

- **Realizan investigaciones:** Los ingenieros de datos llevan a cabo investigaciones en la industria para gestionar cualquier problema que pueda surgir al abordar un problema comercial.

- **Se adaptan a cambios:** Los ingenieros de datos no se basan únicamente en conceptos teóricos de bases de datos. Deben tener el conocimiento y la destreza para trabajar en cualquier entorno de desarrollo, independientemente de su lenguaje de programación.

- **Son capaces de crear modelos e identificar patrones:** Los ingenieros de datos utilizan un modelo de datos descriptivos para la agregación de datos para extraer información histórica. También crean modelos predictivos en los que aplican técnicas de pronóstico para aprender sobre el futuro con información procesable.

- **Automatizan tareas:** Los ingenieros de datos se sumergen en los datos y encuentran tareas en las que la participación manual puede eliminarse con la automatización.

**Analista de datos [12]**

Un analista de datos normalmente es alguien que tiene el conocimiento y las habilidades para convertir datos sin procesar en información y conocimientos, que son de uso importante para tomar decisiones comerciales.

Los analistas de datos son responsables de analizar datos utilizando técnicas estadísticas. Otra tarea importante del analista de datos consiste en encargarse de implementar y mantener bases de datos, recopilar datos de fuentes primarias y secundarias, identificar, analizar e interpretar tendencias de los datos.

Las habilidades de un analista de datos son la estadística, la comunicación, el conocimiento empresarial, la limpieza y preparación de datos, el saber analizar y explorar datos, y la simplificación de la información.

Un analista de datos posee cualidades y realiza funciones como las siguientes:

- Usar herramientas automatizadas para extraer datos de fuentes primarias y secundarias.
- Eliminar datos corruptos y corrección de errores de codificación.
- Desarrollar y mantener bases de datos.
- Realizar análisis para evaluar la calidad y el significado de los datos.
- Filtrar datos mediante la revisión de informes e indicadores de rendimiento para identificar y corregir problemas de códigos.

- Usar herramientas estadísticas para identificar, analizar e interpretar patrones y tendencias en conjuntos de datos complejos que podrían ser útiles para el diagnóstico y la predicción.
- Asignar valor numérico a las funciones comerciales esenciales para que el desempeño comercial pueda evaluarse y compararse durante períodos de tiempo.
- Analizar las tendencias locales, nacionales y globales que impactan tanto a la organización como a la industria.
- Preparar informes para la gerencia que indiquen tendencias, patrones y predicciones utilizando datos relevantes.
- Trabajar con programadores, ingenieros y jefes de administración para identificar oportunidades con el fin de mejorar procesos.
- Proponer modificaciones del sistema y diseñar estrategias de gobierno de datos.
- Preparar informes de análisis finales para que las partes interesadas comprendan los pasos del análisis de datos, lo que les permite tomar decisiones importantes basadas en diversos hechos y tendencias.

**Desarrollador [13]**

Los desarrolladores de bases de datos, también se pueden conocer como diseñadores de bases de datos o programadores de bases de datos. Los desarrolladores son los responsables del diseño, la programación, la construcción y la implementación de nuevas bases de datos, así como de la modificación de las bases de datos que ya existen según las actualizaciones de la plataforma y los cambios en las necesidades de los usuarios y/o clientes.

**DevOps [14]**

Es la integración de dos mundos: el mundo del desarrollo y el de las operaciones. En DevOps se promueve el desarrollo automatizado, la implementación y el monitoreo de la infraestructura. Es un cambio organizacional en el que, en lugar de grupos distribuidos en silos que realizan funciones por separado, los equipos multifuncionales trabajan en entregas continuas de funciones operativas. Este enfoque ayuda a entregar valor de manera más rápida y continua, reduciendo los problemas debidos a la falta de comunicación entre los miembros del equipo y acelerando la resolución de problemas. DevOps significa un cambio de cultura hacia la colaboración entre el desarrollo, la garantía de calidad y las operaciones.

**KPIs**

Los KPI son medidas que pueden ser financieras y no financieras. Las organizaciones generalmente usan estas medidas para descubrir qué tan exitosas fueron en el logro de metas a largo plazo. Los KPIs ayudan a constituir un sistema

efectivo de medición del desempeño y esto con el fin de tener definidos y estandarizados todos los procesos dentro de la organización [15].

Cuando se tiene un enfoque basado en procesos la organización puede dirigir su negocio hacia su deber principal, que es crear valores que satisfagan a sus clientes. Las ventajas de las organizaciones basadas en procesos son [15]:

- Poner las reclamaciones de los clientes en primer lugar.
- Constituir una gestión interfuncional.
- Las actividades se realizan de forma lógica.
- Los empleados cumplen con sus respectivas funciones.
- Se constituye la relación interna proveedores - cliente.

**AWS Well-Architected Framework**

AWS Well-Architected Framework es una arquitectura que se encarga de ayudar al equipo de ciencia de datos a comprender las ventajas y desventajas de las decisiones que se toman al momento de crear sistemas en AWS. Mediante el uso del marco, se logra conocer la variedad de prácticas recomendadas de arquitectura para diseñar y operar cargas de trabajo en la nube de AWS, con el fin de que sean seguras, confiables, eficaces, y rentables. Además de ello, este marco proporciona una forma de medir las arquitecturas que se posean de forma constante en función de las prácticas recomendadas e identificar áreas que se puedan improvisar o mejorar [16].

El AWS Well-Architected Framework se basa en seis pilares que serán explicados a continuación:

**Pilares del AWS Well-Architected Framework** [17]

| Nombre | Descripción |
|---|---|
| **Excelencia operativa** | La capacidad para admitir el desarrollo y ejecutar cargas de trabajo de manera eficaz, obtener información acerca de las operaciones y mejorar continuamente admitiendo procesos y procedimientos para ofrecer valor de negocio. |

| | |
|---|---|
| **Seguridad** | El pilar de la seguridad abarca la capacidad para proteger los datos, sistemas, activos y aprovechar las tecnologías de la nube a fin de mejorar la seguridad. |
| **Fiabilidad** | La capacidad de una carga de trabajo de realizar su función prevista de forma correcta y constante cuando se espera. Esto incluye la capacidad de operar y probar la carga de trabajo a través de su ciclo de vida completo. Este documento ofrece orientación exhaustiva sobre las prácticas recomendadas para implementar cargas de trabajo fiables en AWS. |
| **Eficiencia de rendimiento** | La habilidad de utilizar recursos informáticos de manera eficiente para cumplir con los requisitos del sistema y mantener esa eficiencia a medida que la demanda cambia y la tecnología evoluciona. |
| **Optimización de costos** | Se trata de la capacidad para ejecutar sistemas a fin de entregar valor de negocio al menor precio |
| **Sostenibilidad** | La sostenibilidad como disciplina aborda el impacto medioambiental, económico y social a largo plazo de sus actividades empresariales. La Comisión Mundial sobre el Medio Ambiente y el Desarrollo de las Naciones Unidas define el desarrollo sostenible como «aquel que permite satisfacer las necesidades del presente sin comprometer la habilidad de las futuras generaciones de satisfacer sus necesidades propias». Su organización o negocio puede tener repercusiones negativas en el medioambiente, como emisiones de carbono directas o indirectas, residuos no reciclables y daños a recursos compartidos, como el agua no contaminada. |

Tabla 1. Pilares de AWS Well- Architected Framework [17]

**CI/CD [18]**

CI/CD es un método que se usa con el objetivo de entregar aplicaciones de forma rápida y sencilla a los clientes mediante la introducción de la automatización en las etapas de desarrollo de aplicaciones. Los principales conceptos atribuidos a CI/CD son integración continua, entrega continua e implementación continua.

Específicamente, el proceso de CI/CD incorpora la automatización y monitoreo continuo en todo el ciclo de vida de las aplicaciones desde las fases de integración y prueba hasta la entrega y el despliegue.

**Knowledge Discovery in Database [19]**

Los pasos del proceso de KDD inician con la selección donde de un dataset principal hay que seleccionar un subconjunto de variables que apoyan en la exploración del fenómeno que se está estudiando, comúnmente las bases de datos son "ruidosas" o contienen datos inexactos o faltantes, entonces durante la etapa de preprocesamiento, los datos se limpian. Esto implica la eliminación de "valores atípicos"; decidir estrategias para manejar los campos de datos faltantes; tener en cuenta la información de secuencia de tiempo y la normalización de datos aplicable.
En la fase de transformación se intenta limitar o reducir el número de elementos de datos que se evalúan manteniendo la validez de los datos. Durante esta etapa, los datos se organizan, se convierten de un tipo a otro (es decir, se cambia de nominal a numérico) y se definen atributos nuevos o "derivados". En este punto, los datos se someten a uno o varios métodos de minería de datos, como clasificación, regresión o agrupación. El paso final es la interpretación y documentación de los resultados de los pasos anteriores. Las acciones en esta etapa podrían consistir en volver a un paso anterior en el proceso KDD para refinar aún más el conocimiento adquirido, o traducir el conocimiento en una forma comprensible para el usuario.

**SEMMA [20]**

SEMMA se define como una herramienta que ayuda a los usuarios en los procesos de selección, exploración y modelación de cantidades significativas de datos almacenados, para así poder responder a preguntas o predecir eventos que pueden pasar. Se puede identificar a la metodología SEMMA como un conjunto de herramientas funcionales, enfocándose más en los aspectos del desarrollo del modelo de minería de datos

- **Sample: Extracción de una muestra representativa.**

  En esta primera fase de la metodología, se realiza la extracción de un conjunto de datos que sean una buena representación de la población a

analizar, esto se hace con el objetivo de facilitar los procesos de minado sobre los datos, reduciendo los tiempos que se necesita para determinar la información valiosa para el negocio.

- **Explore: Exploración de los datos en la muestra.**

  En esta fase, se hace un recorrido a través de los datos extraídos en la muestra para detectar, identificar y eliminar datos anómalos, ayudando a refinar los procesos de descubrimiento de información en fases siguientes del proceso.

- **Modify: Modificación de los datos.**

  Esta modificación de los datos se puede realizar creando, seleccionando y transformando las variables en las cuales se va a enfocar el proceso de selección del modelo.

- **Model: Modelación de los datos.**

  En esta fase, las herramientas de software se encargan de realizar una búsqueda completa de combinaciones de datos que predicen de una manera confiable los resultados buscados.

- **Assess: Evaluación de los datos obtenidos.**

  Después de que la fase de modelación presente los resultados obtenidos de la aplicación de los métodos de minería de datos al conjunto de datos, se deberá realizar un análisis de los resultados para ver si estos fueron exitosos de acuerdo a las entradas que se tuvieron para analizar el problema.

**CRISP-DM**

Está metodología proporciona un vistazo general del ciclo de vida de un proyecto de minería de datos. Cada fase se divide en un conjunto de tareas, las cuales se relacionan entre sí, esas tareas a su vez se componen de actividades específicas y de un conjunto de resultados concretos **[20]**.

**Fases de la metodología CRISP-DM [21]**

- **Comprensión del negocio**

Esta primera fase es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.

- **Comprensión de los datos**

  Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis.

- **Preparación de los datos**

  En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, éstas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos.

- **Modelado**

  En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios: Debe ser apropiada para el problema, disponer de los datos adecuados, cumplir los requisitos del problema, cumplir un tiempo adecuado para obtener un modelo y se debe tener un conocimiento previo de la técnica.

- **Evaluación**

  En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis.

- **Despliegue o implantación**

  En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, las tareas que componen esta fase son:

• Planear la implantación
• Planear la monitorización y mantenimiento: Si los modelos resultantes del proceso de minería de datos son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.
• Producir el informe final
• Revisión del proyecto

**DDS: Data Driven Scrum [22]**

Data Driven Scrum es un marco de flujo continuo para la ciencia de datos ágil que integra la estructura de Scrum y el flujo continuo de Kanban.

Los conceptos que se utilizan en el marco de DDS son:

- **Backlog Item (BItem):** Es una lista priorizada de BItems (es decir, trabajo por hacer).
- **Ítem Backlog (IBB):** Es el lugar donde cada ítem (en el registro de ítems) se divide en tareas, con al menos una tarea de creación, una tarea de observación y una tarea de análisis para cada ítem.
- **Task Board:** Es una representación visual de los elementos de trabajo actualmente en curso. Para que se inicie el trabajo en un elemento, las tareas para ese elemento se mueven del IBB al tablero de tareas y se muestran en el tablero, generalmente en la columna 'to do'. El tablero de tareas tiene varias columnas (como mínimo, 'to do', 'in progress', 'done') y cada tarea fluye a través del tablero, mostrando así visualmente el trabajo que se está realizando dentro del equipo.
- **Iterations:** Una iteración es una colección de uno o más elementos de la lista de trabajos pendientes. Debido al desglose de tareas (en el IBB) para el elemento, cada iteración tendrá una o más tareas de observación y análisis. La información obtenida de la iteración debe tener un valor comercial derivado de lo que se crea o del análisis de la tarea completada.

## 2.2. Trabajos relacionados

Para la ejecución del estado del arte se utilizó el modelo de mapeo sistemático propuesto por Kai Petersen, para lo cual se exploraron 4 bases de datos para la extracción de información: Scopus, Web of Science, ScienceDirect, en las cuales se encontraron alrededor de 35 investigaciones y 1035 en la base de datos de Google Scholar. Sin embargo, la información más importante ha sido extraída de los siguientes 4 artículos:

**MLOps: una taxonomía y una metodología [23]**

En este documento los autores proponen una metodología para proyectos de MLOps (Machine Learning Operations) y una taxonomía para agrupar trabajos de investigación relacionados a MLOps. Este documento representa un avance con respecto a la investigación realizada en el año 2021 debido que para ese entonces, no existía ni una sola metodología propuesta con respecto a este tema. Es importante resaltar que a pesar que MLOps impulsa la implementación de procesos de aprendizaje automático y ciencia de datos más rápidos, más óptimos, confiables y productivos, es un área poco estudiada y discutida a nivel académico.

La metodología propuesta por los actores está estructurada de la siguiente manera (Figura 1):

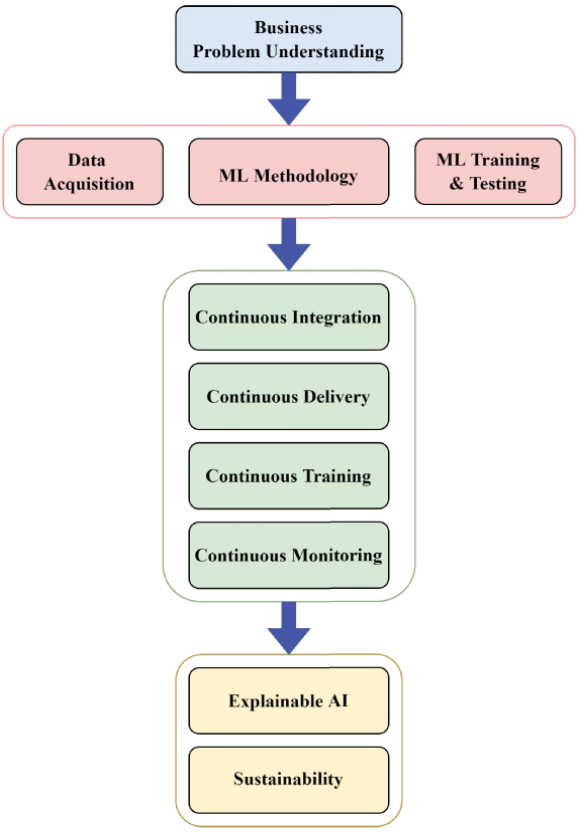


Figura 1. Metodología propuesta para proyectos de MLOps [23]

**A. Comprensión del problema empresarial**

Consiste en establecer un entendimiento del negocio y los KPIs para definir los criterios de éxito y de esta manera resolver el problema correctamente. Los autores mencionan que esta fase no es técnica, y por esa razón es importante la

comunicación entre los científicos de datos o los especialistas en analitica de datos y los expertos empresariales.

**B. Adquisición de datos**

Las tareas para la adquisición de datos consisten en:

- Extracción de datos: Donde se seleccionan e integran los datos relevantes para el modelo de ML.

- Análisis de datos: Se realiza un EDA (Análisis exploratorio de datos) con el fin de identificar las tendencias que se presentan en el conjunto de datos y de esta manera obtener insights importantes para así realizar una buena toma de decisiones.

- Preparación de datos: Esta preparación incluye la limpieza de datos y la división en entrenamiento, validación y conjunto de pruebas. En este paso también se incluye la transformación de datos y la ingeniería de características necesarias para el modelo.

Es importante resaltar que en este documento se mencionan dos metodologías principales que pueden ser usadas cuando no hay suficientes datos para entrenar el modelo:

- El Data Augmentation: Técnica que aumenta el número de datos disponibles mediante la inserción de copias de los datos.
- Transfer Learning: Es un técnica que permite reutilizar una red neuronal que ya ha sido previamente entrenada en un problema similar.

**C. Metodología de ML (Machine Learning)**

Cuando se ha realizado una correcta adquisición de datos, el paso a seguir es la selección de los mejores algoritmos de ML para resolver el problema. Generalmente, el equipo de ciencia de datos realiza una investigación exhaustiva para obtener un fructífero estado del arte con base al problema específico y prueba un enfoque de abajo hacia arriba para resolverlo. ML es empírico por naturaleza. En ML se prueban diferentes características, modelos, parámetros y configuraciones de hiper parámetros para lograr encontrar lo que funciona mejor.

**D. Entrenamiento y pruebas de ML**

Este proceso es iterativo, debido a que el equipo de ciencia de datos prueba varios algoritmos, funciones e hiper parámetros. Después de esto, una vez se han elegido

los mejores modelos de ML, se vuelven a entrenar y probar. Los modelos se evalúan utilizando diferentes métodos de validación como: Holdout validation, Cross-validation, Bootstrap validation.

**E. Integración continua**

Este es el primer paso para iniciar el camino de la entrega continua. CI permite a las empresas mejorar la calidad del software y la productividad de los equipos. Esta práctica incluye la creación y prueba automatizadas de software.

**F. Entrega continua**

El propósito de la entrega continua es garantizar que una aplicación esté lista para la producción después de pasar correctamente los controles de calidad y las pruebas automatizadas. Dicho esto, los autores mencionan que CD produce la reducción del riesgo de implementación, ya que emplea un conjunto de prácticas como CI y automatización de implementación para entregar automáticamente software en producción.

**G. Entrenamiento Continuo**

Durante el entrenamiento continuo es necesario almacenar una mayor cantidad de datos y configurarlos de la misma forma en que se entrena el modelo. Esto conlleva a la detección de valores atípicos para lograr comprender cuándo la distribución de datos discrepa de los datos de entrenamiento. CT se encarga de volver a entrenar y servir automáticamente a los modelos. El entrenamiento continuo es una parte importante en MLOps porque vuelve a entrenar de forma automática y continua los modelos antes de volver a implementarlos.

**H. Monitoreo Continuo**

En esta etapa es importante administrar los riesgos que poseen los modelos en producción. Esto con el objetivo de verificar el rendimiento y la precisión del modelo, debido a que en algunos casos puede que disminuya. Una vez que el modelo está en producción, requiere de validación o pruebas continuas porque hay patrones en los datos que podrían cambiar con el tiempo y esto genera que el modelo se vuelva menos preciso. Esto se puede dar porque los datos utilizados en el entrenamiento del modelo ya no son iguales a los nuevos datos existentes en producción.

**I. IA explicable**

Los autores mencionan que los métodos de aprendizaje profundo poseen una mayor complejidad. Debido a que en estos casos, cuando el modelo se convierte en una aplicación real en producción, se comienza a estudiar la "explicabilidad" de los

modelos con el objetivo de responder a preguntas comerciales. Los autores mencionan que la explicabilidad se define como “el grado en que un ser humano puede comprender la causa de una decisión” [23]. Es decir es más fácil identificar las relaciones de causa y efecto dentro de las entradas y salidas del sistema cuando un sistema de ML es explicable.

**J. Sostenibilidad: Huella de Carbono**

Los autores mencionan la huella de carbono como un punto importante para el cambio climático con respecto al uso habitual de modelos de DL (Deep Learning) en proyectos del mundo real, debido a que corresponde a un inmenso crecimiento en la computación y la energía requerida. Sin embargo, en este documento se menciona que es posible mitigar esta tendencia explorando cómo mejorar la eficiencia energética en los modelos Deep Learning. Por lo tanto, es esencial que los científicos de datos sepan su huella energética y de carbono, para que puedan tomar medidas activas para reducirlas siempre que sea posible.

La anterior representación de la metodología propuesta por los autores representa un esquema interesante para tener en cuenta al momento de realizar proyectos de ciencias de datos. La diferencia entre esta metodología y el modelo de referencia propuesta en este trabajo de grado consiste en la serie de etapas que se han estructurado. La metodología presentada por los autores consiste en un diagrama de flujo, desarrollado de manera vertical y que posee 4 bloques grandes, que a su vez se desencadenan en procesos que los autores recomiendan al momento de llevar a cabo un proyecto de ciencia de datos. El modelo de referencia realizado en este proyecto se organiza por medio de 5 etapas que desencadenan una serie de subetapas y algunas de ellas se regresan a una etapa anterior si es necesario. El flujo del modelo de referencia se puede observar en la Figura 2.

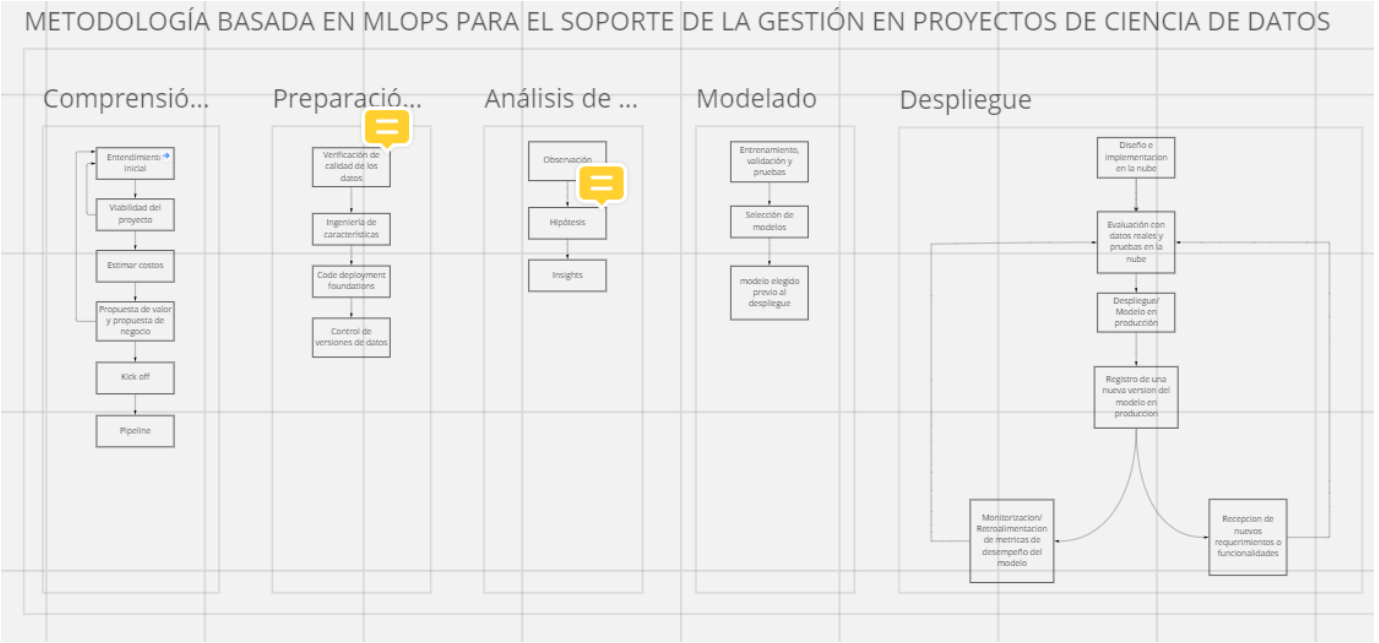


Figura 2. Modelo de referencia representado en la plataforma Miro [fuente propia]

Existen diferencias notables como:

- Kick off - Presentado en el modelo de referencia propuesto
- La huella de carbono - Presentado en la metodología presentada por los autores
- Costos - Presentado en el modelo de referencia propuesto
- Desarrollo de una pipeline - Presentado en el modelo de referencia propuesto
- Retroalimentación de métricas de desempeño del modelo - Presentado en el modelo de referencia propuesto
- IA explicable - Presentado en la metodología presentada por los autores

Con lo anterior dicho es importante tomar en cuenta el tipo de proyecto que se va a llevar a cabo para usar la guía más conveniente para el éxito del mismo.

**Machine Learning Operations (MLOps): Overview, Definition, and Architecture [24]**

Los autores de este paper aplicaron un enfoque de método mixto, el cual consiste en 3 pasos. En el primer paso se realiza una revisión de la literatura del tema a investigar, en este caso MLOps (Machine Learning Operations). Después se identifican las herramientas relevantes en el campo de MLOps esto con el fin de comprender de una mejor manera los componentes técnicos. Finalmente, se llevaron a cabo entrevistas a ocho expertos en el campo de MLOps en donde mencionan que descubrieron cuatro aspectos esenciales de MLOps que son: Los

principios, los componentes, los roles, y la arquitectura como se muestra en la figura 3.

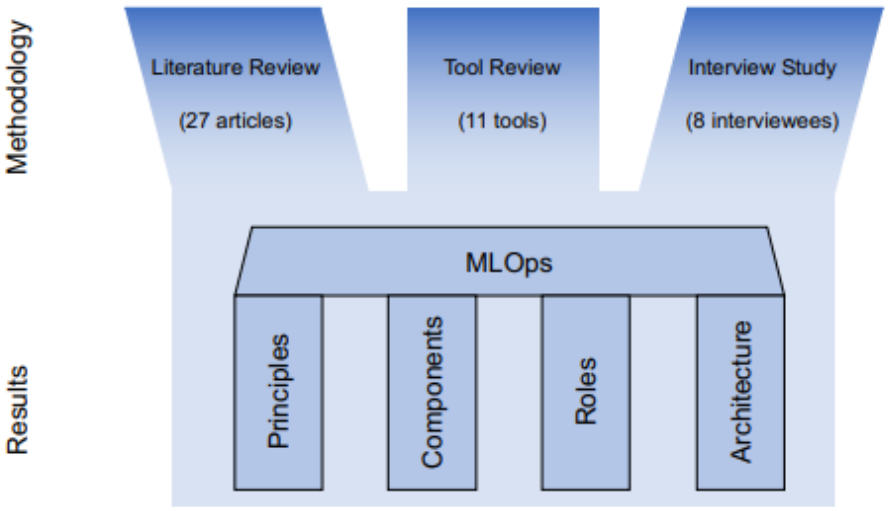


Figura 3. Resumen de la metodología [24]

Además de ello, la figura 3 muestra también el número de artículos seleccionados según criterios de inclusión y exclusión, las herramientas encontradas y la cantidad de entrevistas realizadas.

Los autores agregan una figura extra, en donde se presenta la implementación de principios de los componentes técnicos, es decir, la guía de cómo se deben realizar las cosas en MLOps. Los autores identifican nueve principios necesarios para realizar MLOps mostrados en la figura 4.

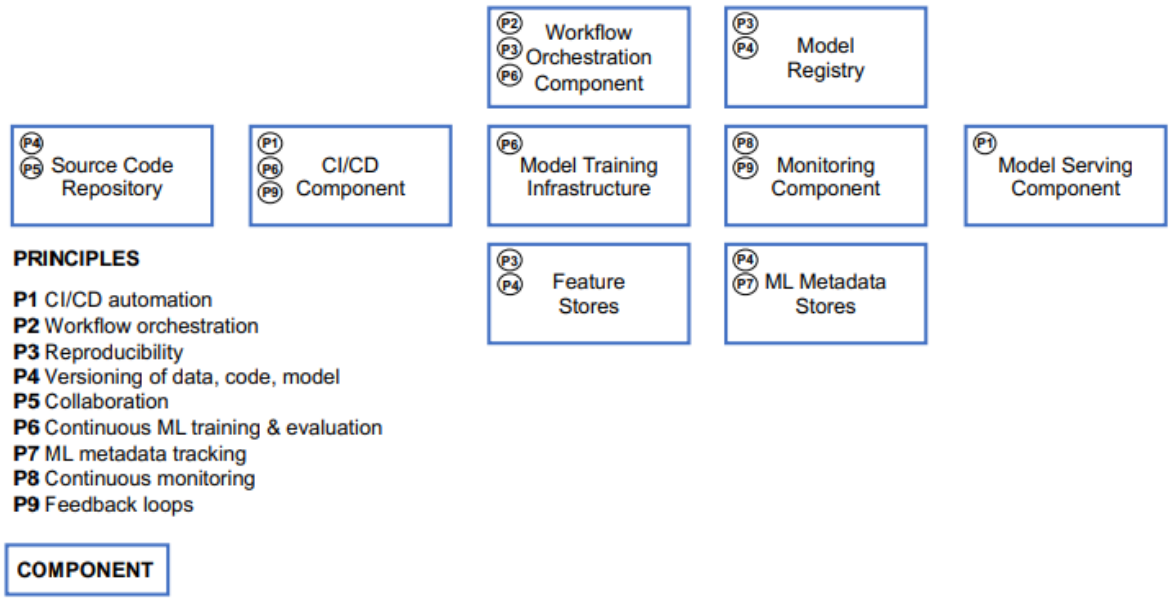

Figura 4 . Implementación de principios vinculados a los componentes técnicos con los que están asociados [24]

Finalmente, los autores proporcionan una ilustración muy completa mostrada en la figura 5. Esta ilustración corresponde al flujo de trabajo y arquitectura de MLOps de extremo a extremo con componentes y roles funcionales.

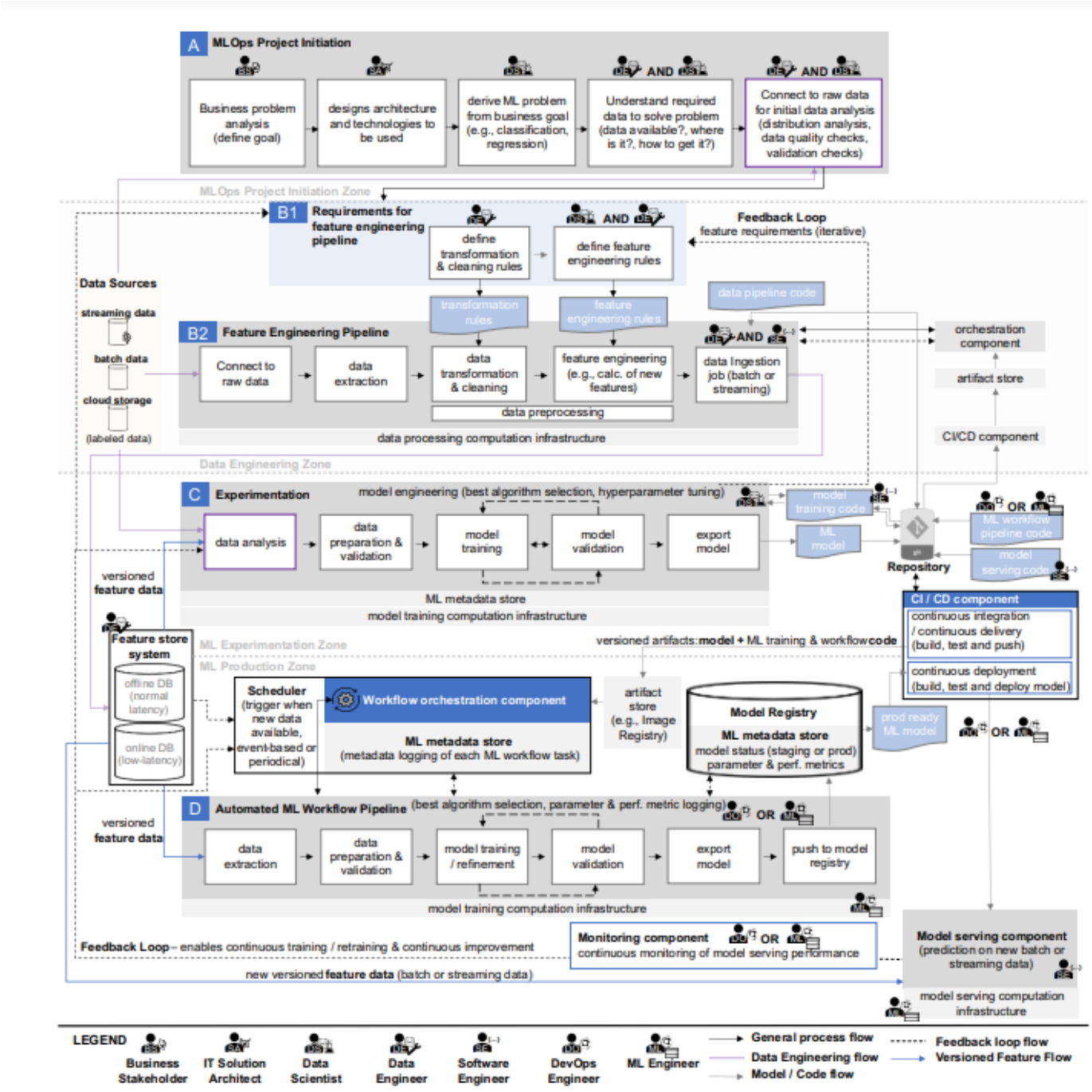

Figura 5. Flujo de trabajo y arquitectura de MLOps de extremo a extremo con componentes y roles funcionales [24]

Los autores explican que el diseño de esta arquitectura está representada secuencialmente. Esta secuencia representa el orden en la que se ejecutan las diferentes tareas en cada etapa. Dicha arquitectura está diseñada para no ser dependiente de la tecnología. Por lo tanto es posible elegir las tecnologías y los marcos que se adapten mejor a las necesidades del cliente.
En la figura 5 se muestra un proceso integral, en donde incluye (A) los pasos de iniciación del proyecto MLOps; (B) la pipeline de ingeniería de funciones, incluida la incorporación de datos al almacén de funciones; (C) la experimentación; y (D) la pipeline del flujo de trabajo de ML automatizado hasta el servicio del modelo.

Es esencial destacar que este paper presenta una interesante recopilación de una ardua investigación realizada por los autores. Este paper tiene una relación directa con la ejecución de este proyecto, y presenta variedad de puntos muy similares a la propuesta de este modelo de referencia debido a que aborda uno de los objetivos específicos planteados al iniciar este proyecto el cual consiste en estructurar una metodología para la gestión de proyectos de ciencia de datos basado en la aproximación MLOps. Con ello, se puede concluir que las pocas diferencias entre esta metodología presentada por los autores y el modelo de referencia realizado en este proyecto son:

- Costos - Presentado en el modelo de referencia propuesto
- Planteamiento de una hipótesis para el análisis exploratorio de datos - Presentado en el modelo de referencia propuesto

Con lo anterior mencionado es válido destacar que ambos modelos de referencia son muy similares y que se debe analizar las características del proyecto que se va a efectuar para utilizar la guía más conveniente con el fin de lograr el éxito del mismo.

**Designing an open-source cloud-native MLOps pipeline [25]**

En el documento se muestra el diseño de una pipeline de MLOps nativa de la nube de código abierto. Según se menciona por los autores, dicha pipeline podría ser usada en la mayoría de proyectos de aprendizaje automático. Es posible ejecutar dicha pipeline en Kubernetes locales, así como en entornos de Kubernetes de varios proveedores de la nube.

En este documento también se consideran preguntas de investigación relevantes para llevar a cabo el proyecto las cuales son:

- ¿Qué se requiere de una pipeline MLOps moderna?
- ¿Qué tan factible es diseñar e implementar una pipeline de MLOps con herramientas nativas de la nube de código abierto?
- ¿Qué tan independiente del proveedor de la nube puede ser la solución?

Con estas preguntas planteadas, los autores realizaron un pipeline de MLOps que corresponde a una solución de las mismas. Para ello, se determinaron requisitos de la pipeline de MLOps. Estos requisitos son la limpieza de datos, etiquetado de datos, ingeniería de características, entrenamiento de modelos, evaluación de modelos, implementación de modelos y monitoreo de modelos. Estos requisitos son ilustrados en la figura 6. Además de ello, la pipeline propuesta e implementada por los autores es capaz de manejar todos estos pasos con herramientas de código abierto y nativas de la nube.

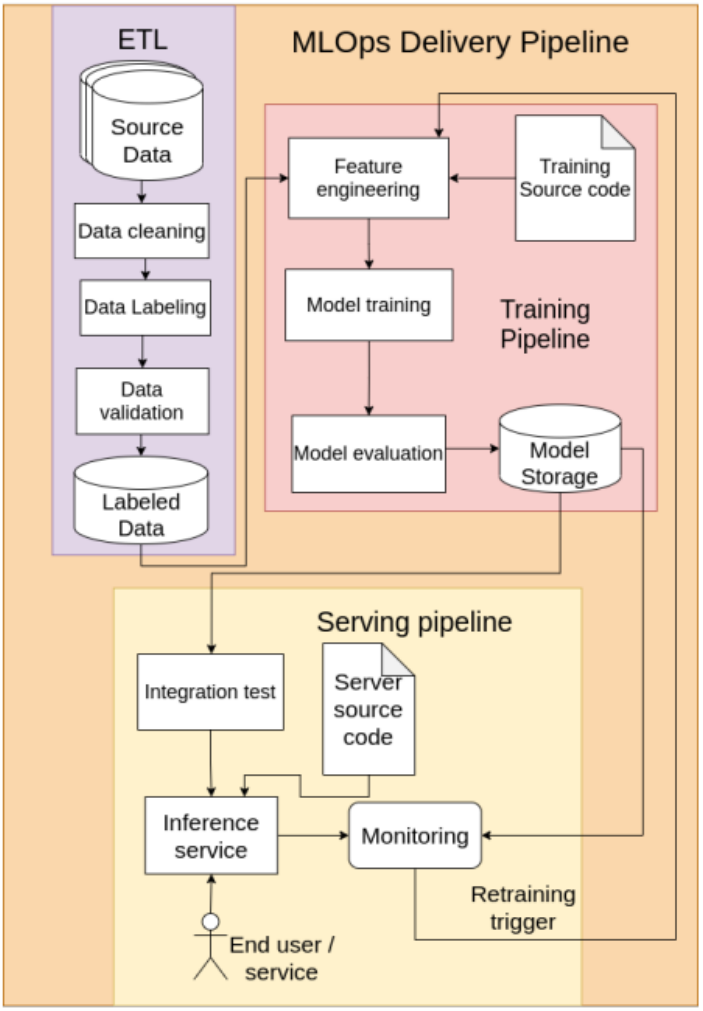


Figura 6. Pipeline de MLOps [25]

Este documento presenta una MLOps delivery pipeline construida por los autores, la cual representa o se asemeja a varias subetapas que se presentan en el modelo de

referencia presentado en esta monografía. Existen varias etapas similares como la limpieza de datos, validación de datos, entrenamiento de modelos, ingeniería de características, entre otras. Sin embargo, es posible notar que existe información clave que está dentro del modelo de referencia el cual no se encuentra en esta pipeline. Como lo son:

- Etapa de comprensión del negocio, esto incluye todas las subetapas internas de este.
- Etapa de análisis de datos
- Subetapa de retroalimentación de métricas.

Sin embargo, es esencial resaltar que al tratarse de una pipeline, no es necesario que posea un trayecto largo, completo y complejo como el del modelo de referencia. Por ende, es importante atribuir a los autores el crédito de que la pipeline ayuda a aumentar el rendimiento del desarrollo en la automatización de CI/CD de proyectos de aprendizaje automático. Y que el modelo de referencia está creado para seguir un proyecto desde que inicia hasta que termina.

**AN INTELLIGENT DEVOPS PLATFORM RESEARCH AND DESIGN BASED ON MACHINE LEARNING [26]**

En este documento los autores han realizado una extensa investigación sobre DevOps y machine learning, para así entender claramente las bases de MLOps (Machine Learning Operations). El documento presenta un estudio en donde se puede extraer información importante para proyectos con cimientos en MLOps, ya que menciona algunos de los beneficios de MLOps y las razones por las cuales es conveniente usarla. Las razones mencionadas en este documento son:

- Colaboración
- Automatización
- Innovación rápida
- Implementación rápida y sencilla
- Gestión eficaz del ciclo de vida

Sin embargo, también aclara que los modelos de aprendizaje automático al ser modelos más complejos que el desarrollo de software típico, también implican una mayor complejidad y un esfuerzo extra en la realización de diferentes tareas como la extracción de datos, la configuración y aprovisionamiento de infraestructura, monitorización, entre otras.

No obstante, el resultado de automatizar DevOps es el aumento de calidad del modelo, minimización de desperdicio y adaptación de ajustes tempranos, un ciclo de retroalimentación rápida, entre otros. Los autores mencionan y recomiendan el uso

de MLOps ya que de esta manera las empresas mejorarán significativamente la eficiencia de sus operaciones y producción.

A partir de este estudio, se pueden encontrar fuentes y conceptos similares tanto en este artículo como en la monografía presentada en este documento, en donde se puede interpretar que la revolución de la tecnología se centra en sistemas inteligentes de aprendizaje automático. Además de ello, se extraen conclusiones similares como que los sistemas que utilizan los principios de CI/CD contribuyen con la eficiencia, la eficacia, ciclos de retroalimentación rápidos y adaptación de nuevos ajustes en los sistemas, por tanto mejoran la calidad de los sistemas.

**Diferencias principales entre este modelo de referencia y los trabajos relacionados:**

| **MODELO DE REFERENCIA** | **TRABAJO RELACIONADO** |
|---|---|
| Presenta puntos esenciales como:<br><br>- Kick off<br>- Costos<br>- Desarrollo de una pipeline<br>- Retroalimentación de métricas de desempeño del modelo | **MLOps: una taxonomía y una metodología**<br><br>Presenta los siguientes puntos:<br><br>- La huella de carbono<br>- IA explicable |
| Presenta puntos esenciales como:<br><br>- Costos<br>- Planteamiento de una hipótesis para el análisis exploratorio de datos | **Machine Learning Operations (MLOps): Overview, Definition, and Architecture**<br><br>Presenta los siguientes puntos:<br><br>- Flujo de trabajo con roles funcionales |
| Presenta puntos esenciales como:<br><br>- Etapa de comprensión del negocio, esto incluye todas las subetapas internas de este.<br>- Etapa de análisis de datos<br>- Subetapa de retroalimentación de métricas. | **Designing an open-source cloud-native MLOps pipeline**<br><br>Presenta un pipeline que posee varias fases similares a este modelo relacionado, sin embargo no tiene en cuenta etapas como la comprensión del negocio, análisis de datos y retroalimentación de métricas. |

| | |
|---|---|
| Presenta un modelo de referencia a llevar a cabo cuando se realizan proyectos de ciencia de datos. | **An intelligent DEVOPS platform research and design based on machine learning**<br><br>Presenta un estudio en donde se puede extraer información importante para proyectos con cimientos en MLOps. |

# Capítulo 3

# Prácticas internas de la empresa

## 3.1. Identificación de las prácticas actuales de la empresa

La empresa WIZIT MIND BLOWING SOLUTIONS S.A.S (previamente conocida como “The Bit Bang Company”) posee 3 áreas para brindar servicios de desarrollo de software.

La primera es el área de ingeniería, en donde se diseñan, crean y prueban sistemas software, ya sea aplicaciones web, aplicaciones móviles, entre otras. La segunda área que comprende la empresa, es el área de analítica de datos, en donde se interpretan datos de carácter cualitativo o cuantitativo que han sido recopilados previamente en un determinado periodo de tiempo, con el propósito de ser analizados y usados en la generación de soluciones a los diferentes proyectos a cargo. Finalmente, la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S ha decidido abrir paso a una tercera área llamada área de innovación. Que consiste en buscar nuevos métodos ya sea de comercialización u organización para mejorar diferentes aspectos de la empresa, y por consecuencia mejorar los productos que se ofrecen.

Dentro de la empresa, existen líderes que controlan cada una de las áreas, para ello, cada equipo debe realizar una reunión cada viernes, esto con el objetivo de llevar un seguimiento de cada uno de los proyectos. Recientemente, la empresa ha comenzado a hacer uso de la plataforma “Monday” para administrar los proyectos. Esta plataforma permite que los equipos de trabajo planifiquen y gestionen el proceso de cada proyecto, esto para mantener una eficiente alineación entre cada equipo de trabajo.

Ahora bien, con el fin de indagar a fondo en el área de analitica de datos, uno de los objetivos comprendidos por este proyecto es identificar las prácticas actuales de la empresa, con respecto al desarrollo de proyectos de ciencias de datos que se describirán a detalle en la siguiente arquitectura presentada en la Figura 7:

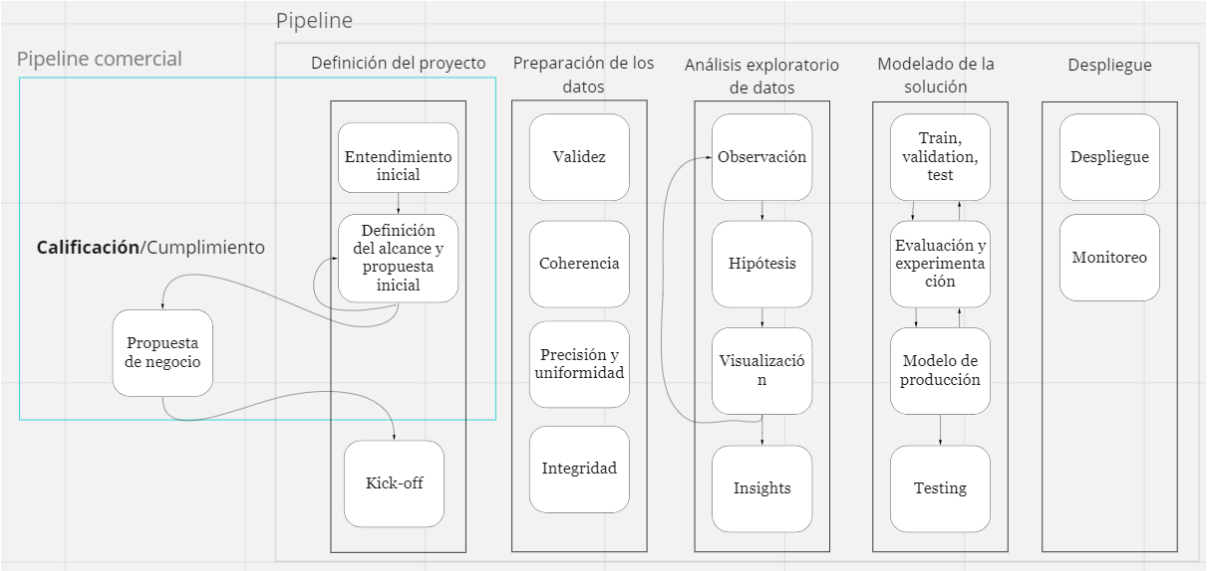


Figura 7. Arquitectura usada en el área de analítica de la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S. para el desarrollo de proyectos de ciencia de datos [fuente propia de la empresa]

Para el desarrollo de esta arquitectura se utilizó la plataforma “Miro”, debido a que agiliza la organización de los proyectos. Esta herramienta se encuentra en la nube, lo que facilita el trabajo remoto, es colaborativa y tiene la ventaja de simplicidad de uso.

En la primera etapa se realiza la definición del proyecto, en donde se obtiene un entendimiento inicial, lo cual implica hacer una primera aproximación con el cliente para entender el problema de manera superficial, y con ello comprender los requerimientos y los resultados que el cliente demanda para el proyecto. El reto en esta primera fase es descubrir factores relevantes que afectarán el proyecto, tanto en su desarrollo como en su fase final. Para ello, el equipo de trabajo de WIZIT MIND BLOWING SOLUTIONS S.A.S debe indagar en ciertos aspectos como:

- Descripción general del proyecto de investigación
- Descripción del problema que debe resolverse
- ¿Por qué es importante este problema?
- ¿Cuáles son los conjuntos de datos a los que los participantes pueden acceder para resolver este problema?
- Enumere y describa el conjunto de datos específico
- ¿Qué tipo de conjuntos de datos externos pueden ser de interés para el problema?
- Resultados esperados del proyecto

Ahora bien, después de responder estas preguntas y teniendo una buena

percepción del proyecto en el entendimiento inicial, se define el alcance del proyecto y se hace un modelo de propuesta inicial. Para este modelo, se elabora una primera versión de la propuesta técnica de éste, y cubre varios aspectos como tiempos de desarrollo, la arquitectura de solución en la cual se exponen a nivel técnico los elementos que implica el desarrollo del proyecto, el cronograma de actividades, los insumos que serán necesarios para la ejecución del mismo y además de ello, se cumplen 8 puntos importantes para la documentación de este, los cuales son:

1. Problema de negocio
2. Impacto del negocio
3. Datos
4. Métodos
5. Interfaz
6. Entregables
7. Cronograma
8. Riesgos

Es importante resaltar que este modelo inicial podría tener cambios, como nuevos requerimientos internos o externos, o distintas modificaciones que podrían presentarse a lo largo del progreso del proyecto. Dicho esto, es necesario interactuar con el cliente múltiples veces para llegar a un acuerdo con el fin de esclarecer la propuesta de negocio. Esta propuesta presenta al cliente el costo del proyecto conforme a lo que se haya solicitado previamente, en donde se incluyen impuestos, y demás costos que repercuten en el precio final del producto. Una vez se haya terminado este, se prosigue con el kick-off del proyecto, en el que se tiene certeza de todos los detalles del proyecto, y se presentan al cliente para así poder dar inicio a las actividades del proyecto. Es importante mencionar que existen 2 reuniones principales que se realizan en esta fase. La primera reunión es la interna, en la que se entrega el proyecto al equipo de ingeniería, y la segunda es una reunión personal con el cliente. El Kick-off consiste en una presentación que incluye los siguientes puntos:

- Objetivo del proyecto
- Fases del proyecto
- Solución propuesta
- Entregables del proyecto los cuales desembocan los siguientes subpuntos:
    - Arquitectura de despliegue
    - Análisis de datos
    - Análisis de datos y modelos preliminares seleccionados
    - Modelo ML
    - Evaluación y ajustes
    - Documentación
    - Productos finales

- Equipo de trabajo
- Canales de comunicación
- Reuniones de seguimiento y consultoría
- Obligaciones

Considerando los puntos anteriormente mencionados y teniendo claridad de la ruta descrita como “definición del proyecto”, se prosigue a la siguiente etapa que consiste en la preparación de los datos, en donde se observa la validez que poseen los datos que el cliente provee a la empresa. Este paso pretende asegurar que los datos representan las reglas de negocio, es decir, los datos recientes y antiguos que el cliente entrega a la empresa, deben ser adecuados para el proyecto. De esta manera, se verifica si hay datos fuera de rango, espacios en blanco, balanceo de datos, etc. Con esto dicho, es válido resaltar que si el cliente no provee datos útiles, la preparación de los datos no se podrá realizar.

Una vez se comprueba la validez de los datos, se prosigue a verificar la coherencia de los mismos. Este proceso se realiza con herramientas proporcionadas por Python. Las actividades principales de esta etapa son la identificación de instancias duplicadas y la forma de la representación de los datos, es decir que los datos deberían representarse de la misma manera en toda la estructura de datos, por lo que el formato de los insumos entregados debería ser el mismo. No obstante, si se presenta el caso en donde no se cumple este requisito, entonces se debe trasladar todo a un mismo formato, para así poder acceder a toda la información y lograr manejarla eficientemente. Además los datos deberían ser consistentes, es decir, deben reflejar la realidad con lo que representan, para así tener certeza de la precisión y la uniformidad de estos. Ahora bien, efectuados todos los procesos anteriores, se prosigue a examinar la integridad de los datos, en donde se verifica la calidad de los datos y se busca asegurar que los datos estén limpios para que exista la menor cantidad posible de datos perdidos. Sin embargo, la integridad de los datos podría ser afectada por un mal proceso de captura de datos, o desde la fuente de datos, por eso es de vital importancia tener especial cuidado en cada etapa del proceso de preparación de datos.

Finalizando este paso, se continúa con la tercera etapa en la cual se lleva a cabo un análisis exploratorio de datos, por lo cual es necesario observar los datos, es decir entender los datos de manera estadística, y así analizar las tendencias de estos, seguido a ello se procede a plantear diferentes hipótesis con el objetivo de comprobar si las variables en el estudio son útiles para hacer la proyección. Luego, con todos los datos previamente recolectados, se procede a examinar cuál información es de valor y cuál no, para tomar decisiones correspondientes a los datos que se usarán para preparar el algoritmo y evaluarlos con el objetivo de obtener insights de negocio. Para esta etapa los insumos son el conjunto de datos procesado, y la salida son los reportes, el conjunto de datos para entrenamiento y el pipeline de procesamiento.

Al tener los datos limpios y listos para dicho algoritmo, se continúa con la etapa de modelado de solución en donde se inicia a entrenar el algoritmo con una porción de

los datos, y se valida con otros. Aquí se pretende dividir el conjunto de datos final en 3 subconjuntos: Train, validation and test (Entrenamiento, validación y pruebas).
Este proceso tiene como insumo el conjunto de datos procesado, y tiene 3 salidas las cuales son el conjunto de datos de entrenamiento, el conjunto de datos de validación y el conjunto de datos de evaluación. Seguido a ello, se continúa con el proceso llamado evaluación y experimentación en donde se elige candidatos de modelos de ML, para posteriormente realizar en ellos tareas de entrenamiento y proseguir a la evaluación en la cual se toman datos nuevos, y se analiza el comportamiento del algoritmo, apoyado en los resultados que arroje. Los insumos son el código de entrenamiento de cada modelo, y el conjunto de datos de entrenamiento y validación. Por otra parte, la salida deberían ser los modelos candidatos.

Después de pasar por estas etapas se realiza el modelo en producción, en donde se elige el modelo de machine learning que más se adecua al proyecto, esta elección se realiza con base en distintas métricas de evaluación.

Es importante resaltar que en este paso los insumos o entradas son los modelos candidatos, el conjunto de datos de evaluación y las métricas de evaluación, mientras que la salida son los modelos de producción y el código del modelo seleccionado. Al finalizar este proceso se procede a hacer el testing en donde se comprueba que el código del modelo esté funcionando correctamente en el ambiente de producción. En el testing o pruebas, los insumos son el modelo de producción, código del modelo de producción y casos de prueba (manuales o automatizados), y la salida son el reporte de pruebas (manual y automático) y el reporte de recomendaciones.

Por último se realiza el despliegue, el cual pretende construir el sistema de ML que utiliza apropiadamente el modelo de ML construido, en esta etapa el proyecto debe pasar por cada área de seguimiento de proyectos que tiene la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S para verificar el óptimo funcionamiento de este. Inicia con el área de ingeniería, que en la empresa está considerada como el área de más alto nivel, y continúa con el área de analitica, en donde el director asigna los miembros de cada equipo y el líder respectivo dependiendo de las habilidades de cada integrante y los requerimientos del proyecto, dichos equipos usan la herramienta de google sheets para la correcta gestión y seguimiento de las actividades. Sin embargo, esta última etapa no posee ningún modelo de referencia o conjunto de lineamientos estandarizados, así que se usa una API REST en la construcción del modelo para que el cliente obtenga la predicción o especificación, en donde se monta una API sobre internet que recibe peticiones http y contesta las predicciones.

Ahora bien para el despliegue se suelen usar los contenedores de docker, no obstante, no se cuenta con ninguna estrategia de despliegue (canary deployment [27], blue/green deployment [28]) ni de seguimiento al desempeño del modelo. Los

insumos para esta fase son modelo de ML de producción y el código de la aplicación, mientras que la salida es el sistema de ML en producción.

Por otra parte, la empresa ha agregado el uso de las prácticas de Git para el desarrollo de proyectos software, con el fin de controlar las diferentes versiones de las aplicaciones, tanto nuevas como pasadas y así llevar un registro de cada cambio que se realice. Git es el sistema de control de aplicaciones más usado por desarrolladores, debido a que permite realizar diferentes acciones como organizar y administrar el trabajo para múltiples desarrolladores, manejar las versiones de aplicaciones que se estén desarrollando, guardar el trabajo, etc [29].

En Git se manejan algunos conceptos importantes que serán mencionados a continuación:

**Repositorio o Repo:**

Es un archivo o varios archivos que contienen la información de los proyectos, como Github, Bitbucket, Gitlab [30].

**Branches (Ramas):**

Es importante entender que cada repositorio contiene una línea de tiempo en donde se registran los cambios que se hagan en cada archivo, estos registros serían las ramas.

Las ramas se definen como los registros en donde se marcan los diferentes cambios que los desarrolladores están realizando en la aplicación.

Ahora bien, los repositorios contienen una rama master, la cual es la rama principal. Así que, a partir de esta rama, es posible crear otras ramas, lo que corresponde a crear otras líneas de tiempo.

En cada una de las ramas configuradas es posible gestionar cambios, al igual que en la rama master, sin que se vea tergiversado el contenido de la rama master.
Las ventajas de crear y configurar nuevas ramas es que siempre se tendrá una rama principal inalterada, lo cual permite hacer cambios al proyecto sin que se vea afectado (lo que brinda seguridad al equipo de trabajo). Además, permite testear nuevas funcionalidades que se requieran para el proyecto, antes de ser integradas a la rama principal. Un dato importante de mencionar, es que las ramas pueden ser mezcladas (Merge), borradas y creadas, de esta forma se actualiza el contenido en la rama principal.

Además, la empresa WIZIT MIND SOLUTIONS S.A.S sigue el flujo de trabajo Git Flow [30], que contribuye a un uso y manejo adecuado de cada una de las ramas que se van construyendo. Gitflow clasifica las ramas de la siguiente manera:

**Master**

Es la rama principal, de despliegue y producción, la cual contiene todas las versiones estables del proyecto [30].

**Hotfix**

Es una rama de apoyo, y surge a partir de la rama master. Esta rama se crea con el objetivo de arreglar un error que se debe arreglar urgentemente. Una vez se haya solucionado el problema, se incluye dicho contenido en la rama master para arreglar el incidente que se había presentado, para así tener una versión modificada de producción [30]. Es importante resaltar que esta versión debe ser marcada con un tag en la rama master.

**Release**

Es una rama de apoyo y surge a partir de la rama develop. Esta rama contiene el código de la versión que será liberada próximamente. El objetivo de esta rama es habilitar pruebas en un ambiente real, como las pruebas del cliente. Una vez se haya finalizado con las pruebas y éstas sean exitosas, esta rama se combina o se incluye en la rama develop y la rama master, para que salga a producción [30].

Por tanto, esta rama se crea para hacer la liberación y se borra al tenerla integrada con la rama Master.

**Develop**

Es una de las ramas principales, debido a que es la rama que contiene el código. Esta rama es de prueba y desarrollo, es decir, es la rama en donde se hacen cambios en el código para añadir nuevas funcionalidades de la aplicación. Esta rama puede incorporarse tanto en una rama release como en una master para su posterior despliegue [30].

Cabe mencionar que no se aconseja guardar cambios directamente en la rama de desarrollo, excepto si son cambios que no afectan la lógica del código.

**Feature**

Es una rama de apoyo. Esta rama al igual que la rama release, debe surgir de la rama develop y almacena el código de desarrollo con las nuevas características (features) que hayan sido previamente solicitadas. Una vez se haya terminado con el desarrollo de las nuevas funcionalidades, esta rama se mezcla con la rama develop para que los cambios queden actualizados y sincronizados en la última versión del código [30].

Con lo anterior dicho, en la figura 8 se puede apreciar un ejemplo de flujo de trabajo mientras se desarrolla un proyecto.


Figura 8. Gestión de versiones con Git usado en la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S [30]

La versión 0.1 es en donde se crea el proyecto y aplicación. Después se crea una rama develop para comenzar con el desarrollo del proyecto, y a sus vez se generan dos ramas feature (con nombres diferentes, pero con la palabra feature en ellas) para el desarrollo de nuevas características. En este proceso se observa como la primera rama feature ha sido finalizada con éxito y combinada con Develop.

También se observa que se ha creado una rama hotfix, es decir que ocurrió un incidente que debe ser arreglado urgentemente, después de solucionar dicho error, la rama hotfix se mezcla con las ramas principales master y develop para luego ser eliminada.
Por último, se genera una rama release a partir de la rama develop. Esta rama se crea con el fin de realizar las pruebas finales y tomar la decisión de aprobar o no el nuevo código.

# Capítulo 4

# Estructuración del modelo de referencia propuesto

Para la realización de este trabajo de grado se realizaron algunas actividades previas con el fin de estructurar la mejor versión posible del modelo de referencia propuesto para la empresa, las actividades son las siguientes:

## 4.1. Actividad 1: Diagrama de flujo con la estructura del modelo de referencia propuesto

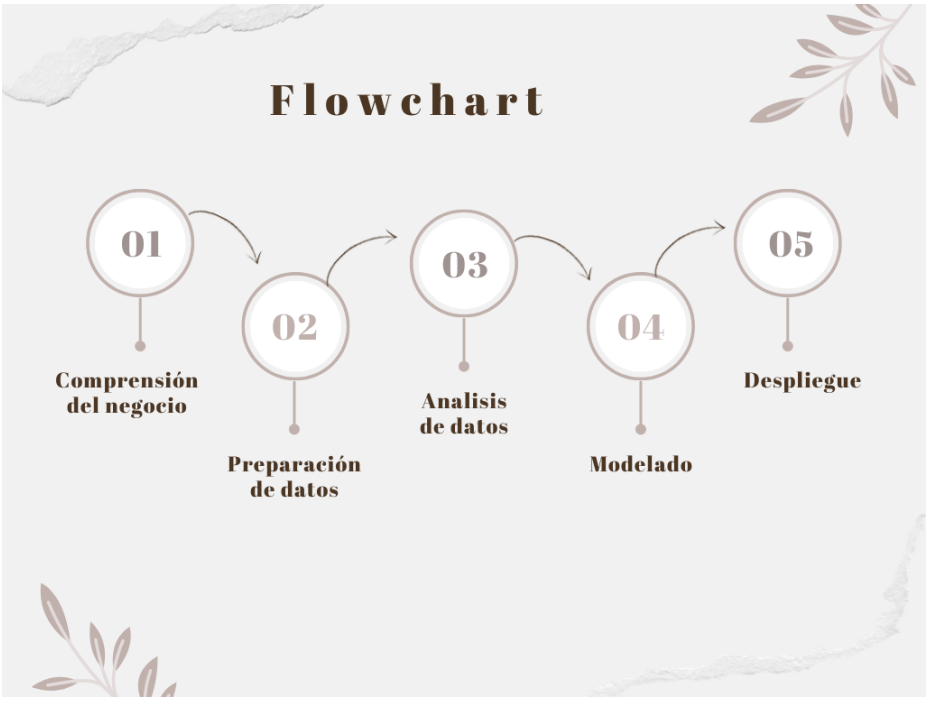


Figura 9. Diseño de diagrama de flujo del modelo de referencia propuesto con base a los objetivos planteados y las exigencias de la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S [fuente propia]

Para la construcción de esta arquitectura se realizó una búsqueda previa sobre las diferentes metodologías ágiles existentes para proyectos de ciencias de datos. La metodología en la que se ha basado dicha arquitectura es DDS (Data Driven Scrum), la cual combina varias prácticas de Scrum y Kanban para apoyar a los científicos de datos, e ingenieros de datos. Se eligió esta metodología debido a que varios marcos ágiles como lo son Scrum, Kanban, Crisp DM, entre otros, no son eficientes para los proyectos de ciencias de datos, por lo que no cubren las complejidades que un proyecto de ciencia de datos puede tener. Algunos de los conceptos que se utilizan en el marco de DDS son [22]:

- Backlog Item (BItem)
- Ítem Backlog (IBB)
- Task Board
- Iterations

## 4.2. Actividad 2: Realizar una primera versión de un esquema utilizando la herramienta miro.

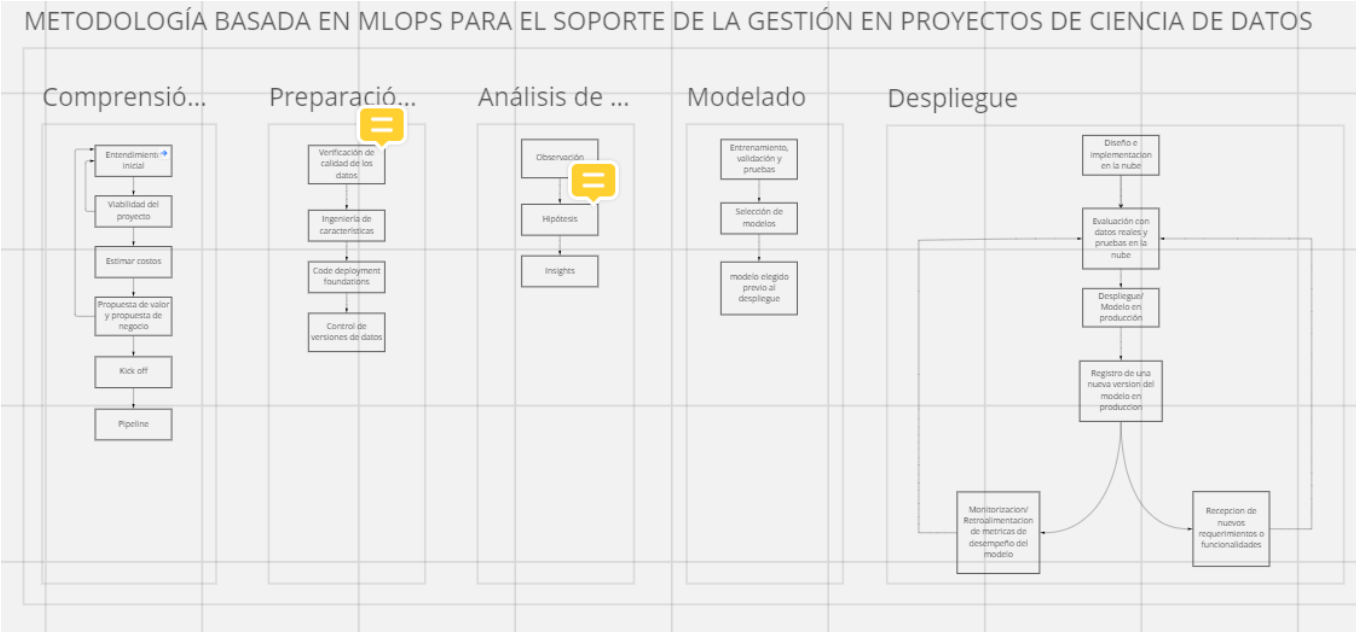


Figura 10. Primera versión del modelo de referencia propuesto [fuente propia]

## 4.3. Actividad 3: Descripción paso a paso del modelo de referencia propuesto

Esta actividad incluye imágenes referentes a cada sub-etapa, y explicación detallada de cada una de las 5 etapas con las sub-etapas correspondientes.

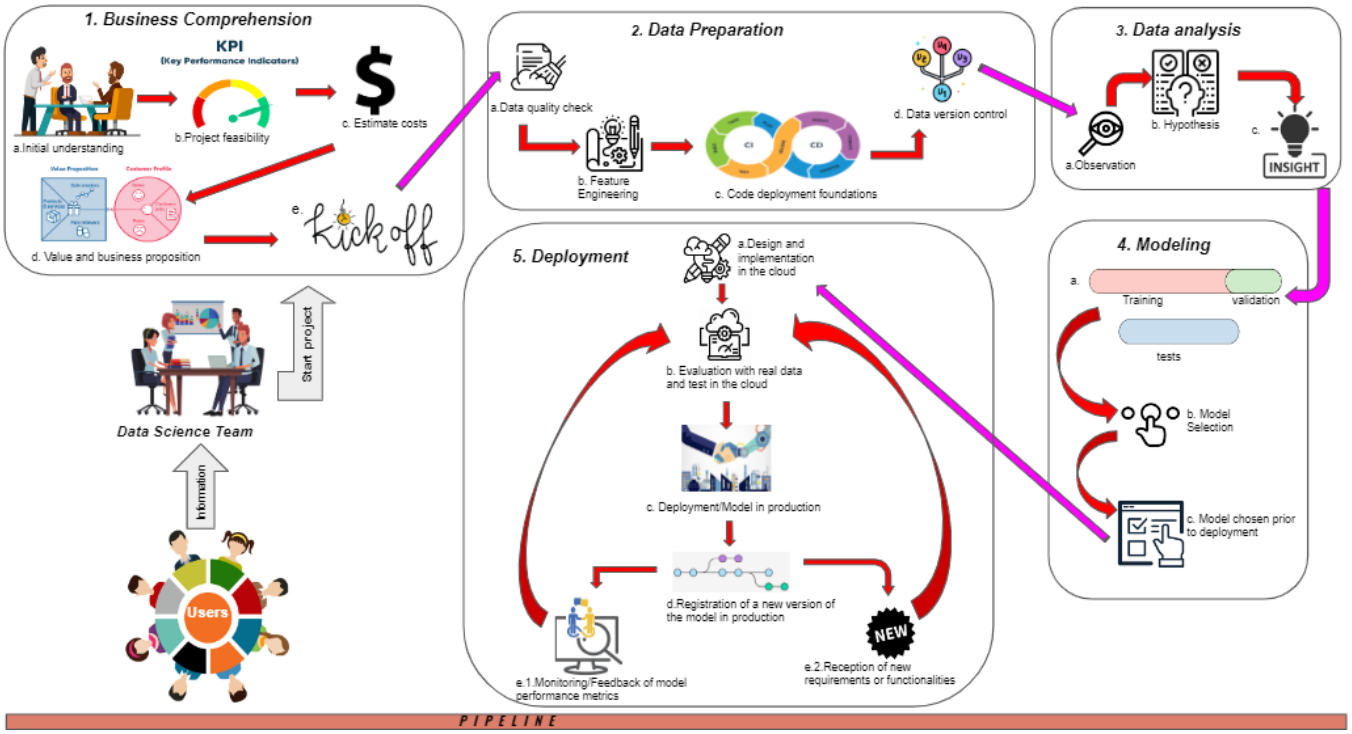


Figura 11. MODELO DE REFERENCIA BASADO EN MLOPS PARA EL SOPORTE DE LA GESTIÓN EN PROYECTOS DE CIENCIA DE DATOS [fuente propia]

**PRIMERA ETAPA: Comprensión del negocio**

La primera etapa se titula **Comprensión del negocio** la cual comprende 6 subetapas que son:

- Entendimiento inicial
- Viabilidad del proyecto
- Estimación de costos
- Propuesta de valor y de negocio
- Kick-off
- Pipeline

**Entendimiento inicial**

Como su nombre lo indica, en esta subfase se entiende el problema de manera superficial, se realiza por medio de la primera reunión con el cliente. El objetivo es que el cliente brinde la información necesaria y suficiente para que el equipo encargado del proyecto logre comprender los requerimientos que dicho cliente solicita y los resultados que se esperan. Para ello, es importante obtener respuesta a una serie de preguntas, que son:

- ¿El problema a resolver es de un cliente o de la empresa misma?
- ¿Se requiere implementar el modelo de ML en servicios de la nube?
- ¿Cuál es el problema a resolver?
- ¿Por qué este problema afecta a la empresa?
- ¿A qué datos se tiene acceso para resolver este problema?
- ¿Hay conjuntos de datos externos que puedan ser relevantes para la solución del problema? Si es así, ¿Cuáles son?
- ¿Qué resultados se esperan? ¿El resultado final es un modelo de ML, un reporte analítico, un dashboard?

**Viabilidad del proyecto**

Aquí se recopilan criterios de éxito técnico junto con criterios de éxito comercial. Estos criterios deben ser medibles. Por lo tanto, es necesario definir indicadores clave de rendimiento (KPIs). Los KPIs son medidas que indican el rendimiento o desempeño del proyecto que se está trabajando. Dicha medida suele expresarse con porcentajes. Los KPIs permiten observar el estado actual de un negocio con respecto a un área en concreto, y a partir de ahí, actuar para optimizar las estrategias [31].

Las principales ventajas del uso de indicadores KPI son [32]:

- Medición constante, se realiza con el objetivo de actuar de forma flexible y rápida en la optimización de la estrategia o proceso a desempeñar. Algunas veces se realiza esta medición constante en tiempo real.

- La adaptación: El negocio debería poder adaptarse a continuos cambios que se puedan presentar en el mercado, o cambios que se generen como nuevos clientes, nuevo mercado competencia, nuevas oportunidades, entre otros.

- Se genera motivación en los empleados y equipos de trabajo para conseguir los objetivos fijados.
- Las actividades se realizan de forma lógica lo que brinda tranquilidad a los inversores, directores y demás grandes cargos relacionados con el negocio que generalmente no se encuentran en el día a día de trabajo.

**Recomendaciones**

Aplicar el marco de machine learning canvas [33]. Este marco permite a todas las partes interesadas especificar la disponibilidad de datos, las restricciones reglamentarias y los requisitos de la aplicación, como solidez, escalabilidad, y demanda de recursos.

**Recomendación de amazon**

Amazon Quick Sight es un servicio de análisis empresarial muy eficaz, de fácil uso, y administrado en la nube que ayuda a los empleados de una organización a realizar una correcta y completa compilación de visualizaciones, una rápida obtención de información empresarial a partir de sus datos en cualquier momento y con la facilidad de realizarlo en cualquier dispositivo y un buen análisis de información ya que conecta a los datos en la nube y combina datos de muchas fuentes diferentes [34]. En amazon Quick Sight un KPI muestra una comparación de valores, los dos valores que se comparan y una barra de progreso.

**Estimación de costos**

Esta propuesta presenta al cliente el costo del proyecto conforme a lo que se haya solicitado previamente, en donde se incluyen impuestos, costo de la adquisición, almacenamiento, procesamiento de datos, recursos humanos y demás costos que repercuten en el precio final del producto.

**Recomendación de AWS**

Se recomienda seguir el pilar de optimización de costos AWS Well-Architected Framework [35] ya que este documento incluye información acerca de presupuestos, administración y optimización de costos, creación de informes para analizar la rentabilidad del trabajo, implementación de objetivos, entre otros. Además recomienda herramientas a usar dependiendo la sección que se encuentre, por ejemplo, para la notificación de costos y creación de informes recomienda herramientas como:

- AWS Cost Explorer
- AWS Trusted Advisor
- AWS Budgets

Otra recomendación esencial para estimar costos es la calculadora de estimación de precios de AWS la cual proporciona una estimación de las cuotas y los cargos de AWS, sin embargo dicha estimación no incluye los impuestos que se puedan aplicar

a estos conceptos. AWS Pricing Calculator proporciona información sobre precios a título únicamente informativo. Si los precios que figuran en las páginas de marketing son diferentes de los precios que utiliza AWS Pricing Calculator, AWS respeta los precios de las páginas de marketing [36].

**Propuesta de valor y de negocio**

La pregunta que surge en esta subetapa es: ¿Por qué un cliente se beneficiaría del uso de nuestro software o servicio?

El producto o servicio que se esté desarrollando debe crear valor para un usuario final. De igual manera, el proyecto que se esté realizando debe tener una propuesta de valor que atraiga y motive al cliente a usar los servicios o comprar los productos de la empresa.

Por otra parte en la propuesta de negocio se presenta al cliente el costo del proyecto conforme a lo que se haya solicitado previamente, en donde se incluyen impuestos, y demás costos que repercuten en el precio final del producto, este precio se debe tener definido en la etapa nombrada estimación de costos.

**Recomendación**

Value proposition canvas: Es una herramienta que ayuda a investigar más a fondo los intereses de los clientes y la creación de valor, la siguiente figura muestra el modelo mencionado anteriormente:

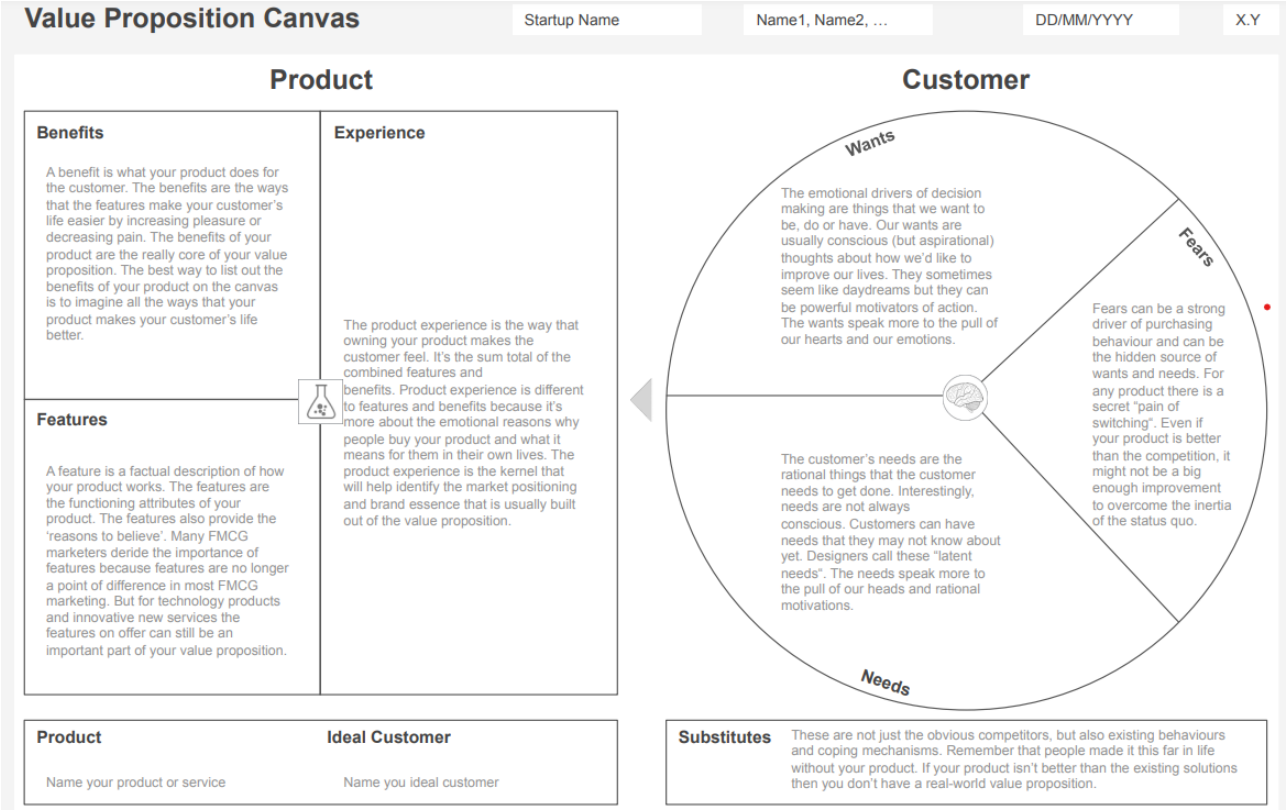


Figura 12. Value proposition canvas

**Recomendación de AWS**

**Propuesta de valor de Amazon [37]**

Amazon es considerada la tienda online número uno en la compra-venta de artículos por internet, ya sea entre una entidad oficial y el cliente en cuestión, o entre los usuarios de la misma web. Este sitio web se adapta continuamente a las necesidades del cliente en varios factores como: precio, usabilidad, reviews o uso previo por otros consumidores y características.

En cuanto a la propuesta de valor hay tres puntos importantes que presenta amazon que son: **mapa del cliente, mapa de valor y encaje entre ambas partes.**

**Mapa del cliente de Amazon,** el cliente es el encargado de realizar la búsqueda y de comercializar aquello que ya no quiere, no necesita, o porque decide conseguir un dinero extra por la venta de algo que ya no usa frecuentemente. Para ello, Amazon implementa la sección de gestionar productos al uso, además de adquirir un artículo según la temporalidad de envío que se necesite, con amplia variedad de opciones de entrega.

**Mapa de valor de Amazon,** Amazon se asegura de cumplir las expectativas del consumidor. Este es uno de los factores más importantes de Amazon, ya que

permite la creación de valor de cliente a la medida justa que el cliente crea conveniente.

**Encaje entre ambas partes,** se refiere a una continua adaptación de las necesidades que exige el consumidor, porque lleva a cabo una continua renovación en los productos y tecnológica, y esto se puede notar debido a que cuando ocurre cualquier problema en la ejecución de la compra o búsqueda, Amazon busca solucionarlo para la siguiente entrega y ofrece distintas características a cabo como Amazon Premium, para aquellos que busquen una entrega más rápida por una módica carga mensual.

**Kick-off**

En este punto se han determinado todos los detalles del proyecto, y se presentan al cliente para dar inicio a las actividades del proyecto. En este punto se deben llevar a cabo 2 reuniones principales. La primera reunión es con el equipo de ingeniería, y la segunda es una reunión personal con el cliente. El Kick-off es una presentación que incluye los siguientes puntos:

- Objetivo del proyecto
- Fases del proyecto
- Entregables y solución propuesta
    - Arquitectura de despliegue
    - Análisis de datos
    - Modelo ML (Producto o servicio en la nube)
- Equipo de trabajo
- Canales de comunicación
- Reuniones de seguimiento y consultoría

**Project pipeline**

Es importante crear un pipeline para gestionar los proyectos debido a que de esta manera es posible monitorear el estado de los proyectos actuales en una sola ventana. Esto con el fin de obtener una descripción detallada de cada uno de los proyectos que se están trabajando y así priorizar rápidamente proyectos de alto impacto y manejar cualquier obstáculo en el camino [38].

Se recomienda crear una pipeline del proyecto ya que es una manera sencilla de monitorear el progreso del equipo e identificar cualquier problema en el flujo de trabajo. Se puede crear y realizar un seguimiento de un ciclo continuo de ideas, aprobación, implementación, producción, revisión y finalización de proyectos, todo administrado en un solo lugar. Además de ello se puede ver qué proyectos están en camino de alcanzar sus objetivos y aquellos que necesitan ajustes o extensiones [38].

Project Pipeline Management implica 5 etapas para garantizar que se cree, discuta y evalúe una cantidad conveniente de  propuestas de proyectos para cumplir con los objetivos que se hayan planteado.

**Las 5 etapas de la gestión de proyectos son [39]**

**1. Ideación**

En esta etapa se genera una lluvia de ideas. La ideación es importante para capturar las propuestas de proyectos necesarias para crear proyectos de alta calidad. En donde se recomienda que cada integrante del equipo comparta aportes únicos y nuevos.

**2. Proyectos propuestos**

Después de generar la lluvia de ideas, se extraen las mejores ideas para trabajar en ellas, y poder discutir factores importantes como el alcance del proyecto, el presupuesto y los resultados deseados.

**3. Planificación**

Se debe planificar adecuadamente con el fin de dividir el proyecto en múltiples tareas, y así poder asignarlas a miembros clave del equipo y crear un plan para monitorear el progreso de este.

Algunos pasos cruciales a seguir durante la etapa de planificación:

- Crear tareas de proyecto y subtareas
- Crear un equipo y asignar los roles y responsabilidades dependiendo de las habilidades de cada integrante.
- Asignar tareas y subtareas a los miembros del equipo
- Asignar fechas de vencimiento para cada tarea
- Organizar reuniones para un seguimiento continuo

**4. Proyecto en curso**

En esta etapa se recomienda que el equipo:

- Actualice el progreso de las tareas con sus respectivas fechas de vencimiento
- Se organicen reuniones de equipo periódicas
- Realice un seguimiento de los indicadores clave de rendimiento (KPI)
- Identificar áreas problemáticas para poder resolverlas

**5. Proyectos completados**

En esta etapa los proyectos ya están completados y listos para ser entregados al cliente. En esta etapa es esencial realizar una evaluación de los resultados, tener un análisis del desempeño y descubrir áreas en las que se podría haber tenido un mejor desempeño.

**Recomendación**

Monday

**SEGUNDA ETAPA: Preparación de datos**

La preparación de datos comprende 5 subetapas que son:

- Verificación de calidad de los datos
    - Limpieza de datos
- Ingeniería de características
- Code deployment foundations (Define las herramientas a usar. Se debe ejecutar en paralelo con las etapas de verificación de calidad e ingeniería de características)
- Control de versiones de datos (La estandarización de datos se encuentra inmersa en este paso)

**Verificación de calidad de los datos**

En este paso se verifica la validez que poseen los datos que el cliente le entrega a la empresa. Este paso pretende garantizar que los datos cumplan las reglas de negocio, es decir, los datos que el cliente provee a la empresa, deben ser apropiados para el proyecto. De esta manera, se verifica si hay espacios en blanco, datos fuera de rango, balanceo de datos, etc. Con esto dicho, se debe mencionar que si el cliente no entrega datos útiles, no será posible realizar la preparación de datos.

Una vez se verifica la validez de los datos, se prosigue a corroborar la coherencia de estos. Este proceso se lleva a cabo con herramientas proporcionadas por Python. Las principales actividades de esta etapa son  la forma de la representación de los datos y la identificación de instancias duplicadas, es decir que los datos deberían representarse de la misma manera en toda la estructura de datos, por lo que el formato de los insumos entregados debe ser el mismo. No obstante, si este requisito no se cumple, entonces todo debe ser trasladado a un mismo formato, para así poder acceder a toda la información y lograr manejarla eficientemente.

Además los datos deben ser consistentes, lo que significa que deben reflejar la realidad con lo que representan, para así tener certeza de la precisión y la uniformidad de estos. Ahora bien, después de realizar todos los procesos anteriores, se prosigue a verificar la integridad de los datos, en donde se verifica la calidad de los datos y se busca lograr que los datos estén limpios para que exista una baja cantidad de datos perdidos. Sin embargo, la integridad de los datos podría ser afectada por un mal proceso de captura de datos, o desde la fuente de datos, por eso es esencial tener precaución en cada etapa del proceso de preparación de datos.

**Ingeniería de características [40]**

La ingeniería de características es un proceso que se utiliza para seleccionar y transformar variables cuando se crea un modelo predictivo mediante el aprendizaje automático. Por lo general, la ingeniería de características incluye la creación, la transformación, y la extracción de características.

- La extracción de características identifica las características en el conjunto de datos que son relevantes para el problema en cuestión.
- La transformación de características administra el reemplazo de las características que faltan o no son válidas.
- La creación de características es el proceso de crear nuevas características a partir de las existentes. En general, el objetivo es reducir la dimensionalidad de las características.

**Code deployment foundations**

Esta subetapa se encarga de aplicar las mejores prácticas de DevOps a las actividades de MLOps. Sin embargo, si existen brechas en las prácticas tradicionales de DevOps, se debe mejorar primero antes de comenzar con actividades más complejas, como el control de versiones de datos y modelos, la capacitación continua de modelos o el almacenamiento de funciones. En este paso se realizan las siguientes preguntas:

- ¿Cómo se mantiene el código?
- ¿Qué sistema de control de versiones de código fuente se utiliza?
- ¿Cómo se monitoriza el rendimiento del sistema?
- En cuanto a la automatización de implementación y prueba: ¿Cuál es la canalización de CI/CD para el código base? ¿Qué herramientas se utilizan para ello?. Algunos ejemplos de herramientas son: Jenkins, AWS, Azure, DevOps, GitLab.

**Control de versiones de datos**

Es un mecanismo para gestionar y rastrear los cambios que se generen en el código de software, ya que se tiene control sobre qué versión del conjunto de datos se utilizó para entrenar un modelo específico. En este paso es importante responder a las siguientes preguntas:

- ¿Este control de versiones de datos es opcional u obligatorio?
- ¿Qué fuentes de datos están disponibles? (p. ej., datos propios, públicos, pagados)
- ¿Cuál es el almacenamiento para los datos anteriores?

**TERCERA ETAPA: Análisis de datos**

El análisis de datos consta de los siguientes 3 sub-etapas:

- Observación
- Aplicabilidad de machine learning (Hipótesis)
- Insights

**Observación**

Es el proceso por el cual se determinan patrones específicos en los datos que se poseen. Estos patrones se observan de manera estadística, de diferentes maneras, ya sea en tablas, histogramas, gráficos entre otros. El objetivo de esta subetapa es entender la tendencia que tienen los datos para así poder interpretarlos y analizarlos correctamente.

**Hipótesis**

Seguido a ello se procede a plantear diferentes hipótesis con el objetivo de comprobar si las variables en el estudio son útiles para hacer la proyección. Dado el caso que se desee implementar un modelo de Machine Learning es importante verificar la aplicabilidad de ML, para saber esto se pueden responder las siguientes preguntas:

- ¿Qué lenguaje de programación utilizar para el análisis? (R, Python, Scala, Julia, SQL?)
- ¿Existen requisitos de infraestructura para la formación de modelos? Ejemplo: capacidad de almacenamiento, rentabilidad, seguridad, rendimiento informático (recursos informáticos).
- ¿Qué métricas de evaluación comercial (KPI) y específicas de ML deben calcularse? [41]

- Reproducibilidad: Para que un trabajo se pueda reproducir, es necesario capturar el entorno informático donde se realizó para que otros puedan replicarlo. Con ello surgen las preguntas:
    - ¿Se está usando alguna herramienta para crear un entorno reproducible? Ejemplo: conda, pip-tools, entre otros [42].
    - ¿Qué metadatos sobre los experimentos de ML se recopilan? (conjuntos de datos, hiper parámetros).

**Insights**

Con todos los datos que han sido recolectados, se procede a verificar cuál información es de valor y cuál no, ya que de esta manera es posible tomar decisiones correspondientes a los datos que se deberán usar para preparar el algoritmo y validarlos con el objetivo de obtener insights de negocio.

NOTA: Las técnicas más robustas para manejar los datos faltantes son Knn para imputación simple y MICE para imputación múltiple para no introducir sesgo [43]

**Explicación paso a paso de un análisis exploratorio para datos estructurados [44]**

Primero se debe crear una pregunta que corresponda a la solución del proyecto, luego se debe observar el dataset, es decir, se debe mirar el tamaño para determinar las características o variables (Las columnas de la tabla) y observar o dar un primer barrido de los datos registrados en las tablas (Las filas del dataset), para así tener una primera idea de los datos.

Después se prosigue a analizar en detalle el dataset, por tanto se debe definir a qué tipo de variable pertenecen dichos datos. Aquí se presentan dos opciones:

- Las variables numéricas: Pueden ser discretas (cuando toman valores enteros) o continuos (cuando toman cualquier valor dentro de un intervalo)
- Las variables categóricas: Pueden ser nominales (se usan para etiquetar el dato pero no pueden ser ordenados, ni medidos), binarios (Indican una de dos posibles categorías) o datos ordinales (corresponden al orden en que vienen representados los datos).

El siguiente paso es iniciar con la descripción estadística que depende del tipo de datos que se tenga en cada una de las variables y para esto se usan dos tipos de medidas:

- Las medidas de tendencia central
- Las medidas de variabilidad

Las medidas de tendencia central proveen una idea general del valor típico que pueden tener los datos, las principales son:

- La media
- La mediana

En donde la media representa el promedio de los datos y se puede aplicar a datos discretos y continuos, sin embargo la desventaja es que es muy sensible a valores atípicos, por lo tanto calcular el promedio no resulta muy conveniente. Sin embargo, la mediana es la solución a este inconveniente y se puede aplicar para datos ordinales o discretos. Para calcularla se debe organizar los datos de manera ascendente y la mitad de los datos estarán por debajo de este valor y la otra mitad por encima de este valor.

No obstante, no es suficiente con conocer la media o la mediana de la distribución porque también se debe tener en cuenta de qué tan agrupados o dispersos están los datos. Para determinar esto se utilizan las medidas de variabilidad donde las principales son la desviación estándar y el rango intercuartiles y nos indican que tanto se alejan los datos del valor medio o de la mediana respectivamente. Sin embargo, la desventaja de la desviación estándar es la misma de la media. Es muy sensible a los valores atípicos. Pero para ello existe una alternativa que es el rango intercuartil, el cual representa la diferencia entre el percentil 75 y el percentil 25.

Ahora bien, es importante resaltar que la limitación de las medidas centrales y de variabilidad es que son solo un número, lo que representa una idea muy general del comportamiento de los datos, así que el siguiente paso a seguir es visualizar los datos para poder encontrar más detalles, para datos continuos y discretos se puede calcular y dibujar el histograma que se obtiene tras organizar los datos en diferentes grupos y realizar el conteo de número de datos en cada uno, con el histograma se puede verificar si la distribución es normal es decir si tiene forma de campana o si está sesgada, como una campana asimétrica. la desventaja es que no permite ver los valores atípicos porque esos valores quedan enmascarados cuando los datos se introducen en uno de sus bits, así que la alternativa es usar los diagramas de caja o box plots que se pueden utilizar tanto en datos continuos como discretos, en un boxplots se dibujan los percentiles. En la representación del boxplots es posible notar que las barras superior e inferior corresponden a los percentiles 75 y 25, mientras que la línea en medio de la caja es la mediana. Por fuera de la caja hay dos líneas conectadas por líneas punteadas que se llaman whiskers y cada una de ellas es igual al percentil 75 o 25 más o menos 1.5 veces el rango intercuartil.

Es importante tener en cuenta los outliers porque son un elemento fundamental en el análisis exploratorio de los datos.

Ahora bien, cuando se tienen datos categóricos, la opción es visualizar un gráfico de barras, en donde se puede mostrar por ejemplo el conteo de ocurrencias en diferentes categorías o el porcentaje que estas categorías representan en un total de datos.

Con lo anterior mencionado, es importante indicar que el análisis univariado corresponde a analizar y visualizar una sola variable, pero también se puede empezar a observar interacciones y posibles relaciones entre dos o más variables lo que se conoce como el análisis bivariado y multivariado en donde se analiza si existe alguna tendencia lineal, es decir si el aumento de una variable genera el aumento o disminución de la otra. O también es posible calcular el índice de correlación de estas dos variables, donde un valor cercano a 1 indica una relación lineal, un valor cercano a -1 indica una relación lineal inversa, y un valor cercano a 0 indica que no hay una correlación linear entre los datos.

También se puede comparar una variable numérica con una variable categórica y usar gráficos de barra para la visualización o usar una gráfica de violín qué es similar a un boxplot, pero tiene una ventaja y es que además de mostrar la mediana y los límites de los cuartiles, incluye una gráfica de densidad de la distribución qué representa una gráfica continua del histograma.

Nota: también se puede comparar dos variables categóricas usando gráficos de barras apiladas.

Por otra parte, en el análisis multivariado se compara simultáneamente todos los posibles pares de variables para encontrar algún tipo de relación. Para cada comparación se calcula el índice de correlación entre diferentes pares de variables y se dibujan los resultados en una matriz de correlación. En donde, en la diagonal principal de dicha matriz se poseen valores iguales a 1 debido a que se está comparando una variable consigo misma y lo que se encuentra por fuera de esta línea es donde se debe centrar la atención, ya que es lo que se debe analizar.

El último paso de este EDA es la sumarización donde se extraen las conclusiones más importantes de todo el análisis realizado para identificar las variables o características que están correlacionadas y definir cuales son más relevantes

**CUARTA ETAPA: Modelado**

El modelado consta de las siguientes subetapas:

- Entrenamiento, validación y pruebas
- Selección de modelos
- Modelo elegido previo al despliegue

## Entrenamiento, validación y pruebas

Este proceso tiene como insumo el conjunto de datos procesado, y dado dicho conjunto de datos, éste a su vez se va dividir entre tres conjuntos: Conjunto de entrenamiento, conjunto de validación y conjunto de pruebas.

Usualmente el conjunto de entrenamiento es el conjunto más grande, oscila entre un 60% a un 80% del conjunto de datos de entrada. Dicho esto, es esencial recolectar los datos de entrenamiento necesarios para obtener información y establecer los parámetros del modelo a fin de optimizar el proceso de entrenamiento. Con lo anterior citado es importante mencionar que las configuraciones que se pueden ajustar para controlar el comportamiento del algoritmo de machine learning y la arquitectura de modelo resultante se conocen como hiper parámetros. La cantidad y el tipo de hiper parámetros en los algoritmos de aprendizaje automático son específicos de cada modelo. Algunos ejemplos de hiper parámetros utilizados habitualmente son: la tasa de aprendizaje, las capas ocultas, las unidades ocultas y las funciones de activación. El ajuste o la optimización de hiper parámetros es el proceso de selección de la arquitectura de modelo óptima.

Otra parte del conjunto de datos será para la validación, la cual ayuda a evaluar o validar los hiper parámetros en lugar de ajustarlos como se hizo en el entrenamiento. Por último se tiene el conjunto de datos de prueba el cual ayuda a evaluar la generalización del modelo de machine learning. Se necesita evaluar la eficiencia con la que este modelo realiza las predicciones cuando se tienen datos nuevos que son desconocidos y no vistos por el modelo anteriormente. Este paso es importante ya que se debe validar para determinar el desempeño y la precisión del modelo. Se pueden generar varios modelos mediante distintos métodos y evaluar cada uno de ellos. Por lo general, en un algoritmo de machine learning se calculan varias métricas, como los errores en el entrenamiento, sobreajuste, la precisión de la predicción, entre otras. Estas métricas ayudan a determinar si el modelo aprende correctamente y si se generaliza de forma adecuada para realizar predicciones sobre los datos desconocidos. Las métricas informadas por el algoritmo dependen del problema empresarial y de la técnica de aprendizaje automático que se utilizó. Por ejemplo, un algoritmo de clasificación se puede medir con una matriz de confusión que detecte verdaderos o falsos positivos y verdaderos o falsos negativos, mientras que un algoritmo de regresión se puede medir a través de la desviación cuadrática media (RMSE).

## Selección de modelos

Previamente se han limpiado, preparado y se han entendido los datos. Por ende, en esta etapa se tienen identificados uno o varios modelos que están funcionando bien con los datos, por esa razón esta subetapa ha sido titulada como selección de modelos.

Es importante mencionar que existen variedad de criterios para realizar esta selección correctamente, entre ellos los más utilizados son:
El criterio de información de Akaike y el factor Bayes y/o el criterio de información Bayesiano(que en cierta medida se aproxima al factor Bayes) [45].

Para la selección de modelos es preciso responder a las siguientes preguntas:

- ¿Qué algoritmo de Machine Learning se debe usar? El algoritmo que se selecciona depende principalmente de dos aspectos diferentes del escenario de ciencia de datos:

    - ¿Qué se quiere hacer con los datos? Es decir, ¿A qué problema se debe dar respuesta aprendiendo de los datos pasados?
    - ¿Cuáles son los requisitos del escenario de ciencia de datos? En concreto, ¿Cuál es la linealidad, la precisión, el tiempo de entrenamiento, el número de parámetros? [46]

**Modelo elegido previo al despliegue**

En esta subetapa se tiene el modelo que se ha decidido desplegar. Este modelo debe cumplir con todos los requerimientos dados por el cliente, y haber ejecutado todas las etapas correctamente. Además de ello, en la subetapa anterior titulada selección de los modelos, entre todos los modelos existentes, este modelo elegido es el que más acorde esté con las necesidades del cliente y por ende es posible y conveniente desplegarlo.

**QUINTA ETAPA: Despliegue**

El despliegue comprende 6 subetapas que son:

- Diseño e implementación en la nube
- Evaluación con datos reales y pruebas en la nube
- Despliegue/Modelo en producción
- Registro de una nueva versión del modelo en producción
- Monitorización/ Retroalimentación de métricas de desempeño del modelo
- Recepción de nuevos requerimientos o funcionalidades

**Diseño e implementación de la solución en la nube**

Esta subetapa es de suma importancia porque es donde se toman las decisiones sobre la arquitectura del modelo. Es decir, cual es la mejor forma para usar el modelo predictivo. Así que, por lo general, en los sistemas de predicción basados en el uso de algoritmos de Machine Learning se generan las preguntas: ¿Cuál es la forma en la que se va a usar esta predicción? ¿Quiénes son los clientes que van a

consumir este sistema?, ¿Cómo se espera que dicho cliente reciba la información?, entre otras. Es decir, cuestionarse el cómo se consumirá el modelo y de qué forma se hace más eficiente. Un ejemplo más específico de preguntas podría ser: ¿El cliente solo recibirá un reporte en un documento de excel el cual le muestra la predicción, o el cliente podrá verlo a través de una página web?. Si se habla de una clasificación, ¿Se debe clasificar un día o todas las semanas?. Se deben generar las preguntas necesarias para entender y suplir los requerimientos del cliente.

Es importante resaltar que el paso de diseñar la solución generalmente se realiza en esta etapa final pero puede suceder que en algunos casos se realiza desde la primera etapa de comprensión del negocio. En realidad, depende de la forma en que se haya manejado el proyecto, muchas veces no se diseña e implementa nada en la nube antes de estar seguros que hay un modelo que cumple con la tarea que se necesita y pueda suplir las necesidades del cliente. Porque algunas veces los datos no son suficientes, o la calidad de los datos no es buena, por ende el modelo no es tan bueno como se esperaba entonces el cliente es el que decide si se requiere el modelo en la nube o no. Por esa razón, esta sub-etapa se encuentra en esta etapa de despliegue.

**Evaluación con datos reales y pruebas en la nube**

En este paso se ha seleccionado y elegido el modelo más adecuado para realizar pruebas en la nube con datos reales y de esta manera determinar si su rendimiento y precisión permitirán cumplir los objetivos del negocio. Existen varias formas de realizar esta evaluación y prueba en la nube, uno de los más conocidos es gitflow workflow, pero existen variedad de herramientas para realizar este proceso y así poder tener las diferentes versiones del modelo en un ambiente controlado en la nube, y con ello no generar ningún impacto a los usuarios. Después de hacer estas pruebas y estar completamente seguros de que se tiene una versión estable y eficiente del modelo entonces ya se despliega, es decir se saca a producción.

**Despliegue/Modelo en producción**

El despliegue de un modelo de ML consiste en llevarlo de la etapa de desarrollo a la de producción para que esté en disposición de un usuario final. Este aspecto ha cobrado mucha importancia en los últimos años para el desarrollo de aplicaciones a nivel industrial.

Debido a que MLOps representa un conjunto de prácticas que busca lograr el despliegue y mantenimiento de modelos de ML de manera confiable y eficiente. El despliegue es una parte central de este proceso. Consiste en tomar el modelo entrenado y hacerlo accesible a un usuario final.

La elección del tipo de despliegue dependerá del uso del modelo y de varios requerimientos de diseño que son:

- **Tipo de predicción a realizar:** Puede ser en tiempo real o por lotes.

    - En tiempo real las predicciones son realizadas y devueltas al usuario en el menor tiempo posible después de recibir la solicitud.
    - Cuando se predice por lotes, el objetivo es procesar una gran cantidad de información de entrada y el modelo genera las predicciones de forma asíncrona, es decir esta predicción no es inmediata

- **Latencia:** Tiempo de respuesta requerido desde que se envía la solicitud al modelo en producción hasta que se recibe la predicción. La idea es que sea lo más pequeña posible.

- **Rendimiento:** Número de solicitudes por segundo que puede soportar el sistema donde está alojado el modelo.

- **Complejidad del modelo:** Se refiere al tamaño del modelo de machine learning. Es decir, que tanta capacidad de cómputo y almacenamiento requiere el modelo. Cuando el modelo es muy complejo se requieren muchos recursos computacionales.

Para esto, es importante tener en cuenta las alternativas de despliegue que por lo general se dividen en 2 grandes grupos que son:

- **En la nube:** Se refiere a que el cómputo requerido para las predicciones se realizan en servidores remotos alojados en la nube, por lo cual los datos y las predicciones se transfieren a través de internet. Se recomienda utilizar esta alternativa cuando los modelos son complejos o cuando la latencia no es un problema porque podemos aceptar retardos en el envío de los datos y las recepción de las predicciones. O también porque las predicciones se hacen por lotes.

    Al desplegar en la nube se puede acceder al modelo de dos maneras:

    - La primera es almacenando los datos de entrada al modelo en una base de datos y programando el sistema para que cada cierto tiempo tome esos datos y logre generar las predicciones. Dichas predicciones resultantes también son almacenadas en una base de datos y periódicamente son entregadas a una aplicación cliente que será encargada de llevarlas al usuario final.

    - La segunda forma es empaquetando el modelo en una API que le permitirá recibir solicitudes hechas por el usuario final así como entregar resultados de la predicción. Esta forma es usada cuando se requiere hacer predicciones de baja latencia o cuando no es por lotes.

- **On the edge:** Quiere decir que el modelo no va a estar alojado en la nube sino en el mismo dispositivo encargado de recibir las solicitudes. Sin embargo, este dispositivo tiene la capacidad de cómputo muy limitada como por ejemplo: un teléfono móvil, una tablet o un smartwatch. Esta alternativa se usa cuando no hay modelos muy complejos, cuando se requieren bajos niveles de latencia o si por cuestiones de seguridad no resulta conveniente enviar información a través de internet. Es importante mencionar que en este despliegue la predicción sólo puede ser en tiempo real, debido a que en dispositivos móviles los recursos son muy limitados para ser desplegado por lotes.

**Registro de una nueva versión del modelo en producción (Registro de modelos y control de versiones del modelo)**

La razón común para la actualización del modelo de aprendizaje automático es la "decadencia del modelo", donde el rendimiento del modelo disminuye con el tiempo a medida que llegan nuevos datos.

Es necesario garantizar la compatibilidad con versiones anteriores revirtiendo los modelos creados anteriormente. Además, mediante el seguimiento de varias versiones del modelo de machine learning, es posible implementar diferentes estrategias de implementación.vgy7

Con el control de versiones de modelos se puede realizar un seguimiento del historial de modelado y de las puntuaciones asociadas con los resultados de ejecución de los modelos a medida que proporciona más o diferentes conjuntos de datos.

**Monitorización/ Retroalimentación de métricas de desempeño del modelo / Recepción de nuevos requerimientos o funcionalidades**

Una vez se llega a la etapa del despliegue, empieza un proceso continuo de monitorización. En donde es esencial vigilar que las métricas de desempeño del modelo estén funcionando correctamente y eso implica que el cliente constantemente provea retroalimentaciones de cómo se encuentra funcionando el sistema de machine learning en la tarea específica que se encuentre realizando.

En esta subetapa también se toma en consideración la posibilidad de generar nuevas funcionalidades que se deban probar para incluir en el sistema. Por

consiguiente se debe recibir retroalimentación ante cualquier cambio. Ya sea un cambio de reentrenamiento porque las métricas de desempeño bajaron de un umbral establecido o ya sea por la aparición de nuevas funcionalidades solicitadas por el cliente. En este paso es importante realizar un registro de control de versiones.

Ahora bien, se han generado las siguientes preguntas para poder vigilar el desempeño del modelo de machine learning:

- ¿Cómo se entregan las predicciones? ¿Qué tanto nivel de gestión me interesa tener? (batch or online)
- ¿Cuál es el tiempo esperado para los cambios? (Tiempo desde el compromiso hasta la producción)
- ¿Qué tipo de entorno consumirá los resultados del modelo (predicciones o clasificaciones)?
- ¿Cuáles son las métricas de desempeño del modelo?
- ¿Cuál es su política de liberación de modelos? Is A/B testing or multi-armed bandits testing required? (p. ej., para medir la efectividad del nuevo modelo en métricas comerciales y decidir qué modelo debe promoverse en el entorno de producción)
- ¿Cuál es su estrategia de implementación aplicadas a machine learning? (Como montar el modelo) (e.g. shadow/canary deployment required?)

**Batch inference:** Se usa este proceso donde la latencia es mayor, donde no se necesita la predicción inmediata en nuestro modelo de inteligencia artificial.
Al ser un proceso asíncrono que no depende del tráfico de usuarios, no hay que tener en cuenta el escalado de solución. y si hay un fallo no hay que actuar de manera inmediata ya que no hay clientes que se vean afectados.

Nota: Las Tecnologías para hacer inferencia en batch son:

- Databricks jobs: Se genera un notebook, o un fichero JAR y se configura para que se ejecute cada determinado tiempo
- MLFlorw Serving Se hace la inferencia con Apache Spark, se hace con un single node.
- Spark ML: Se tiene un Job de inferencia en donde usando Spark se leen los datos y se hace la inferencia y se configura cada x tiempo, para que se ejecute el proceso.
- Azure ML pipeline: Permite tener un pipeline con los distintos procesos que conlleva azure machine learning y que se puede configurar para hacerlo a manera de batch utilizando el parallel Run Step

**Online inference.** ¿Cuándo se requiere la inferencia en línea?

En general, se requiere la inferencia en línea siempre que se necesiten predicciones sincrónicamente.

Nota: El modelo debe desvincularse de las aplicaciones móviles y web. Desvincular el modelo de las aplicaciones permite a los científicos de datos actualizar el modelo, retroceder a versiones anteriores y operar varias estrategias de implementación mucho más fácilmente.

**Recomendación de amazon**

SageMaker Feature Store provee baja latencia (hasta menos de 10 milisegundos) y alto throughput de lectura para servir el modelo sobre nuevos datos que van llegando.

**Edge inference:** Ayuda en casos donde no hay mucha conectividad, aprovechar los dispositivos IOT o los dispositivos móviles.

Las ventajas es que se puede escalar la solución, y el escalado se hace desplegando ese modelo en distintos dispositivos, pero esto conlleva nuevos riesgos y nuevos retos que se deben afrontar, por ejemplo, Si se desea desarrollar modelos que sean capaces de correr en dispositivos en donde no se tiene la capacidad como se tiene en el cluster en la nube, este proceso se va a complicar porque depende de los distintos hardware en donde se quiera desplegar el modelo, así que se debe tener en cuenta la arquitectura del modelo. se deben tener procesos muy buenos, para aprovechar las ventajas

Para decidir qué método implementar, se deben tener en cuenta los siguientes factores:

1. **Latencia:** ¿Qué tan rápido requiere una aplicación/usuario los resultados de la predicción del modelo?
2. **Privacidad de datos:** ¿Hay algún problema/preocupación sobre el envío de datos a un back-end?
3. **Conectividad de red:** algunas opciones de implementación requieren acceso a Internet/red. Si el entorno en el que se debe implementar el modelo tiene conectividad de red/internet limitada o nula, las opciones son limitadas.
4. **Costo:** Ciertas opciones de implementación serán más costosas que otras. Piensa en tener un servidor en línea las 24 horas del día, los 7 días de la semana para servir predicciones. ¿Cuál será el costo de operar y mantener este servidor?

# Capítulo 5

# Validación

## 5.1 Herramientas de validación

## 5.1.1 Validación por medio de cuestionario

Para la validación del modelo de referencia se elaboró un cuestionario utilizando la herramienta google forms. Este cuestionario fue validado con 6 proyectos diferentes en los cuales hay 4 ingenieros involucrados y un estudiante de la facultad de ingeniería electrónica y telecomunicaciones de la Universidad del Cauca que se encuentra realizando su práctica profesional en la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S.

El cuestionario plantea una serie de preguntas basadas en las 5 etapas del modelo de referencia. Las preguntas más relevantes de cada etapa son las siguientes:

**Etapa 1:COMPRENSIÓN DEL NEGOCIO**

- ¿Se tiene acceso a un conjunto de datos para resolver este problema?
- ¿Se comprende completamente el requerimiento o necesidad del cliente/empresa y el objetivo del proyecto?

En el caso de que la ruta sea un modelo de ML o un dashboard:

- ¿Se requiere implementar el modelo de ML o el dashboard en servicios de la nube? Si no es así, explique el porqué.
- ¿Se tiene conocimiento de los criterios de éxito técnico y de negocio?

- Ya se ha calculado y determinado el costo de:
- Los componentes físicos que se utilizan para el desarrollo del proyecto
- Almacenamiento y procesamiento de datos
- Recursos humanos
- Herramientas de la nube
- Herramientas de AWS

- ¿Se ha establecido si este negocio es rentable para los posibles consumidores del producto o servicio ofrecido?
- ¿Se ha iniciado con el diseño de una pipeline del proyecto?

**Etapa 2:PREPARACIÓN DE DATOS**

- ¿Se ha analizado si los datos entregados por el cliente son adecuados y útiles para el proyecto? Y si no es así, ¿Se ha realizado el tratamiento debido de los mismos?
- ¿Se han extraído las características relevantes en el conjunto de datos para el problema en cuestión?
- Con respecto a la automatización de implementación y prueba: ¿Se ha establecido la canalización de CI/CD para el código base?

**Etapa 3:ANÁLISIS DE DATOS**

- ¿Se realizó la descripción estadística?
- ¿Se realizó el análisis necesario para identificar si los datos se encuentran dispersos o agrupados?
- ¿Se realizó el análisis respectivo de los percentiles 75, 25, el min y el max?
- ¿Considera usted que se realizó el análisis adecuado para extraer las conclusiones más importantes?
- ¿Se han planteado las hipótesis necesarias para comprobar si las variables en el estudio son útiles para hacer la proyección?
- ¿Se ha evaluado el algoritmo?

**Etapa 4:MODELADO**

- ¿Se ha seleccionado un algoritmo de aprendizaje automático que sea adecuado para el problema a resolver?
- ¿El modelo se encuentra entrenado?

- Las métricas dadas por el algoritmo dependen del problema empresarial y de la técnica de aprendizaje automático que se utilizó. Dicho esto, seleccione la opción más acorde a su modelo de machine learning:
- Clasificación binaria
- Clasificación multiclase
- Regresión
- Clustering o agrupamiento
- Clasificación
- Detección de anomalías

- ¿Se ha evaluado correctamente el modelo?

- ¿Se ha prevenido el sobreajuste?
- ¿Se realizaron las pruebas necesarias para corroborar que el modelo está funcionando correctamente?

**Etapa 5: DEPLOYMENT**

- ¿Se seleccionó el modelo que se va a utilizar para el desarrollo y ejecución del proyecto?
- Describa cómo se evaluó el modelo y las herramientas que se usaron para este proceso, si no fue ese el caso, describa porque no fue necesario hacer este proceso.
- ¿Se tiene el modelo en producción? Es decir que el modelo ya se encuentra en la nube cumpliendo con las funciones requeridas y el cliente puede usarlo.
- ¿El modelo es reproducible? Una de las principales quejas sobre los proyectos de aprendizaje automático es la falta de reproducibilidad. Es importante que el método y los resultados de la fase de modelado sean reproducibles.
- Especifique: Cuál es su estrategia de implementación y porqué se eligió.
- Los datos del mundo real puede que difieran de los datos utilizados para entrenar el modelo, lo que puede generar divergencias en la calidad de los modelos y, eventualmente, generar modelos menos precisos. ¿Se están usando para monitorear el modelo y evitar que la calidad del modelo se deteriore?

## 5.1.2 Validación por medio de puntuación

Además de ello, se ha creado un excel el cual evalúa qué tan avanzado se encuentra el proyecto por medio de una puntuación, que se maneja de la siguiente manera:

Al ser definidas 3 rutas, cada ruta posee su cantidad de puntos específica que son:

- En un dashboard la cantidad de puntos máxima es de 55
- En un reporte analitico la cantidad de puntos máxima es de 76
- En un modelo de Machine Learning la cantidad de puntos máxima es de 115 - 122 dependiendo de la elección de métrica de evaluación elegida.

| Correo electrónico | Puntuación / 140 | Publicación de la puntuación |
|---|---|---|
| sebastian.rojas@thebitbang.company | 91 | Sin publicar |
| daniel6697@unicauca.edu.co | 81 | Sin publicar |
| emmanuel.lasso@thebitbang.company | 59 | Sin publicar |
| emmanuel.lasso@thebitbang.company (1) | 50 | Sin publicar |
| julian.plazas@thebitbang.company | 17 | Sin publicar |
| edwargiron73@gmail.com | 71 | Sin publicar |

Figura 13. Puntuación de los 6 proyectos [fuente propia]

La puntuación obtenida de los 6 proyectos mencionados se puede observar en la figura 13. En donde es importante resaltar que la puntuación más baja, es el resultado de un dashboard, por ende el más alto puntaje que este proyecto pudo obtener es de 55. Y el proyecto de más alta puntuación consiste en un modelo de Machine learning, en donde el puntaje más alto que pudiese haber obtenido es entre 115 y 122 dependiendo de la métrica en la que se haya trabajado el proyecto.


Figura 14. Distribución de las puntuaciones totales [fuente propia]

La figura 14 muestra la distribución del puntaje total de los 6 proyectos, en donde se presenta un valor medio de 59 puntos obtenidos de los 140 posibles puntos, sin embargo es importante resaltar que la máxima puntuación obtenida es entre 115 y 122 puntos, no obstante la herramienta google forms señala que la puntuación máxima es 140 puntos porque no difiere las diferentes rutas que se pueden tomar cuando se toma en consideración la evaluación de las métricas, por tanto dicha herramienta realiza la sumatoria de todas las métricas y por esa razón en la ilustración estadística mostrada en la figura 14 muestra que la cantidad de puntos máxima es 140, a pesar de que la puntuación máxima se encuentra en un rango de 115 a 122 puntos como se puede observar en la figura 15.

| MACHINE LEARNING | | | | | |
|---|---|---|---|---|---|
| Métricas de evaluación: | | | | | |
| Clasificación binaria | Clasificación multiclase | Regresión | Agrupación en clústeres | Clasificación | Detección de anomalías |
| 113 | 113 | 113 | 113 | 113 | 113 |
| 5 | 4 | 4 | 3 | 2 | 9 |
| 118 | 117 | 117 | 116 | 115 | 122 |

Figura 15. Cantidad máxima de puntos de un proyecto de machine learning en base a la métrica elegida [fuente propia]

Con esto dicho, es esencial aclarar que los proyectos que se validan por medio de este cuestionario y la herramienta de evaluación por puntos deben poseer un porcentaje final completado que oscile entre un 75%  y un 100%, este porcentaje corresponde al porcentaje final que un proyecto debe tener para verificar que dicho proyecto cumple efectivamente con las recomendaciones propuestas en este modelo de referencia.

Es importante mencionar que las estadísticas muestran que los puntos se encuentran en un intervalo de 17 a 91 puntos, lo que quiere decir que entre las encuestas realizadas a los 6 proyectos hay un promedio de 54 puntos. Con esto dicho, se sabe entonces que el 50% de la puntuación total corresponde a [115/2 - 122/2] esto es entre [57.2 - 61] puntos, y que el valor promedio no sobrepasa esta puntuación y se encuentra entre un 44.26% - 47.20% por ende, se decidió realizar un análisis detallado de cada uno de los proyectos con el fin de revisar cuales son las recomendaciones que más se tienen en cuenta de este modelo de referencia, y cuales no se toman en consideración o no se consideran cruciales al momento de realizar un proyecto en la empresa WIZIT MIND BLOWING SOLUTIONS.

## 5.2 Análisis detallado de los 6 proyectos evaluados

**Proyecto 1: Proyecto realizado por el ingeniero Juan Sebastián Rojas**

Este proyecto se realizó para la empresa-cliente 1. Es un proyecto que fue desarrollado para una empresa de Medellín. El objetivo era desarrollar un modelo de

predicción de horas estimadas en historias de usuario para desarrolladores de software.

El problema era que este proyecto diseñaba y estimaba matemáticamente las horas que debía tomarle a un desarrollador el realizar una historia de usuario (una serie de tareas según SCRUM). Esa estimación la realizaban mediante un archivo de excel y no siempre era muy preciso y estaba sujeto a imprecisiones introducidas por el error humano (mala estimación de tareas - sobreestimación o subestimación de las horas que puede tomar hacer algo en software). Entonces se generó un modelo de predicción de esas horas estimadas a partir de los datos ingresados por el project manager que fue un árbol de decisión y el modelo fue embebido en los formatos de excel que los integrantes de la empresa-cliente 1 ya usaban para que fuera una transición sencilla.

Este proyecto tiene una puntuación de 91 puntos y la ruta elegida por el ingeniero fue un modelo de ML como se puede observar en la figura 16.

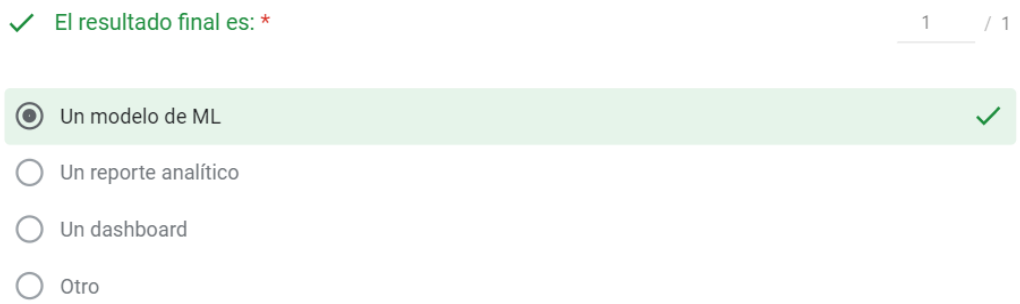


Figura 16. Ruta elegida en el cuestionario de validación [fuente propia]

lo cual ubica la evaluación de puntos en el en un rango máximo de 115 a 122 puntos dependiendo de la métrica de evaluación elegida en el modelado. La métrica elegida por el ingeniero fue regresión como se puede observar en la figura 17.

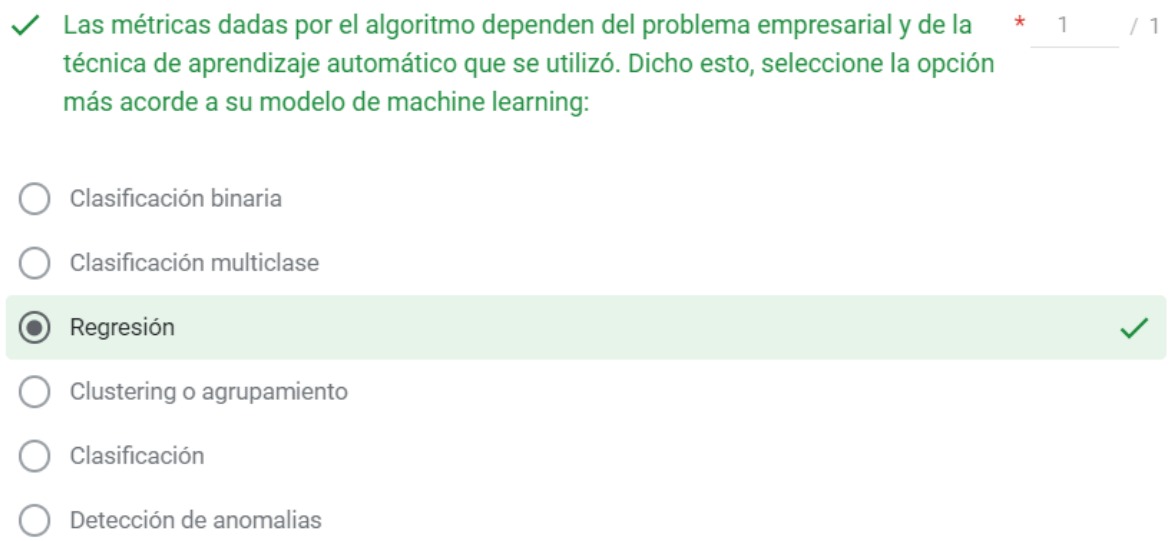


Figura 17. Algoritmo de machine learning [fuente propia]

Esto quiere decir que, la máxima cantidad de puntos que se pudo obtener en este proyecto son 117 como se puede apreciar en la figura 15 mostrada anteriormente.

A simple vista, este proyecto culminó alrededor del 91/117 = 77,77%.

Sin embargo, es importante resaltar que algunas preguntas se deben evaluar manualmente, por tanto es necesario realizar un barrido de las respuestas que se tienen en la herramienta google forms. Al hacer este barrido se observa que:

- La pregunta ilustrada en la figura 18, debe verificarse manualmente debido a que la persona o personas que estén respondiendo el cuestionario pueden elegir una o varias respuestas, y de igual manera la respuesta será correcta debido a que el objetivo principal de esta pregunta es conocer si existen fuentes de datos de entrenamiento disponibles y de qué tipo son estas fuentes de datos (propios, pagados o públicos). Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un total de 92 puntos.

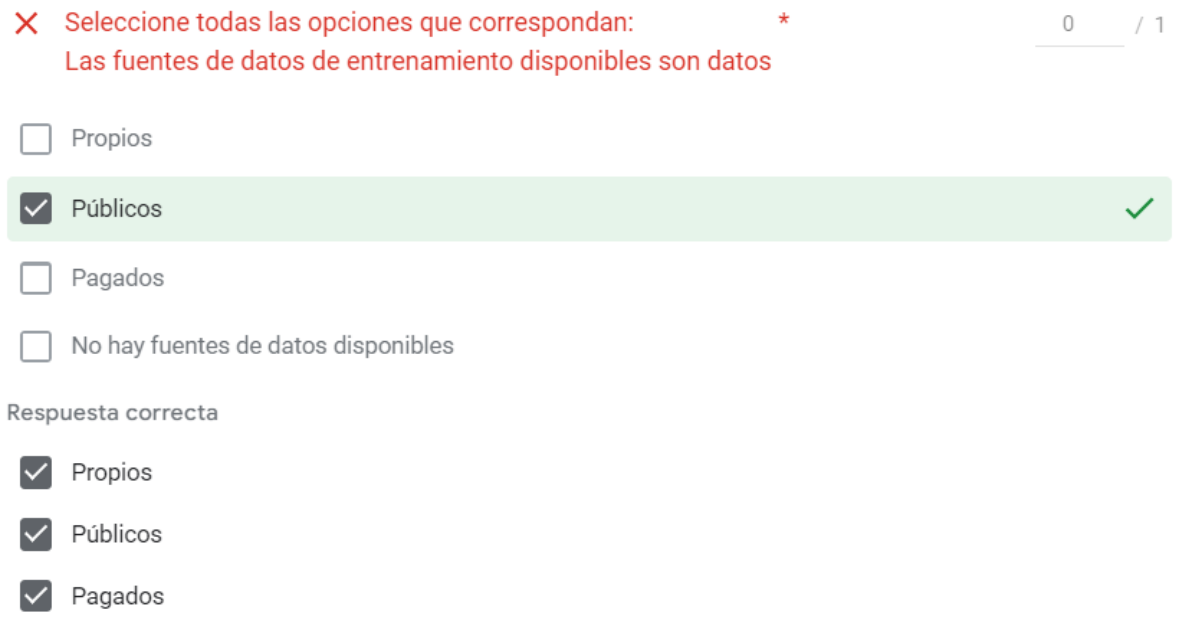


Figura 18. Fuentes de datos de entrenamiento disponibles [fuente propia]

- Un caso similar sucede con la pregunta que se ilustra en la figura 19, debido a que las preguntas de tipo selección múltiple, solamente aparecen correctas con la herramienta cuando se ha hecho la selección de todas las opciones posibles, a pesar de que no es necesario que se seleccionen todas las respuestas para que la pregunta se considere correcta. En este caso, el objetivo es informativo, para conocer las variables con las que se trabajó en el proyecto. En donde las opciones son valores enteros, continuos, binarios, nominales y ordinales, y la única variable que no fue usada por el ingeniero son las variables ordinales. Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un puntaje total de 93 puntos. Sin embargo, es importante mencionar que aunque se hubiese usado un solo tipo de variables y no el resto de las variables presentadas, la respuesta se considera correcta y se debe sumar el punto extra.

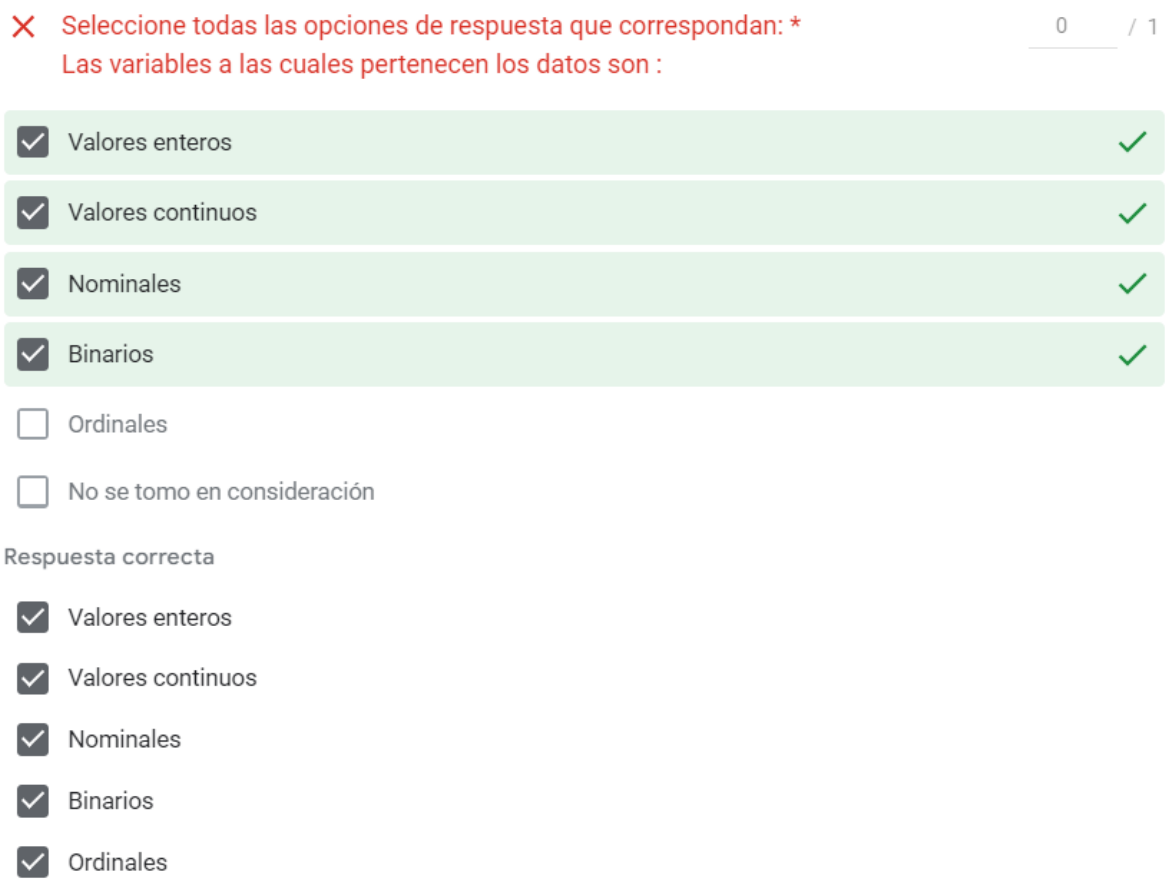


Figura 19. Variables a las cuales pertenecen los datos [fuente propia]

Con el análisis realizado se concluye que el proyecto logró un total de 93 puntos sobre 117 es decir que el porcentaje completado de este proyecto es 79.48%. Esto quiere decir que el proyecto sigue las recomendaciones del modelo propuesto el cual oscila entre un porcentaje de 75% a 100%.

**Proyecto 2: Proyecto realizado por el estudiante Daniel Santiago Vásquez**

Este proyecto consiste en implementar un modelo de Machine Learning siguiendo los lineamientos de MLOps definidos por la empresa que permita evaluar el perfil de riesgo de una persona, sin historial de crédito, para recibir un producto o servicio financiero.

Este proyecto tiene una puntuación de 81 puntos y la ruta elegida por el estudiante de trabajo de grado fue un modelo de ML como se puede comprobar en la figura 20.

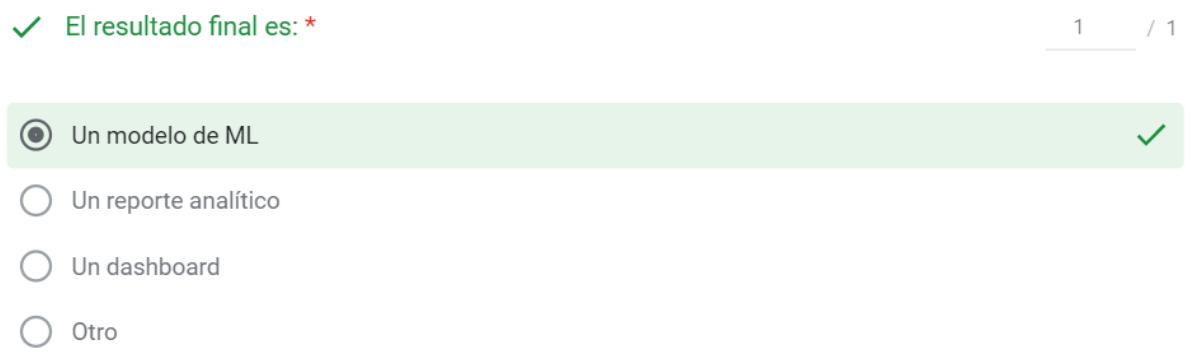


Figura 20. Ruta elegida en el cuestionario de validación [fuente propia]

Lo cual ubica la evaluación de puntos en el en un rango máximo de 115 a 122 puntos dependiendo de la métrica de evaluación elegida en el modelado. La métrica elegida por el estudiante fue clasificación binaria como se puede observar en la figura 21.

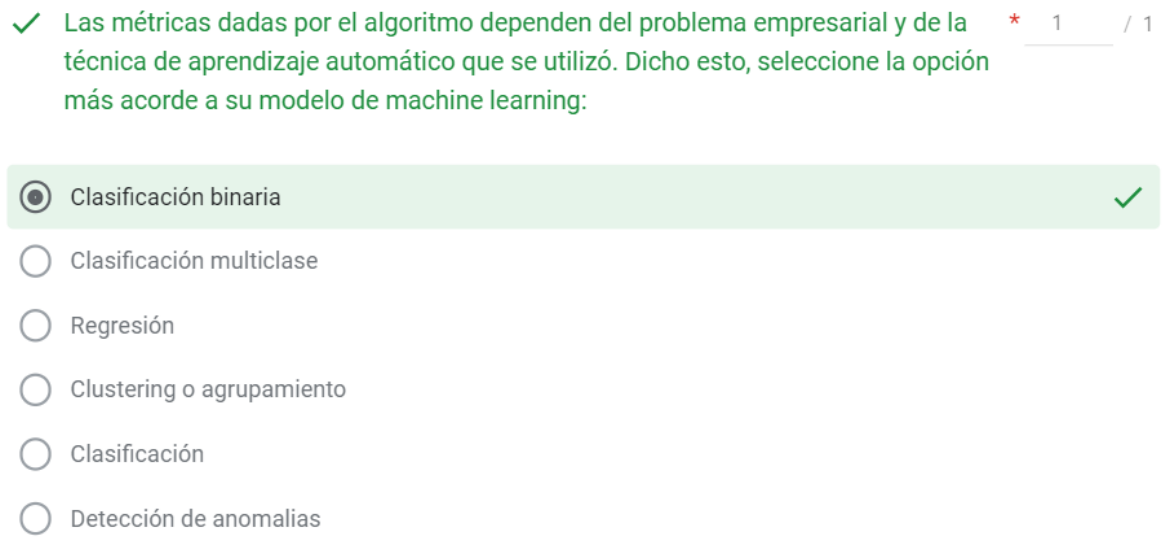


Figura 21. Algoritmo de machine learning [fuente propia]

Esto quiere decir que, la máxima cantidad de puntos que se pudo obtener en este proyecto son 118 como se puede apreciar en la figura 15.

Ahora bien, según la herramienta de google forms este proyecto culminó alrededor del 81/118 = 68,64%.

Sin embargo, al igual que en el proyecto anterior, es válido destacar que es necesario evaluar algunas preguntas manualmente, por tanto es preciso realizar un barrido de las respuestas que se tienen en la herramienta google forms. Al hacer este barrido se observa que:

- La respuesta a la pregunta ilustrada en la figura 22, debe examinarse manualmente ya que la persona o personas que estén respondiendo el cuestionario pueden elegir una o varias respuestas, y de igual manera la respuesta será correcta debido a que el objetivo principal de esta pregunta es conocer si existen fuentes de datos de entrenamiento disponibles y de qué tipo son estas fuentes de datos (propios, pagados o públicos). Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un total de 82 puntos.

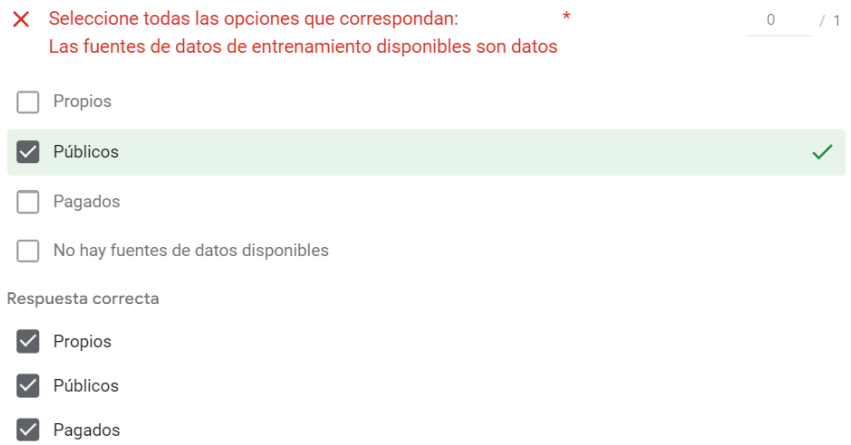


Figura 22. Fuentes de datos de entrenamiento disponibles [fuente propia]

- Un caso muy parecido ocurre con la pregunta que se ilustra en la figura 23, debido a que las preguntas de tipo selección múltiple, solamente se consideran correctas en la herramienta cuando se ha hecho la selección de todas las opciones posibles, a pesar de que no es necesario que se seleccionen todas las respuestas para que la pregunta se considere correcta. En este caso, el objetivo es informativo, para conocer las variables con las que se trabajó en el proyecto. En donde las opciones son valores enteros, continuos, binarios, nominales y ordinales y el estudiante seleccionó 3 de las 5 opciones correctas. Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un puntaje total de 83 puntos. Sin embargo, es importante mencionar que aunque se hubiese usado un solo tipo de variables y no el resto de las variables presentadas, la respuesta se considera correcta y se debe sumar el punto extra correspondiente.

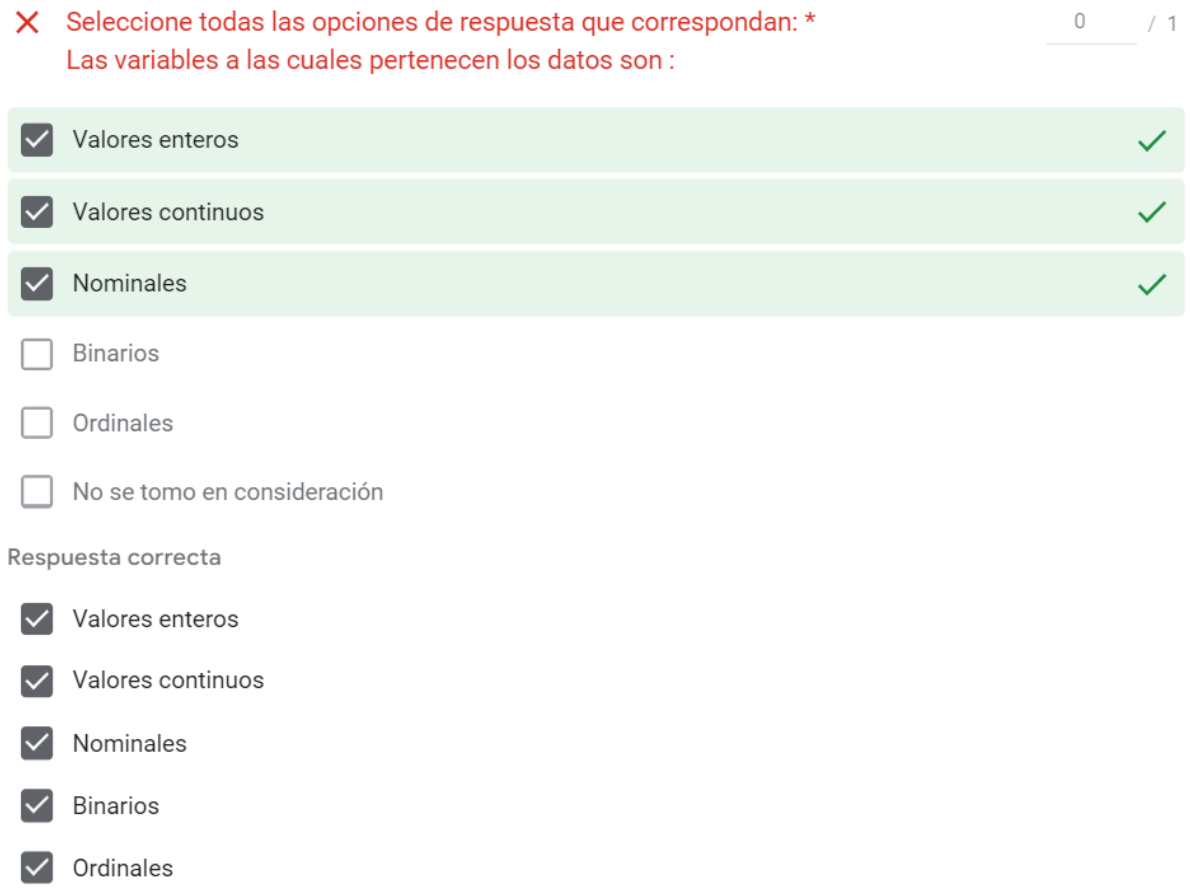


Figura 23. Variables a las cuales pertenecen los datos [fuente propia]

Con el análisis realizado es posible concluir que el proyecto del estudiante Daniel Vasquez logró un total de 83 puntos sobre 118 lo que corresponde a un porcentaje completado en este proyecto de 70.33%. Esto quiere decir que el proyecto no logra el puntaje recomendado por este modelo de referencia, el cual sugiere obtener una puntuación ubicada entre el 75% y 100%. Por ende es importante resaltar que es posible que existan características que podrían mejorarse del proyecto según las propuestas de este modelo de referencia.

**Proyecto 3: Proyecto realizado por el ingeniero Emmanuel Lasso**

Este proyecto se realizó para la empresa-cliente 3. Donde el objetivo es recopilar y analizar datos de interacciones de posts en redes sociales de las cuentas asociadas a gobernaciones. En este se analizaron las interacciones que contribuyen al engagement de cada post y se generaron modelos de datos que permitían verificar las variables que más impacto tienen en el engagement.

Este proyecto tiene una puntuación de 59 puntos y la ruta elegida por el ingeniero fue un reporte analitico como se puede observar en la figura 24:

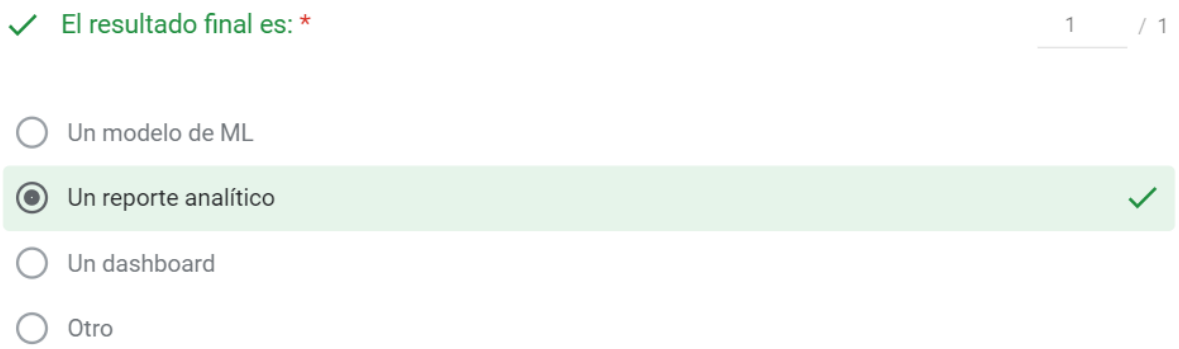


Figura 24. Ruta elegida en el cuestionario de validación [fuente propia]

Esto quiere decir que la máxima cantidad de puntos que se pudo obtener en este proyecto son 76 cómo se logra apreciar en la figura 15 mostrada anteriormente.

A simple vista, este proyecto culminó alrededor del 59/76x100 = 77,63%

Sin embargo, es importante subrayar que algunas preguntas se deben evaluar manualmente, por tanto es necesario realizar un barrido de las respuestas que se tienen en la herramienta google forms. Al hacer este barrido se observa que:

- La pregunta ilustrada en la figura 25, debe verificarse manualmente debido a que la persona o personas que estén respondiendo el cuestionario pueden elegir una o varias respuestas, y de igual manera la respuesta será correcta debido a que el objetivo principal de esta pregunta es conocer si existen fuentes de datos de entrenamiento disponibles y de qué tipo son estas fuentes de datos (propios, pagados o públicos). Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un total de 60 puntos.

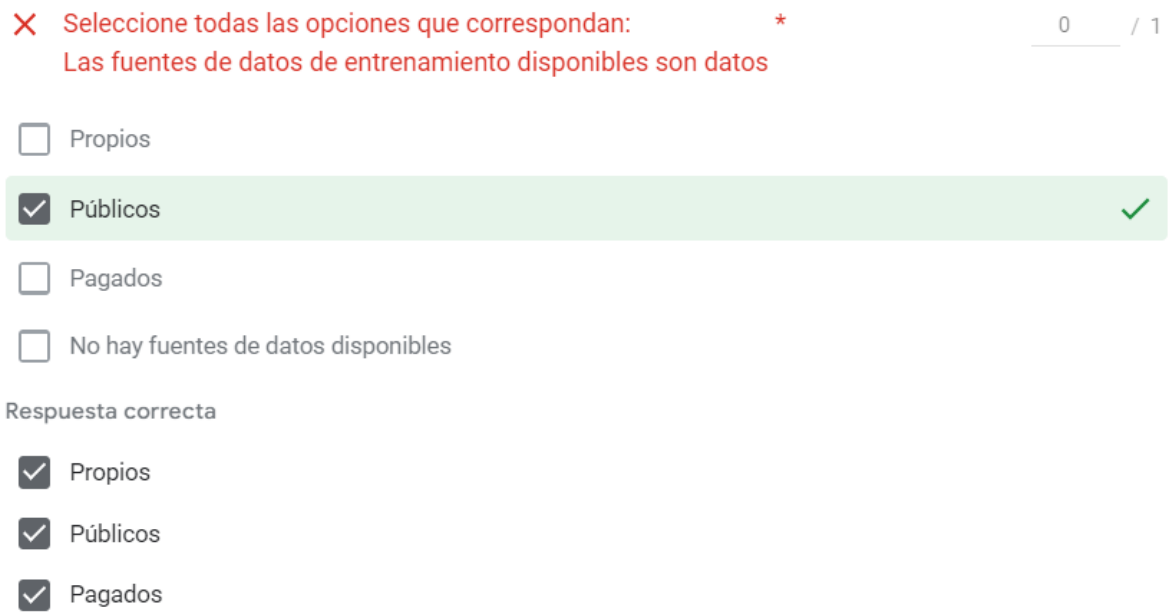


Figura 25. Fuentes de datos de entrenamiento disponibles [fuente propia]

- Existen preguntas en donde se agregan la cantidad de puntos manualmente dependiendo del número de respuestas seleccionadas a diferencia del caso anterior ilustrado en la figura 25. Debido a que las preguntas de tipo selección múltiple, solamente aparecen correctas con la herramienta cuando se ha hecho la selección de todas las opciones posibles, a pesar de que no es necesario que se seleccionen todas las respuestas para que la pregunta se considere correcta. Para el caso de la pregunta formulada en la figura 26, el objetivo es conocer el costo de recursos y herramientas importantes al momento de realizar un proyecto basado en ciencia de datos, porque es importante tener en cuenta todos los detalles para poder darle al cliente una propuesta de negocio racional y conveniente para ambas partes. Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un puntaje total de 61 puntos.

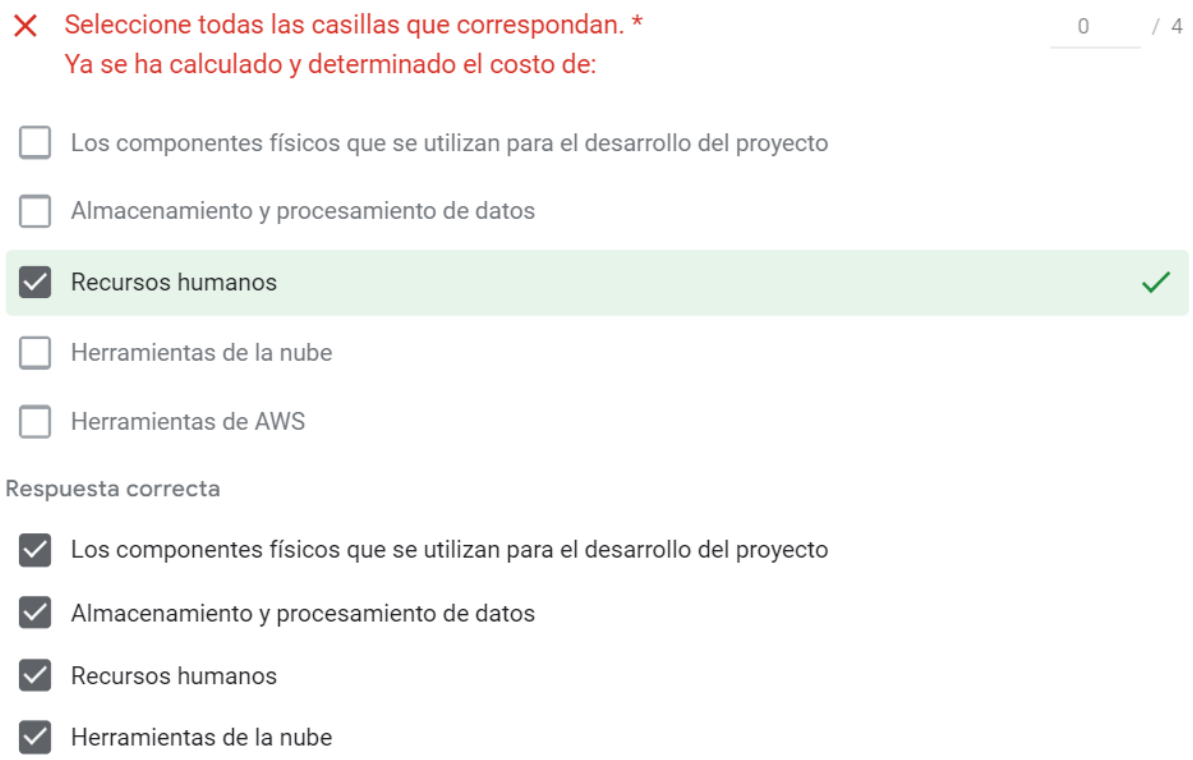


Figura 26. Selección múltiple con múltiples respuestas: Cálculo de costos [fuente propia]

Ahora bien, el mismo caso sucede con la figura 27 ya que los puntos aumentan con cada respuesta seleccionada correcta. Para este caso el ingeniero seleccionó dos de las seis opciones que se presentaban. Por ende se deben sumar dos puntos extra a la puntuación  final. Lo que corresponde a 63 puntos en total.

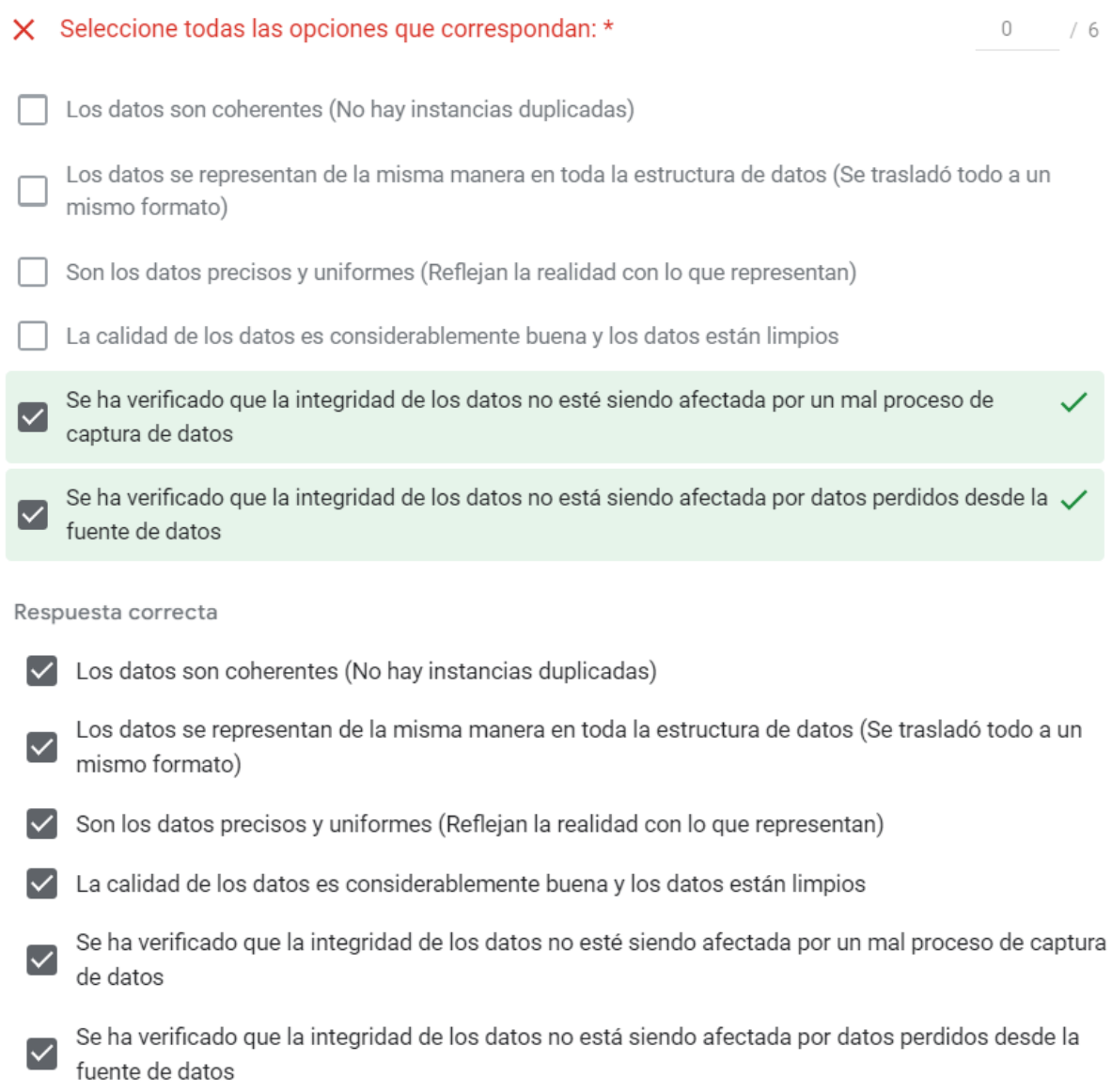


Figura 27. Selección múltiple con múltiples respuestas: Preparación de datos [fuente propia]

Con esto se puede afirmar que en el análisis realizado del proyecto de la empresa-cliente 3, se logró un total de 63 puntos sobre 76. Es decir que el porcentaje completado de este proyecto es 82.89%. Esto quiere decir que el proyecto cumple con los objetivos recomendados en este modelo de referencia propuesto, ya que el puntaje final oscila entre un 75% y un 100%.

**Proyecto 4: Proyecto realizado por el ingeniero Emmanuel Lasso**

Este proyecto se realizó para la empresa-cliente 4. Esta es una plataforma para la gestión de riesgos agrarios en Colombia. Este proyecto está basado en datos espaciales. Las actividades giraron alrededor de la implementación de un modelo

mecanístico llamado Aquacrop, para que este funcionara con datos espaciales. Adicionalmente se realizaron exploraciones y análisis de datos para comprobar integridad matemática del modelo y obtener la sensibilidad de las variables que lo componen

Este proyecto tiene una puntuación de 50 puntos y la ruta elegida por el ingeniero fue un reporte analitico como se puede observar en la figura 28.


Figura 28. Ruta elegida en el cuestionario de validación [fuente propia]

Esto quiere decir que la máxima cantidad de puntos que se pudo obtener en este proyecto son 76 cómo se logra apreciar en la figura 15 mostrada anteriormente.

A primera vista, este proyecto culminó alrededor del 50/76x100 = 65,78%

Sin embargo, al igual que en los proyectos analizados anteriormente, cabe señalar que algunas preguntas se deben verificar manualmente, por tanto es necesario realizar un barrido de las respuestas que se tienen en la herramienta google forms. Al hacer este barrido se observa que:

- La pregunta ilustrada en la figura 29, debe verificarse manualmente debido a que para este caso en específico se otorgan 4 puntos si se han seleccionado las primeras 4 opciones presentadas en el cuestionario. La razón por la cual no se otorga un punto extra si se ha seleccionado la opción “herramientas de AWS” es porque no es necesario que cumpla con este requisito para la realización de un proyecto, sin embargo es posible tenerlo en cuenta por si se encuentra realizando el proyecto con herramientas de AWS debido a que esta guía posee recomendaciones de AWS en la mayoría de subetapas creadas, ya que uno de los objetivos específicos consiste en estructurar los lineamientos basados en la aproximación MLOps de Amazon al interior de la empresa. Por ello, es importante tenerlo presente. Para este caso se debe

sumar cuatro puntos extra a la puntuación final, lo que lleva a un total de 54 puntos.

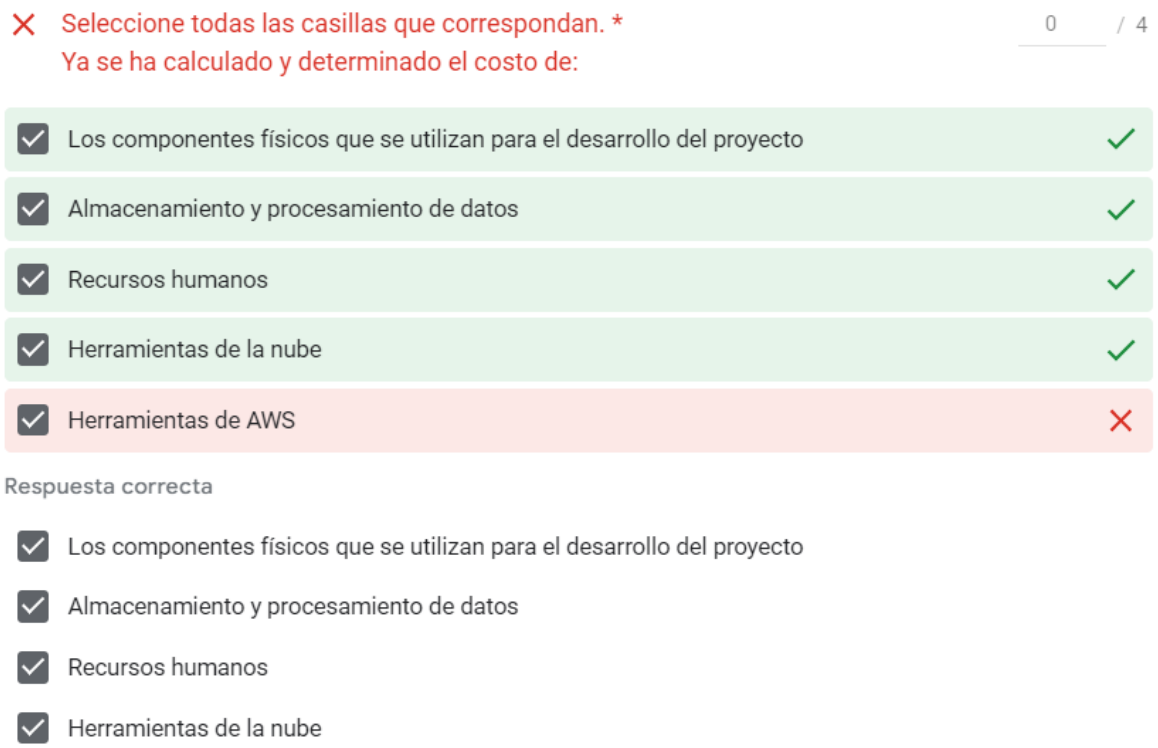


Figura 29. Selección múltiple con múltiples respuestas: Cálculo de costos [fuente propia]

- En la figura 30 la cantidad de puntos aumenta con cada respuesta seleccionada correctamente. Para este caso el ingeniero seleccionó dos de las seis opciones que se presentaban. Por ende se deben sumar dos puntos extra a la puntuación final. Lo que corresponde a 56 puntos en total.

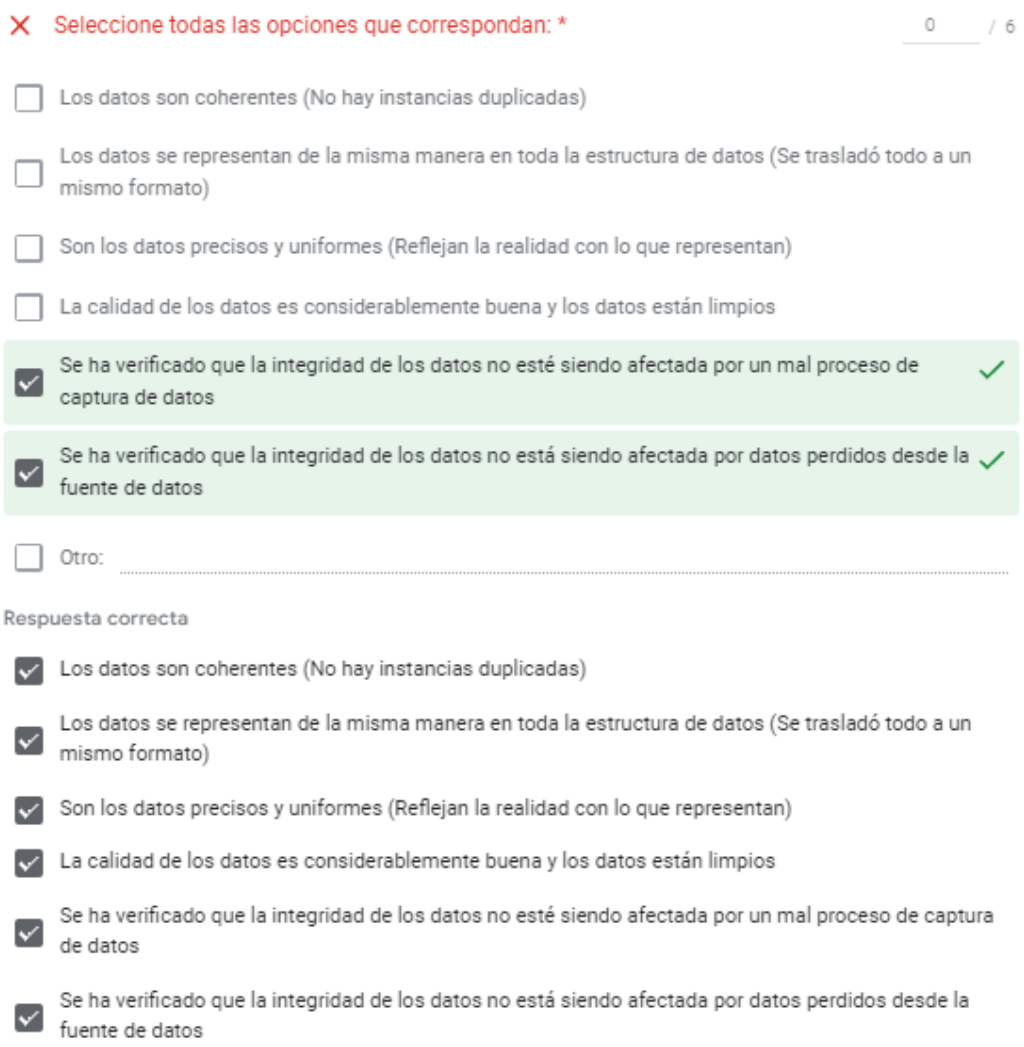


Figura 30. Selección múltiple con múltiples respuestas: Preparación de datos [fuente propia]

- La pregunta ilustrada en la figura 31, corresponde al mismo análisis que se ha realizado con algunos de los proyectos mencionados anteriormente. como en la figura 25 y la figura 18. El objetivo principal de esta pregunta es saber si existen fuentes de datos de entrenamiento disponibles y de qué tipo son estas fuentes de datos (propios, pagados o públicos). Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un total de 57 puntos.

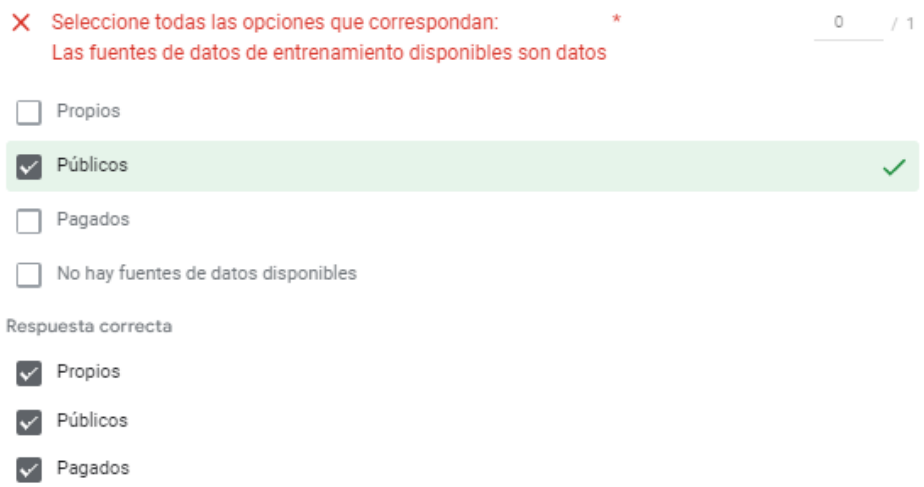


Figura 31. Fuentes de datos de entrenamiento disponibles [fuente propia]

- Un caso muy similar ocurre con la pregunta que se ilustra en la figura 32. En este caso, el objetivo es informativo. En donde las opciones son valores enteros, continuos, binarios, nominales y ordinales y el ingeniero seleccionó 4 de las 5 opciones correctas. Por ello, en este caso se debe sumar un punto extra a la puntuación final, lo que lleva a un puntaje total de 58 puntos.

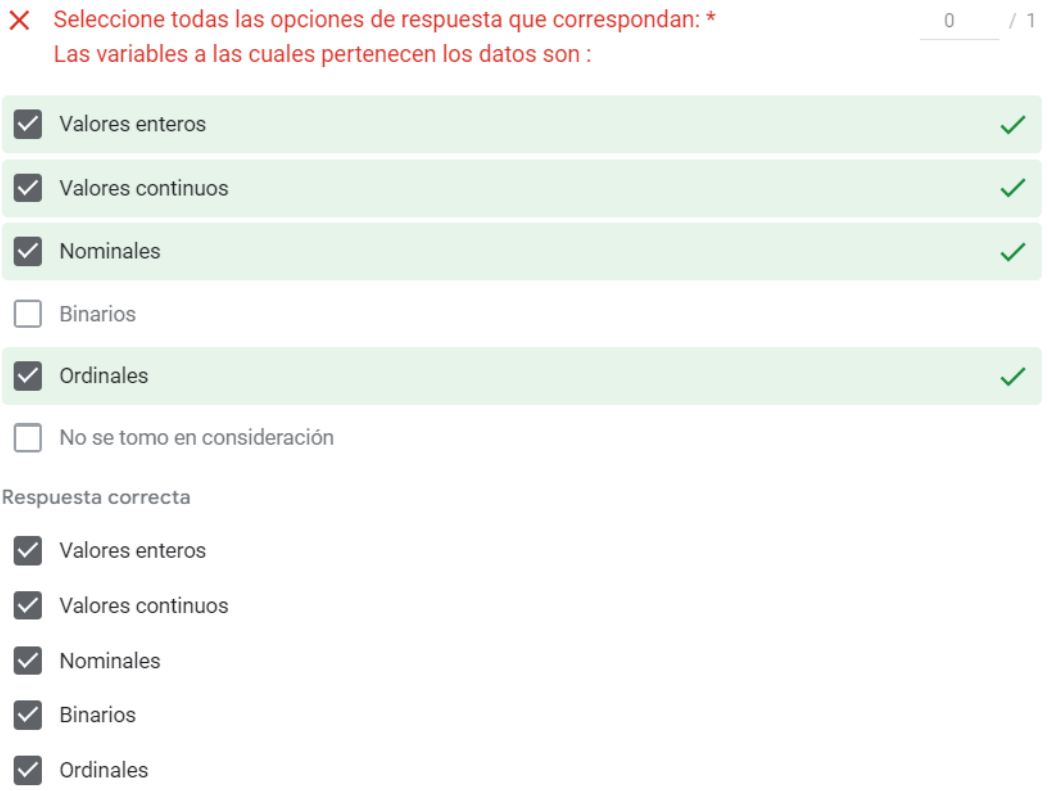


Figura 32. Las variables a las cuales pertenecen los datos [fuente propia]

Con el análisis realizado se concluye que el proyecto para la empresa-cliente 4 logró un total de 58 puntos sobre 76 es decir que el porcentaje real completado de este proyecto es 76.31%. Esto quiere decir que el proyecto cumple con los objetivos recomendados en este modelo de referencia propuesto, ya que el puntaje final oscila entre un 75% y un 100%.

**Proyecto 5: Proyecto realizado por el ingeniero Julián Plazas**

El objetivo de este proyecto es desarrollar un sistema de gestión y despliegue de datos IoT para realizar seguimiento a la productividad de una fábrica de dispositivos de medición.

Este proyecto tiene una puntuación de 17 puntos y la ruta elegida por el ingeniero fue un dashboard como se puede observar en la figura 33.

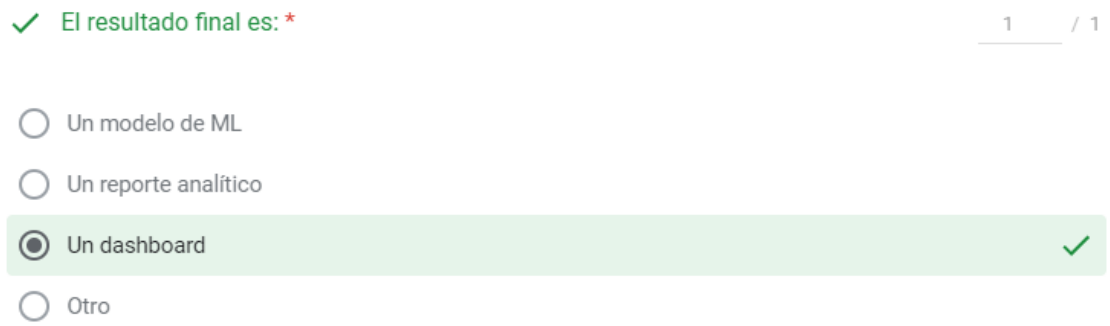


Figura 33. Ruta elegida en el cuestionario de validación [fuente propia]

Esto quiere decir que la máxima cantidad de puntos que se pudo obtener en este proyecto son 55 cómo se logra apreciar en la figura 15 mostrada anteriormente.

A simple vista, este proyecto culminó alrededor del 17/55x100 = 30,90%

Sin embargo, al igual que en los proyectos analizados anteriormente, cabe resaltar que algunas preguntas se deben verificar manualmente, por tanto es necesario realizar un barrido de las respuestas que se tienen en la herramienta google forms. Al hacer este barrido se encuentra que:

- La pregunta ilustrada en la figura 34, debe verificarse manualmente debido a que la opción “los componentes físicos que se utilizan para el desarrollo del proyecto” otorga un punto. Por tanto, el puntaje total hasta este punto es de 18 puntos. Cabe mencionar que el mecanismo es el mismo que en la figura 26 analizada anteriormente.

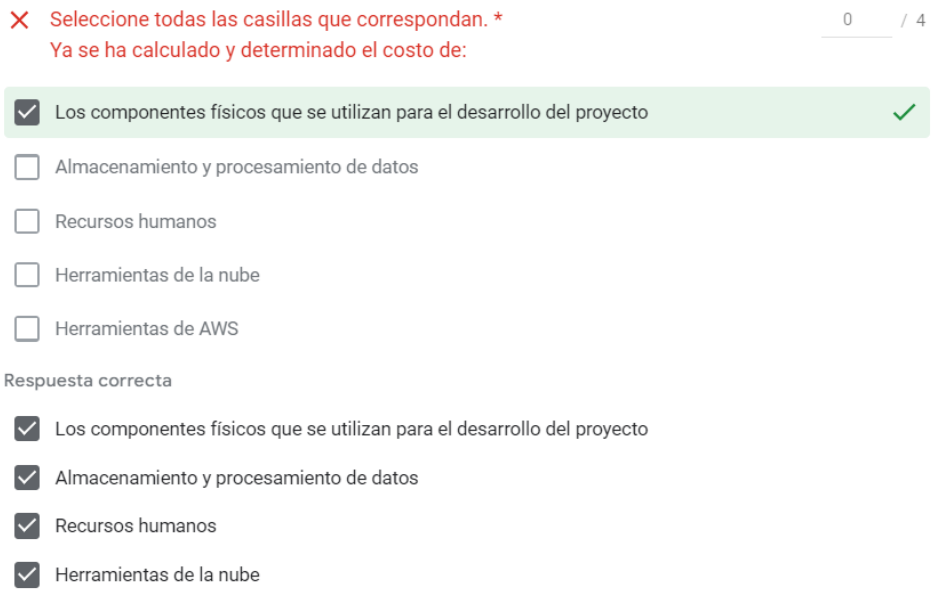


Figura 34. Cálculo de costos [fuente propia]

- Un caso similar sucede con la pregunta que se ilustra en la figura 35, debido a que las preguntas de tipo selección múltiple, solamente son correctas con la herramienta cuando se ha hecho la selección de todas las opciones posibles, a pesar de que no es necesario que se seleccionen todas las respuestas para que la pregunta se considere correcta. En este caso, se otorgan 2 puntos extra. Entonces el puntaje final total es de 20 puntos.

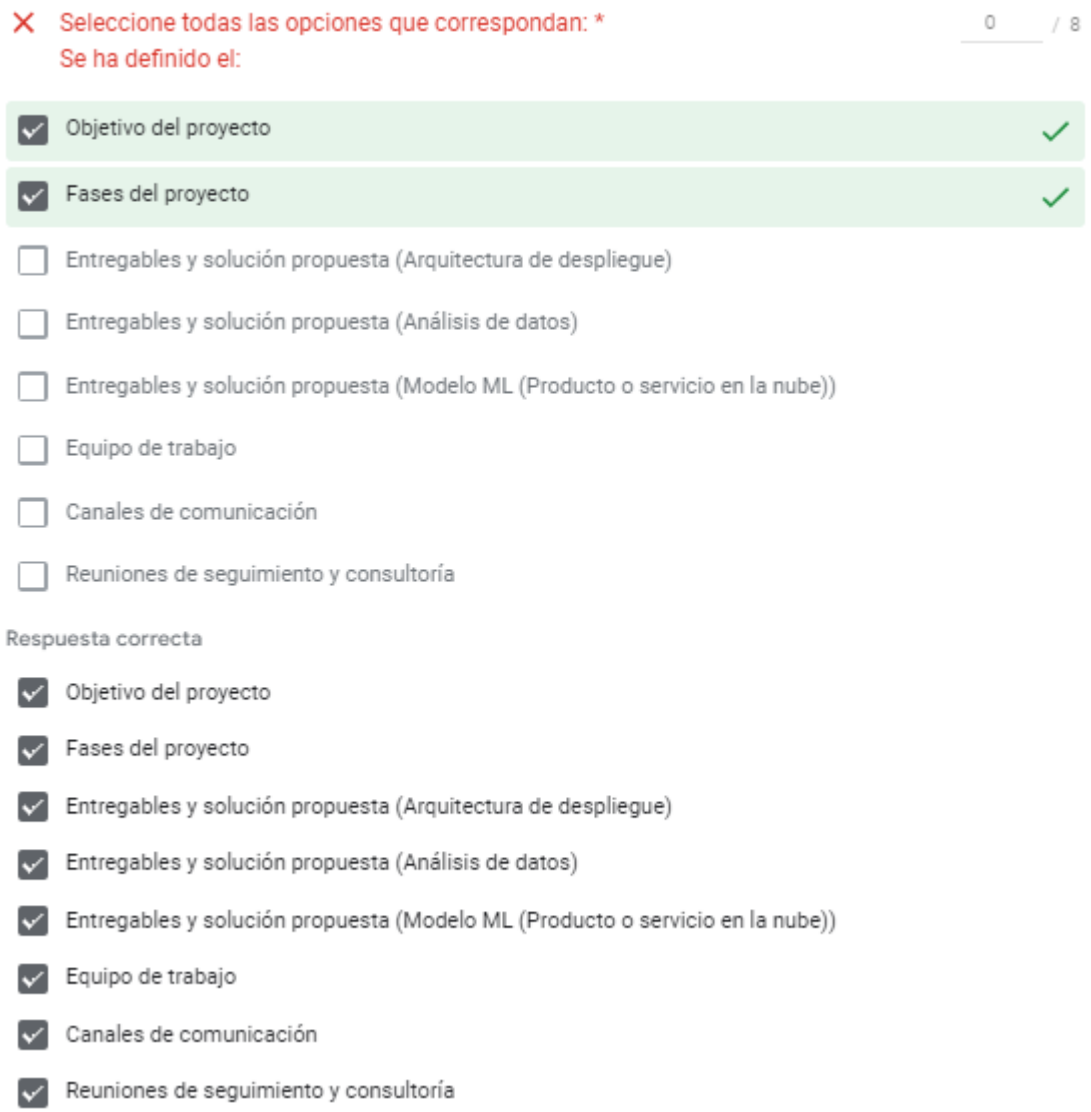


Figura 35. Selección múltiple con múltiples respuestas: Kick off [fuente propia]

Con el análisis realizado se concluye que el proyecto 6 logró un total de 20 puntos sobre 55 es decir que el porcentaje completado de este proyecto es 36.36%. Esto significa que el proyecto no logra el puntaje recomendado por este modelo de referencia, el cual sugiere obtener una puntuación ubicada entre el 75% y 100%. Por consiguiente se puede deducir que es posible que existan características que podrían mejorarse del proyecto según las propuestas de este modelo de referencia.

**Proyecto 6: Proyecto realizado por el ingeniero Edwar Girón**

El último proyecto es titulado “Datany Video” el cual fue una propuesta interna de WIZIT MIND BLOWING SOLUTIONS. Este es un proyecto de visión artificial, que consistió en analizar videos pregrabados o videos en vivo (tiempo real) que le permitiera (a la persona que adquiriera el servicio) obtener información demográfica

y del comportamiento de las personas dentro de un sitio específico. De esta manera, el proyecto ofreció un servicio con múltiples propósitos que se adaptara a las necesidades particulares del cliente.

Este proyecto tiene una puntuación de 71 puntos y la ruta elegida por el ingeniero fue un modelo de ML como se puede observar en la figura 36.

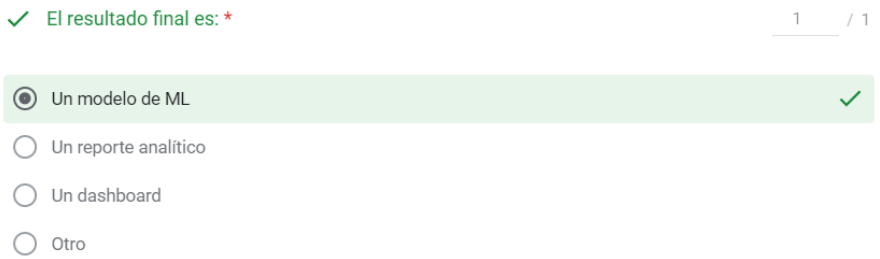


Figura 36. Ruta elegida en el cuestionario de validación [fuente propia]

Lo cual ubica la evaluación de puntos en el en un rango máximo de 115 a 122 puntos dependiendo de la métrica de evaluación elegida en el modelado. La métrica elegida por el ingeniero fue regresión como se puede observar en la figura 37.

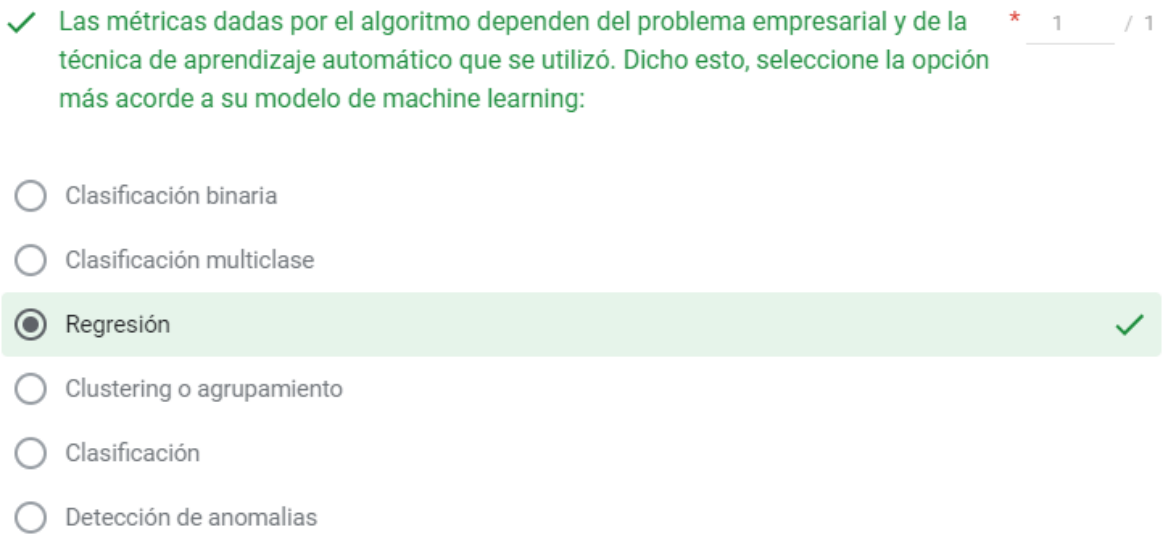


Figura 37. Algoritmo de machine learning [fuente propia]

Esto quiere decir que, la máxima cantidad de puntos que se pudo obtener en este proyecto son 117 como se puede apreciar en la figura 15 mostrada anteriormente.

A simple vista, este proyecto culminó alrededor del 71/117 = 60,68%

No obstante, es importante resaltar que es necesario evaluar algunas preguntas manualmente, por tanto se debe realizar un barrido de las respuestas que se tienen en la herramienta google forms. Al hacer este barrido se observa que:

- Para el caso de pregunta formulada en la figura 38 el objetivo es conocer el costo de recursos y herramientas importantes al momento de realizar un proyecto basado en ciencia de datos, es muy similar al caso ocurrido en la figura 26. Por ello, en este caso se debe sumar tres puntos extra a la puntuación final, lo que lleva a un puntaje total de 74 puntos.

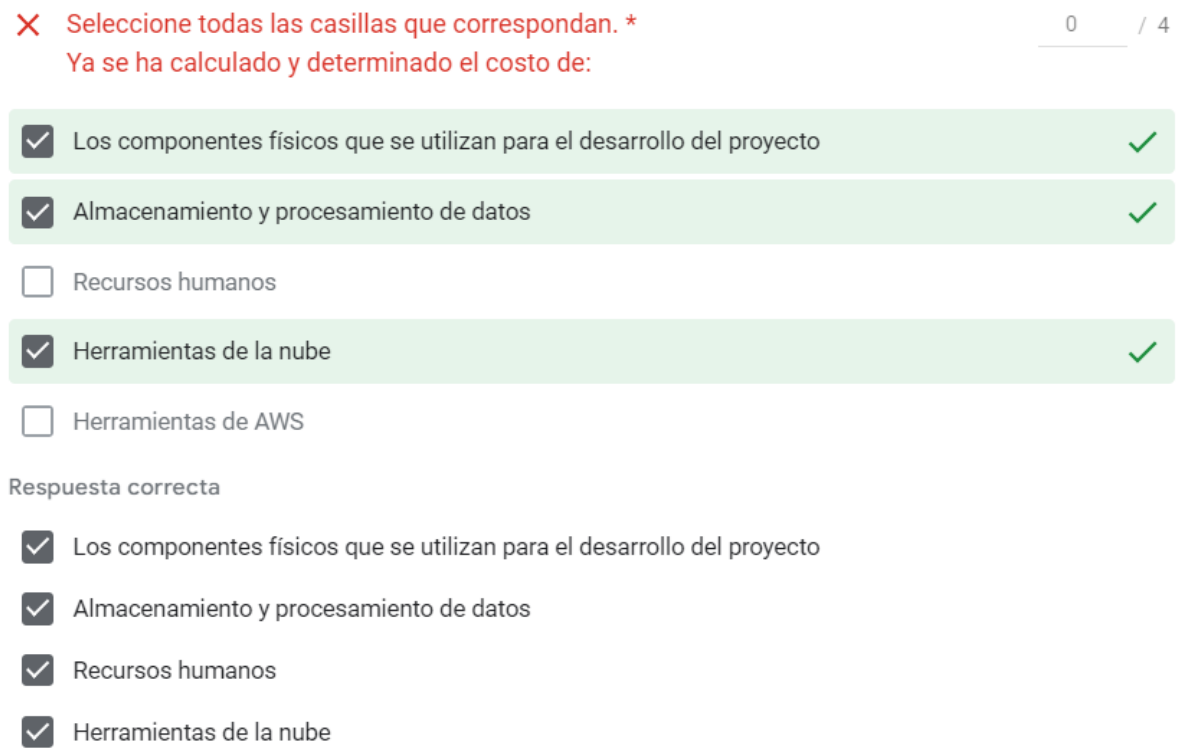


Figura 38. Cálculo de costos [fuente propia]

- Un caso similar sucede con la pregunta que se ilustra en la figura 39, ya que las preguntas de tipo selección múltiple, solamente se consideran correctas con la herramienta cuando se ha hecho la selección de todas las opciones posibles, como lo que ocurrió en el proyecto 3 en la figura 27. Entonces se debe aumentar 4 puntos extra al puntaje final lo que lleva a un total de 77 puntos.

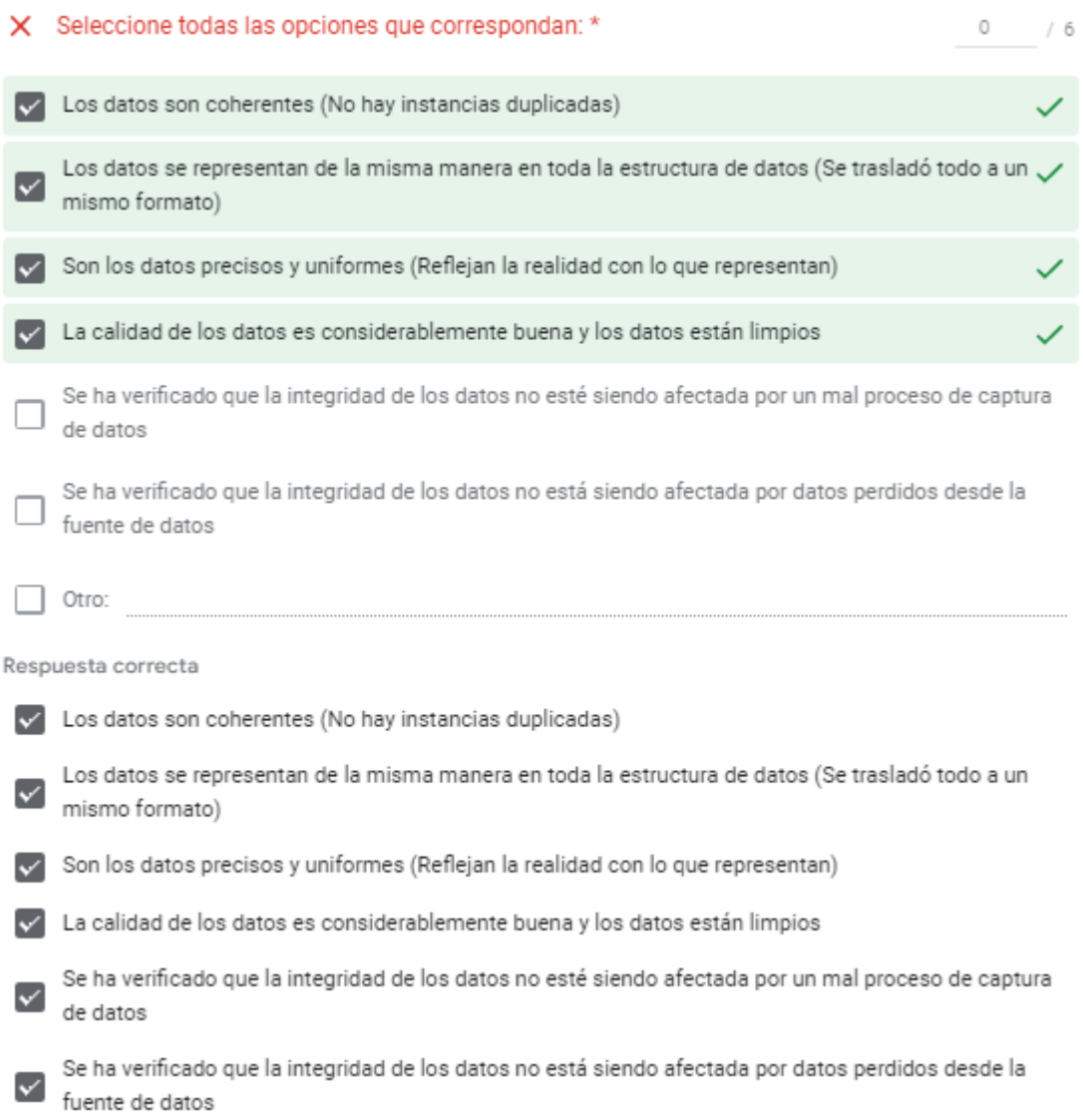


Figura 39. Selección múltiple con múltiples respuestas: Preparación de datos [fuente propia]

- Para las siguientes figuras 40 y 41, el objetivo es informativo, para conocer las variables con las que se trabajó en el proyecto. Por ello, se deben sumar 2 puntos extra a la puntuación final, lo que lleva a un puntaje total de 79 puntos.

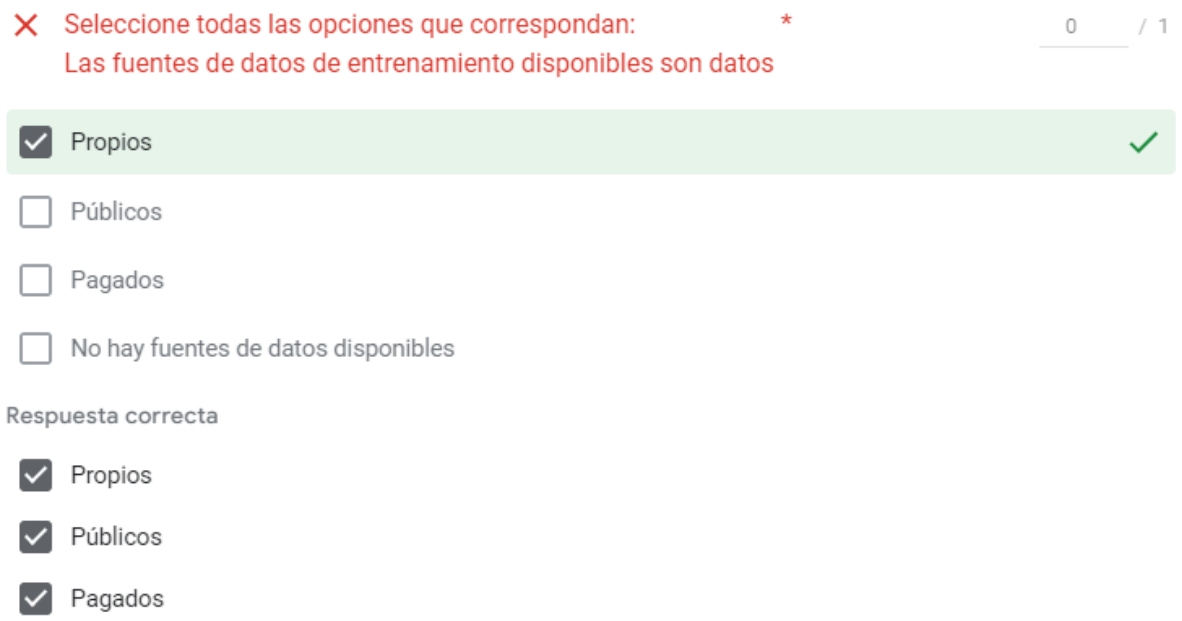


Figura 40. Fuentes de datos de entrenamiento disponibles [fuente propia]

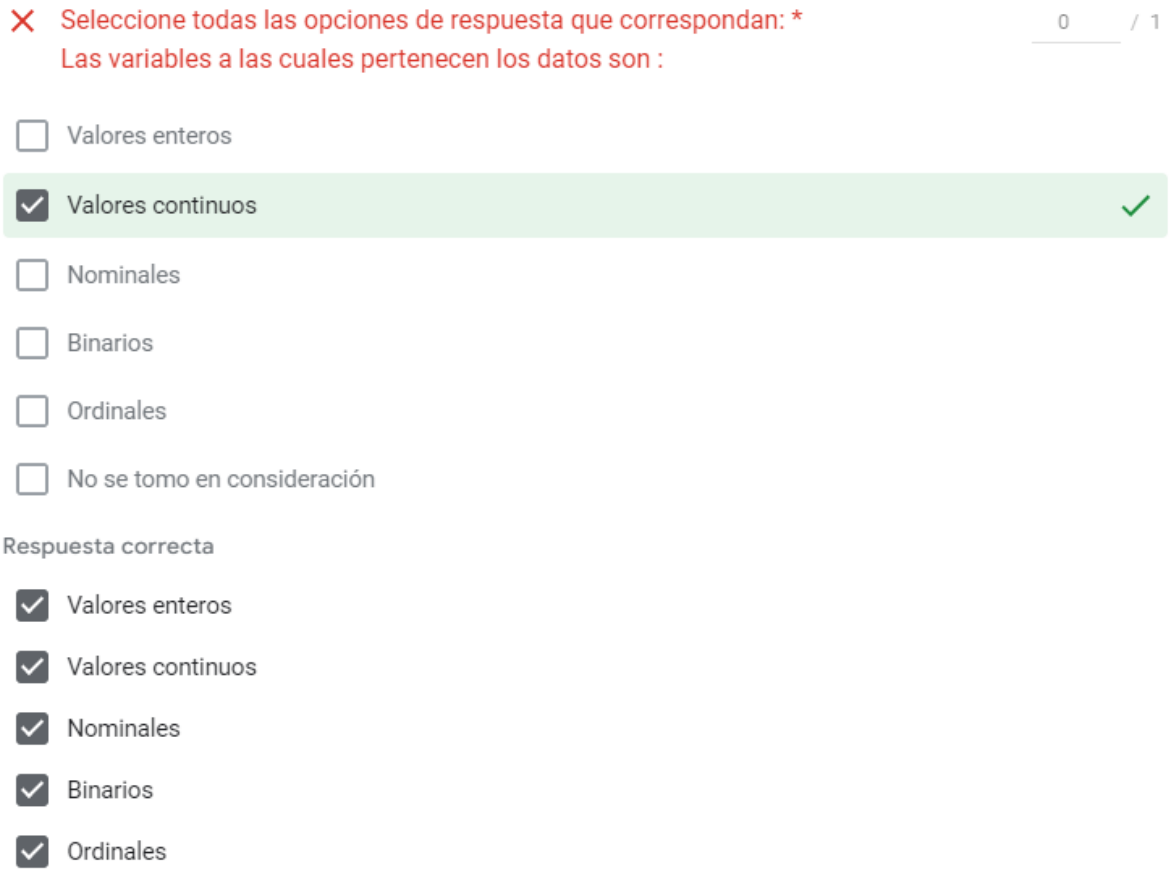


Figura 41. Variables a las cuales pertenecen los datos [fuente propia]

- Debido a que para este proyecto no se trabajó con datos estructurados, no se tenían en consideración las columnas y filas del dataset. Sin embargo, el ingeniero a cargo mencionó que si se examinaron las variables correspondientes al proyecto en cuestión, por tanto en la figura 42 y 43 se consideran estos dos puntos extras al puntaje final. Por tanto la puntuación total es 81 puntos.

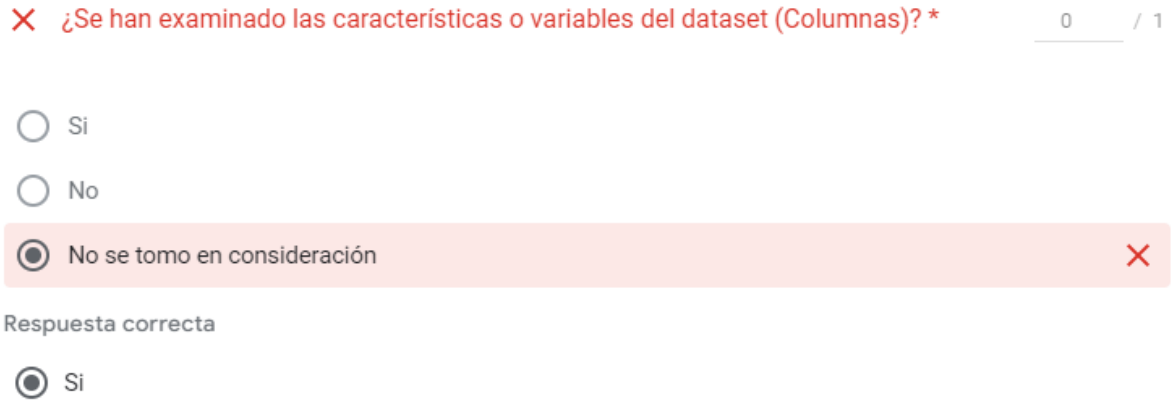


Figura 42. ¿Se han examinado las características o variables del dataset (Columnas)? [fuente propia]

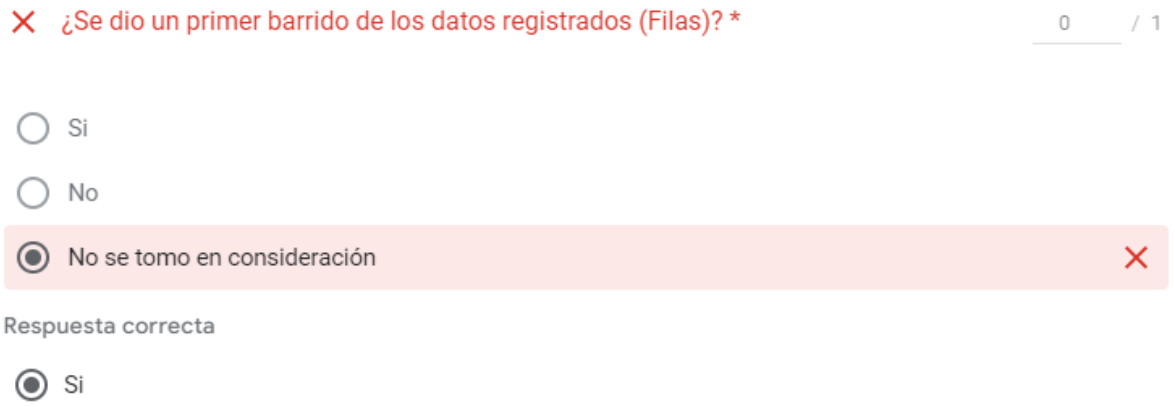


Figura 43. ¿Se dio un primer barrido de los datos registrados (Filas)? [fuente propia]

Con el análisis realizado se concluye que el proyecto “Datany video” logró un total de 81 puntos sobre 117 es decir que el porcentaje completado de este proyecto es 69.23%. Esto quiere decir que el proyecto no sigue las recomendaciones del modelo propuesto el cual oscila entre un porcentaje de 75% a 100%. Sin embargo, para este proyecto en específico, es importante mencionar que al no tratarse de un proyecto con datos estructurados, el modelo de referencia no representaba una guía 100% constituida para dicho proyecto.

## 5.3 Análisis general de las respuestas obtenidas

Ahora bien, se ha realizado un análisis de las respuestas obtenidas en el cuestionario en donde es posible observar que hay algunos indicadores importantes a tomar en consideración ya que la mayoría de proyectos no los cumplen, o no los toma en cuenta. Las preguntas en las cuales más fallan los usuarios que han llenado el cuestionario son:

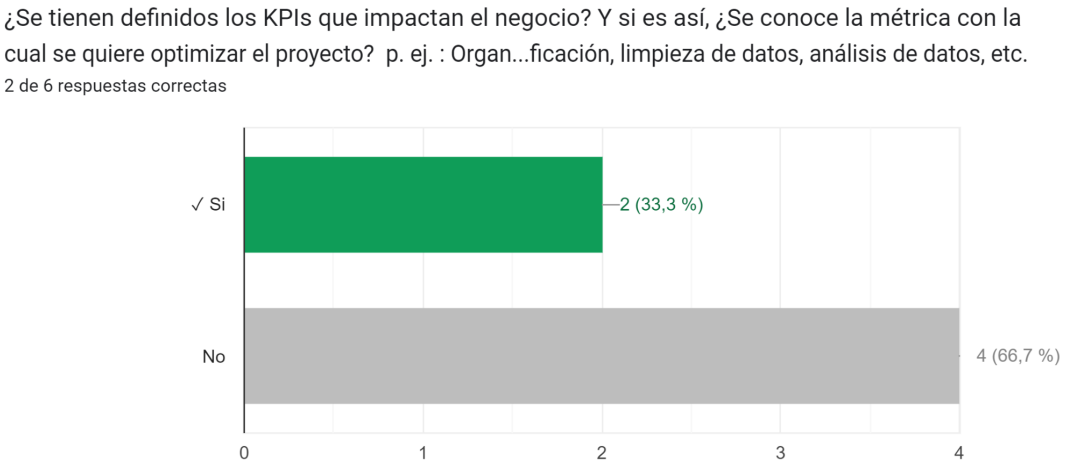


Figura 44. ¿Se tienen definidos los KPIs que impactan el negocio? [fuente propia]

La razón por la cual la mayoría de respuestas es “no” es debido a que en varias ocasiones los clientes no tienen claro cuales son los kpis de interés para ellos, entonces se vuelve una tarea a resolver durante el desarrollo del proyecto.

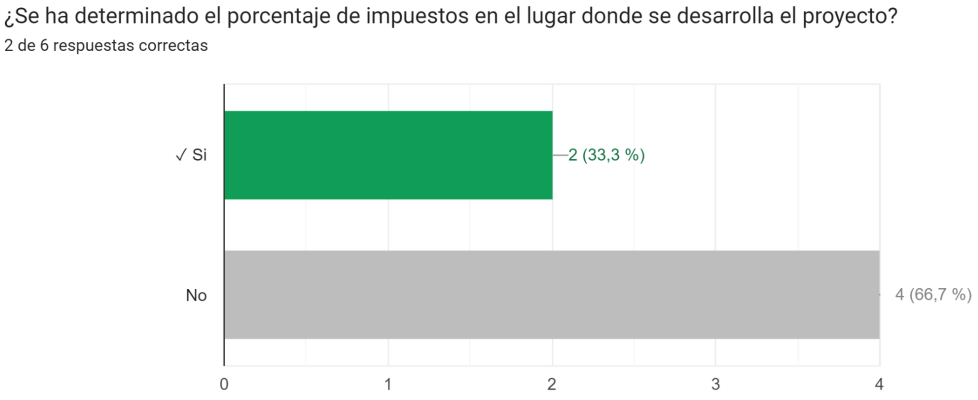


Figura 45. ¿Se ha determinado el porcentaje de impuestos en el lugar donde se desarrolla el proyecto? [fuente propia]

Debido a que el cuestionario fue llenado por el equipo técnico y no el equipo comercial, se desconoce si se había determinado el porcentaje de impuestos en el lugar donde se desarrolló el proyecto. Sin embargo, se generó un feedback por parte de los ingenieros el cual sugiere que el cuestionario sea llenado por ambos equipos, tanto el equipo técnico como el equipo comercial, y de esta manera poder responder correctamente el cuestionario.

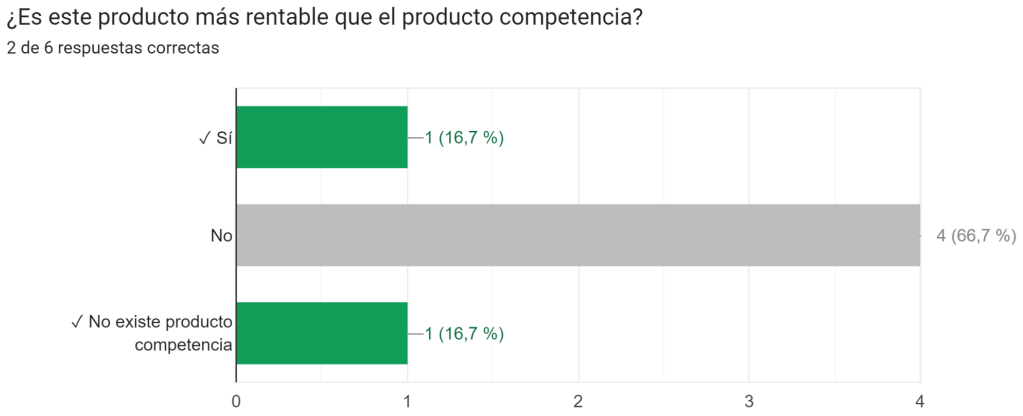


Figura 46. ¿Es este producto más rentable que el producto competencia? [fuente propia]

Esto se genera debido a que generalmente no se hace un estudio de mercado del producto ofertado, la necesidad llega por parte del cliente y se le desarrolla una solución software a la medida de lo que el cliente solicita. Por esta razón la mayoría de las veces la empresa no conoce productos competencia, ya sea porque el producto solicitado es demasiado específico o porque se acomodan a las necesidades del proyecto o porque no hay necesidad de hacer ese tipo de estudio.

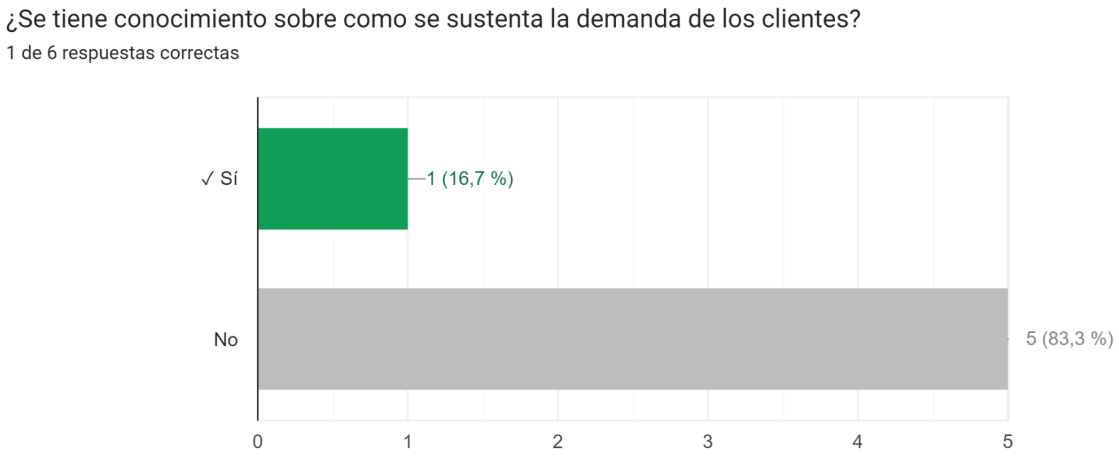


Figura 47. ¿Se tiene conocimiento sobre cómo se sustenta la demanda de los clientes? [fuente propia]

La pregunta hace referencia a que la solución sea sostenible en el tiempo y que se tenga almacenamiento suficiente. La respuesta de los ingenieros es “no” debido a que la empresa WIZIT MIND BLOWING SOLUTIONS realiza productos software a la medida, los cuales, una vez se consigue el resultado esperado se entregan al cliente. Los proyectos generalmente no tienen soporte y continuidad por parte de la empresa, por ende no se le brinda sustento a esa demanda del cliente.

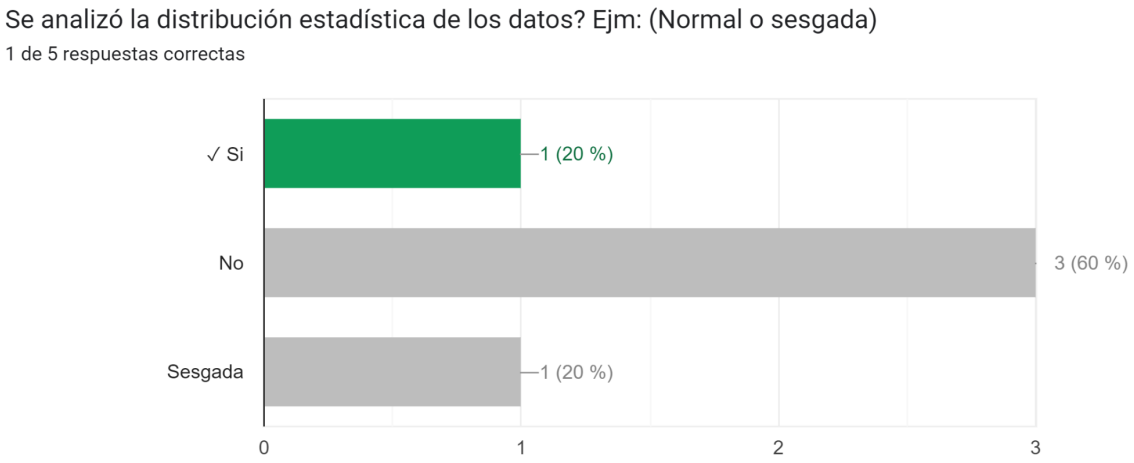


Figura 48. ¿Se analizó la distribución estadística de los datos? [fuente propia]

Dos de los proyectos que respondieron “no” es porque no tenían esa naturaleza de datos. Debido a que estas preguntas están diseñadas para proyectos que manejan datos estructurados, no otro tipo de datos.

La tercera persona que respondió que no, es debido a que efectivamente no se analizó la distribución estadística de los datos.

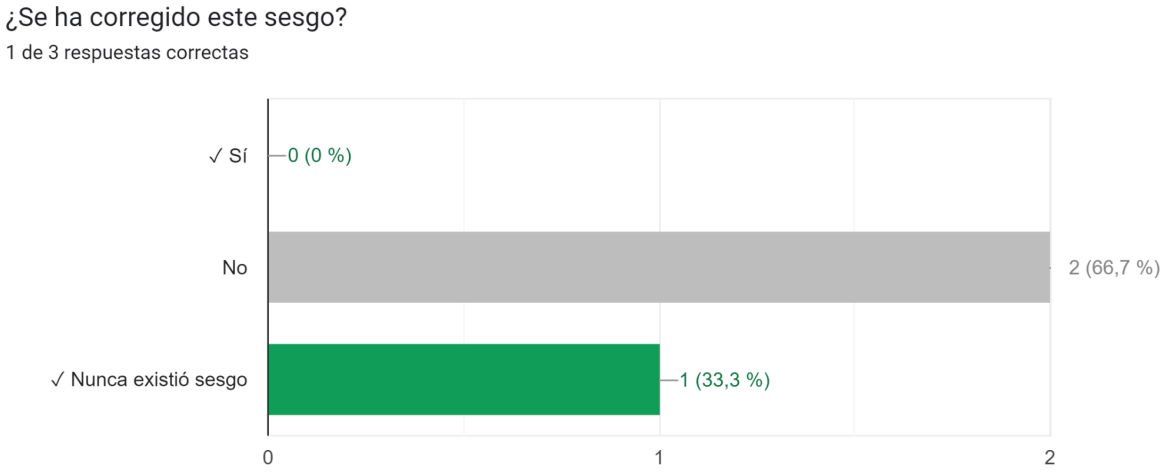


Figura 49. ¿Se ha corregido este sesgo? [fuente propia]

No se corrigió el sesgo.

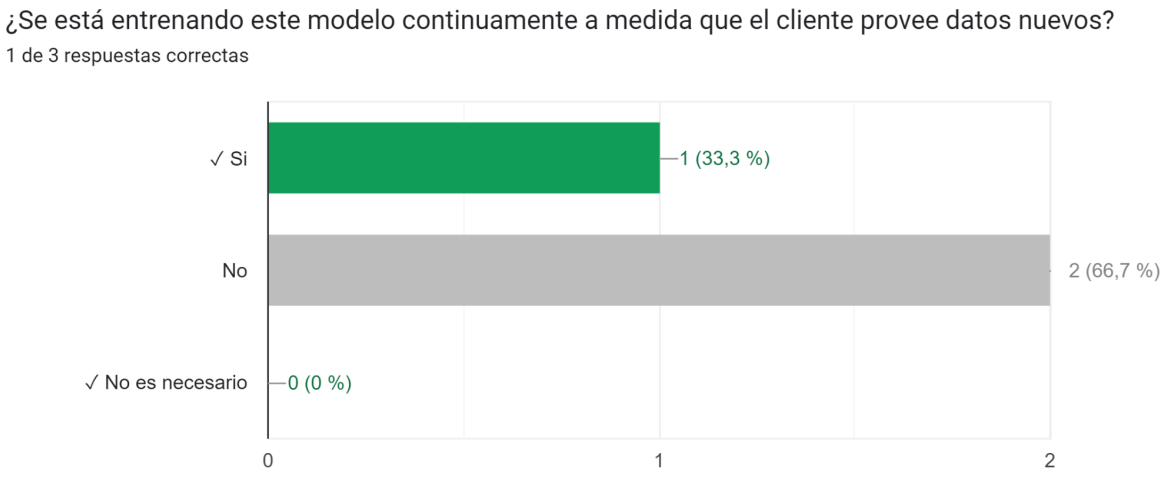


Figura 50. ¿Se está entrenando este modelo continuamente a medida que el cliente provee datos nuevos? [fuente propia]

Los proyectos no tuvieron esa necesidad, debido a que no se le brindaba continuidad y soporte al proyecto.

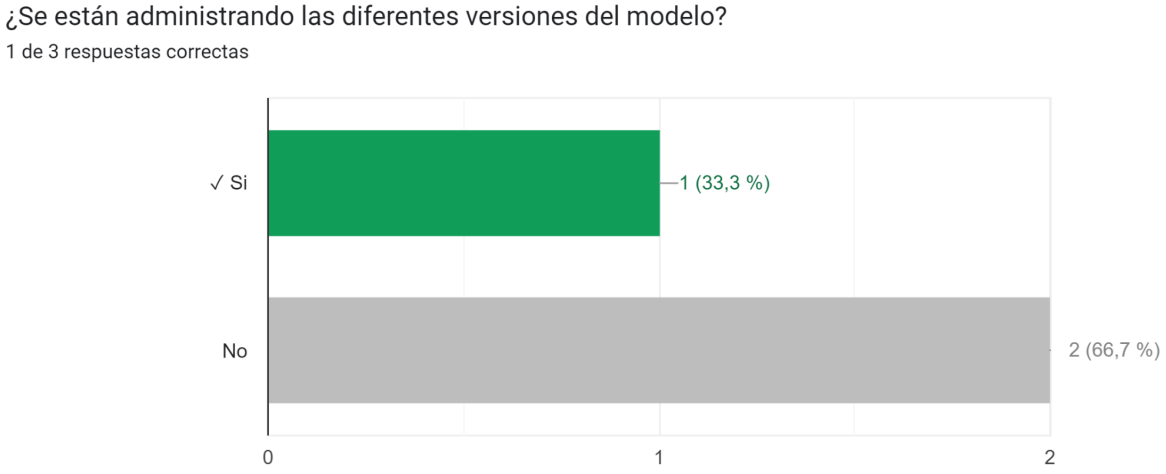


Figura 51. ¿Se están administrando las diferentes versiones del modelo? [fuente propia]

No se administran las diferentes versiones del modelo debido a que una vez el producto software está listo se entrega y no se continúa, ni se realiza ningún tipo de soporte.

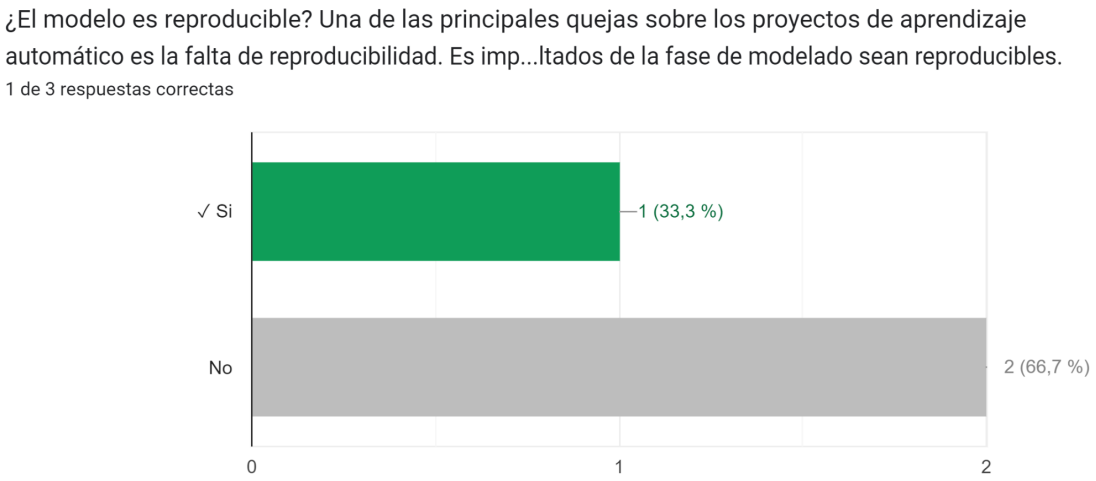


Figura 52. ¿El modelo es reproducible? [fuente propia]

Esto sucede debido a que es modelo hecho a la medida que no es fácilmente reproducible en otros casos.


Figura 53. ¿Se posee alguna estrategia de implementación? [fuente propia]

El modelo se entrega con una arquitectura que generalmente está elaborada sobre los servidores e infraestructura del cliente, por eso no hay ninguna estrategia de implementación para llevarlo a producción.

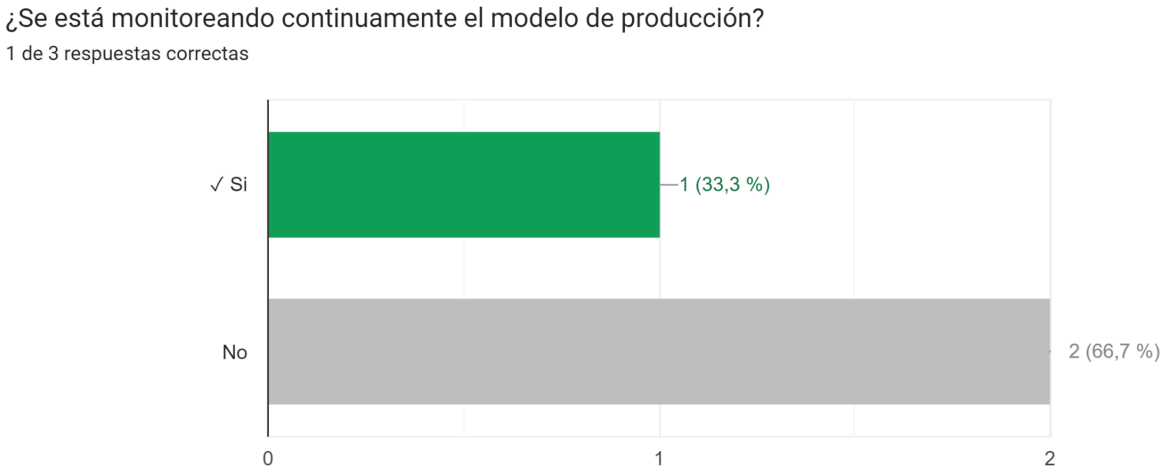


Figura 54. ¿Se está monitoreando continuamente el modelo de producción? [fuente propia]

La gran mayoría de proyectos en la empresa WIZIT MIND BLOWING SOLUTIONS S.A.S generan el producto software lo entregan y no realizan modificaciones, aportes, ni ningún tipo de cambio a los productos después de ser entregados.

Como se puede observar en las gráficas obtenidas por google forms la empresa no toma en cuenta puntos importantes al momento de realizar un proyecto como lo son:

- KPIs
- Costos
- Estrategias de deployment
- Rentabilidad
- Canalización CI/CD
- Monitorización del modelo en producción
- Sesgo
- Reproducibilidad

En este punto se puede evidenciar que la empresa tiene un vacío en el despliegue debido a la naturaleza de su modelo de negocios, ya que no suelen brindar soporte sobre un producto ya entregado, al menos no durante un tiempo constante. Algunas veces se realiza soporte durante un tiempo corto en caso de fallas inesperadas, pero no sobre un servicio en la nube ya que suelen ser productos entregados a la medida y sobre la infraestructura del cliente, lo cual va en contravía a lo que se propone en MLOps. Sin embargo, este modelo de referencia está diseñado para

mejorar estos puntos débiles que se han encontrado en la empresa, y generar una mayor posibilidad de éxito en los proyectos que se estén trabajando.

# Capítulo 6

# Conclusiones y trabajos futuros

## 6.1 Conclusiones

En este documento monográfico se propuso un modelo de referencia basado en MLOps para el soporte de la gestión en proyectos de ciencia de datos. Este modelo está estructurado por 5 etapas que son:

- Comprensión del negocio
- Preparación de datos
- Análisis de datos
- Modelado
- Deployment

Las cuales a su vez poseen subetapas que explican paso a paso el proceso que se propone en este modelo de referencia, además de recomendaciones y conceptos. Esta explicación detallada se encuentra en la sección 4.3.

Para la validación del modelo de referencia se elaboró un cuestionario utilizando la herramienta google forms. Este cuestionario fue validado con 6 proyectos diferentes en los cuales hay 4 ingenieros involucrados y un estudiante de la facultad de ingeniería electrónica y telecomunicaciones de la Universidad del Cauca. Dicho estudiante realizó su proyecto al mismo tiempo que este modelo de referencia estaba en proceso de desarrollo. Ambos proyectos se encuentran orientados hacia el desarrollo y aplicación de MLOps, por ende resultó conveniente que los dos proyectos se llevarán a cabo de manera simultánea. Con lo anterior dicho, el proyecto titulado "**PROPUESTA DE IMPLEMENTACIÓN DE UN PILOTO DE MLOPS SIGUIENDO LOS LINEAMIENTOS DEFINIDOS POR LA EMPRESA WIZIT MIND BLOWING SOLUTIONS S.A.S.**" procuró aplicar este modelo de referencia planteado en un alto porcentaje (70,33%). Sin embargo no fue posible seguirla al pie de la letra debido a que para dicho proyecto no era necesario llevar a cabo todos los puntos proyectados de este modelo de referencia. Sin embargo, es importante

mencionar que los 6 proyectos restantes estaban en su fase final, por tanto este cuestionario tiene 2 objetivos principales:

- El primer objetivo consiste en evaluar el estado de cada proyecto, para ello se ha creado un documento en excel en donde, dependiendo la cantidad de puntos, se puede concluir el estado en el que se encuentra un proyecto, basado en la lista de validación que propone la encuesta. Dicho documento se maneja de la siguiente manera:

    - Las preguntas de “si” o “no” suman un punto o no, dependiendo de la respuesta. La respuesta “si” otorga un punto y la respuesta “no” otorga 0 puntos.
    - Las preguntas de selección múltiple dan un punto por cada selección, es decir, si una pregunta de selección múltiple posee 7 respuestas, entonces si el usuario las selecciona todas, se le otorgará 7 puntos. Sin embargo, es importante mencionar que en este tipo de pregunta el conteo de puntos se debe realizar manualmente debido a que google forms no permite hacer sumatoria de puntos individual cuando se realiza una pregunta en selección múltiple con múltiples respuestas correctas.

- Cabe resaltar que las preguntas en las que se debe cambiar la puntuación manualmente son muy pocas. Además de ello, el documento de excel permite ver el porcentaje en el que se encuentra el proyecto en cada una de las secciones y también permite observar el porcentaje del proyecto en su totalidad hasta ese punto. Cada uno de los puntos tiene su explicación, el objetivo es que las personas de la empresa puedan checar con ello el estado de un proyecto y que tan completo está.

- El segundo objetivo consiste en que el equipo de ingenieros a cargo de realizar la encuesta noten si hay pasos faltantes en el desarrollo del proyecto, con ello analizar si hay características faltantes, o métricas que quizás no se tuvieron en cuenta, y verificar si hay áreas específicas que quizás no han considerado y son relevantes para el éxito del proyecto.

  Es importante resaltar que los integrantes del equipo de ciencia de datos que realizan el cuestionario deben ser los líderes junto con su equipo técnico y de negocio, en vista de que el cuestionario tiene preguntas de todos los dominios, tanto técnicas, como de negocio. Por tanto se necesita el conocimiento de todo el equipo para resolverlo y debido a que cada miembro del equipo posee un rol diferente con una tarea diferente, se debe responder el cuestionario con las personas indicadas para evitar saltarse algún paso del modelo de referencia y poder responder el cuestionario conscientemente.

- Al realizar un análisis de las respuestas generadas por google forms. Se puede concluir que los proyectos cumplen con un 44.26% - 47.20% del modelo de referencia. Donde las áreas más fuertes suelen ser:

- La comprensión del negocio y de requerimientos del usuario
- La adaptabilidad
- Análisis adecuado para realizar un completo EDA (Exploratory data analysis)
- Poseer fuentes de datos de entrenamiento disponibles

  Sin embargo las áreas que más se le dificulta cumplir en la empresa son: Deployment, el área financiera, competitiva y rentable, monitorización de proyectos de machine learning. Para ello:

- Se sugiere que la empresa WIZIT MIND BLOWING SOLUTIONS provea servicios y procure trabajar en la nube.
- Utilice este modelo de referencia.
- Se sugiere además que: el cuestionario se debe realizar en varios momentos durante el desarrollo del proyecto con el fin de observar y tomar en cuenta los cambios que se han presentado en el mismo. Esto se puede notar fácilmente con un análisis rápido de las respuestas que se hayan proveído por el equipo de ciencia de datos. Con ello se concluye que al realizar este experimento en un grupo más amplio, se puede deducir que la herramienta se puede extender a distintos momentos del proyecto, no es necesario llenar el cuestionario solo al final para verificar en qué estado está, porque puede tener interpretaciones distintas dependiendo del estado del proyecto.

- Otra conclusión que se encontró en la reunión con los ingenieros que llenaron la encuesta es que un proyecto de analítica no tiene solamente una ruta. Sino que es un conjunto de actividades que se pueden mezclar dependiendo de las necesidades del proyecto. Y dependiendo de lo que se haya hecho en el proyecto, puede que haya recomendaciones de buenas prácticas para cada una de las etapas con el fin de que se aproxime a una implantación de machine learning operations. Este feedback se considerará en trabajos futuros.

## 6.2. Trabajos futuros

- Aplicar el modelo de referencia en un proyecto nuevo con el objetivo de llevar a cabo todas las etapas y comprobar si el modelo de referencia ayuda a completar exitosamente un proyecto de ciencia de datos siguiendo todas las etapas de principio a fin.
- Generar un modelo de referencia para datos no estructurados y para series de tiempo, debido a que este modelo de referencia está basado en datos estructurados es decir están representados en formato tabular.
- Adicionar un esquema de visualización que permita realizar un seguimiento a las etapas del cuestionario de validación aplicado para los proyectos.

# Referencias bibliográficas


[1] ASTURIAS CORPORACIÓN UNIVERSITARIA®, “Estructuras, metodologías y métodos ágiles y Lean”. 2022.

[2] R. Gurrola y J. G. Rodriguez Rivas, *Ciencia de los Datos, Propuestas y casos de uso*. 2020.

[3] A. Azevedo y M. F. Santos, “KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW”, p. 6.

[4] J. Saltz, “A J U M P S T A R T T O Key concepts to help you successfully execute data science projects”. 2021.

[5] A. Goyal, “Machine Learning Operations | International Journal of Information Technology Insights & Transformations [ISSN: 2581-5172 (online)]”. http://technology.eurekajournals.com/index.php/IJITIT/article/view/655 (consultado el 29 de noviembre de 2022).

[6] I. El Naqa y M. J. Murphy, “What Is Machine Learning?”, en *Machine Learning in Radiation Oncology: Theory and Applications*, I. El Naqa, R. Li, y M. J. Murphy, Eds. Cham: Springer International Publishing, 2015, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

[7] “(PDF) Machine Learning Algorithms -A Review”. https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review (consultado el 1 de diciembre de 2022).

[8] S. Islam, “Definition of Project Management: What Is Project”, Consultado: el 1 de diciembre de 2022. [En línea]. Disponible en: https://www.academia.edu/31616426/Definition_of_Project_Management_What_Is_Project

[9] J. Bern, “Introduction to Data Science”.[8]\uc0\u8220{}Data Science and Machine Learning: Mathematical and Statistical Methods\uc0\u8221{}

[11] “Data Engineer Roles and Responsibilities: Skills Required in 2022”, *Simplilearn.com*, el 3 de diciembre de 2019. https://www.simplilearn.com/data-engineer-role-article (consultado el 1 de diciembre de 2022).

[12] “Job Description for Data Analyst: Responsibilities and Skills Required”, *Simplilearn.com*, el 16 de abril de 2021. https://www.simplilearn.com/data-analyst-job-description-article (consultado el 1 de diciembre de 2022).

[13] “How to Become a Database Developer – A Complete Career Guide”, *DiscoverDataScience.org*. https://www.discoverdatascience.org/career-information/database-developer/ (consultado el 1 de diciembre de 2022).

[14] C. Ebert, G. Gallardo, J. Hernantes, y N. Serrano, “DevOps”, *IEEE Softw.*, vol. 33, núm. 3, pp. 94–100, may 2016, doi: 10.1109/MS.2016.68.

[15] D. Velimirovic, V. Milan, y S. Rade, “Role and importance of key performance indicators measurement”, *Serbian J. Manag.*, vol. 6, mar. 2011, doi: 10.5937/sjm1101063V.

[16] “AWS Well-Architected Framework - AWS Well-Architected Framework”. https://docs.aws.amazon.com/es_es/wellarchitected/latest/framework/welcome.html (consultado el 2 de diciembre de 2022).

[17] “Definiciones - AWS Well-Architected Framework”. https://wa.aws.amazon.com/wat.definition.wa-def.es.html (consultado el 2 de diciembre de 2022).

[18] “What is CI/CD?” https://www.redhat.com/en/topics/devops/what-is-ci-cd (consultado el 5 de diciembre de 2022).

[19] J. of C. S. Ijcsis y C. K. Nwagu, “Knowledge Discovery in Databases (KDD): An Overview”, Consultado: el 1 de diciembre de 2022. [En línea]. Disponible en: https://www.academia.edu/35695793/Knowledge_Discovery_in_Databases_KDD_An_Overview

[20] M. A. R. Herrera y J. D. A. Vasquez, “ESTUDIO SOBRE EL ESTADO DE LAS SOLUCIONES ICT Y DE LOS CASOS PRÁCTICOS DE APLICACIÓN DE LA MINERÍA DE DATOS A NIVEL MUNDIAL EN AL MENOS 5 CASOS REPRESENTATIVOS.”, p. 228.
[21] V. G. Cortina, “Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario”, p. 120.
[22] “DDS Guide: How it works – Data Driven Scrum”. https://datadrivenscrum.com/how-dds-works/ (consultado el 1 de diciembre de 2022).
[23] M. Testi *et al.*, “MLOps: A Taxonomy and a Methodology”, *IEEE Access*, vol. 10, pp. 63606–63618, 2022, doi: 10.1109/ACCESS.2022.3181730.
[24] D. Kreuzberger, N. Kühl, y S. Hirschl, “Machine Learning Operations (MLOps): Overview, Definition, and Architecture”, p. 13.
[25] S. Mäkinen, “Designing an open-source cloud-native MLOps pipeline”. el 12 de marzo de 2021.
[26] D. S. Battina, “AN INTELLIGENT DEVOPS PLATFORM RESEARCH AND DESIGN BASED ON MACHINE LEARNING”, *SSRN Electron. J.*, vol. 6, pp. 68–75, mar. 2019.
[27] T. Fernandez, “What Is Canary Deployment?”, *Semaphore*, el 1 de septiembre de 2020. https://semaphoreci.com/blog/what-is-canary-deployment (consultado el 30 de noviembre de 2022).
[28] “Blue/Green Deployments on AWS”, p. 34.
[29] Á. García, “Qué son GIT y GIT FLOW?”, *Castor Transformación Digital*, el 23 de abril de 2020. https://castor.com.co/integracion-continua-de-software-git-y-git-flow/ (consultado el 30 de noviembre de 2022).
[30] javier, “Tutorial básico de Git y GitHub para uso de control de versiones”, *Blog de Javier Rguez*, el 2 de diciembre de 2016. https://www.javierrguez.com/tutorial-basico-git-github-uso-control-versiones/ (consultado el 30 de noviembre de 2022).
[31] D. Oppenheimer, “10 measures and KPIs for ML success Building the business case for MLOps and management”.
[32] C. Más, “¿Qué son los indicadores KPI y qué tipos existen?”, *Vilma Núñez - Consultora Estratégica de Marketing*, el 1 de noviembre de 2018. https://vilmanunez.com/indicadores-kpi/ (consultado el 2 de diciembre de 2022).
[33] “Machine Learning Canvas”, *OWNML*. https://www.ownml.co/machine-learning-canvas (consultado el 2 de diciembre de 2022).
[34] “¿Qué es Amazon QuickSight? - Amazon QuickSight”. https://docs.aws.amazon.com/es_es/quicksight/latest/user/welcome.html (consultado el 2 de diciembre de 2022).
[35] “Pilar de optimización de costos - AWS Well-Architected Framework”, p. 72.
[36] “AWS Pricing Calculator”. https://calculator.aws/#/ (consultado el 1 de diciembre de 2022).
[37] “La Propuesta de Valor. Qué es, cómo diseñarla y ejemplos”, *Guillermo Fuentes*, el 21 de febrero de 2019. https://guillermofm.com/propuesta-de-valor/ (consultado el 2 de diciembre de 2022).
[38] A. Pick, “Everything you need to know about pipelines for project management”, *monday.com Blog*, el 31 de marzo de 2021. https://monday.com/blog/remote-work/everything-you-need-to-know-about-pipelines-for-project-management/ (consultado el 6 de diciembre de 2022).
[39] “Project Pipeline Management: Everything You Need to Know”, *ProProfs Project Blog*, el 23 de febrero de 2021. https://www.proprofsproject.com/blog/project-pipeline-management/ (consultado el 6 de diciembre de 2022).
[40] “Feature engineering - Machine Learning Lens”. https://docs.aws.amazon.com/es_es/wellarchitected/latest/machine-learning-lens/feature-engineering.html (consultado el 29 de noviembre de 2022).
[41] luisquintanilla, “Métricas de ML.NET - ML.NET”.

https://learn.microsoft.com/es-es/dotnet/machine-learning/resources/metrics (consultado el 29 de noviembre de 2022).
[42] "10 sugerencias principales para Machine Learning reproducible", *Platzi*. https://platzi.com/blog/diez-sugerencias-para-machine-learning/ (consultado el 29 de noviembre de 2022).
[43] *¿Cómo manejar los DATOS FALTANTES?: guía completa*, (el 18 de junio de 2021). Consultado: el 6 de diciembre de 2022. [En línea Video]. Disponible en: https://www.youtube.com/watch?v=ARwHkq4t2q0
[44] *¿Cómo hacer el ANÁLISIS EXPLORATORIO DE DATOS?: guía paso a paso*, (el 11 de junio de 2021). Consultado: el 6 de diciembre de 2022. [En línea Video]. Disponible en: https://www.youtube.com/watch?v=_-KW4gT_oGU
[45] hmong.wiki, "Selección de modelo IntroducciónyDos direcciones de selección de modelo". https://hmong.es/wiki/Model_selection (consultado el 29 de noviembre de 2022).
[46] lgayhardt, "Selección de un algoritmo de Machine Learning - Azure Machine Learning". https://learn.microsoft.com/es-es/azure/machine-learning/how-to-select-algorithms (consultado el 29 de noviembre de 2022).