Tecnológico Universitario "RUMIÑAHUI"

ANALÍTICA DE DATOS DEL MODULO DE TIC´S OBTENIDOS DE LA ENESEM DEL INEC DEL AÑO 2012 AL AÑO 2015 SOBRE EL USO DE EQUIPOS, INTERNET Y PERSONAL PARA DETERMINAR INVERSIONES REALIZADAS EN EL ÁREA POR LAS EMPRESAS DEL ECUADOR.

EJECUCIÓN DE PROYECTOS

ALUMNOS:

FLORES PALAQUIBAY CRISTIAN DANIEL

LESCANO SALAZAR JOSEPH FRANSHUA

TUTORA: ING. ÁLVAREZ MAYRA. MSc

QUITO, SEPTIEMBRE 2023

AGRADECIMIENTO

Queremos comenzar agradeciendo a Dios por habernos brindado la oportunidad de culminar este proyecto y por habernos dado la capacidad y la paciencia para lograrlo.

Agradecemos a nuestra tutora, Ing. Mayra Álvarez, por su invaluable apoyo, su paciencia, su profesionalismo y su dedicación en cada una de las etapas de este proyecto. Su guía, comentarios y enseñanzas han sido de gran valor para mi formación académica y personal.

También queremos agradecer a la coordinadora de mi carrera, Ing. Yngrid Melo Mg., por habernos brindado su apoyo y orientación en las cuestiones administrativas asociadas al proyecto.

Agradecemos a nuestras familias y amigos por habernos brindado su amor, su apoyo y su motivación constante, sin lo cual no podríamos haber logrado este proyecto. Yo Cristian Flores en especial, quiero agradecer a mi esposa e hijos por entender y apoyarme en cada momento durante este proceso.

DEDICATORIA

Este proyecto representa una gran parte de nuestra formación académica y personal, y es por eso que queremos dedicarlo a.

En primer lugar, dedicamos este proyecto a Dios, que siempre nos ha brindado su protección y guía en cada etapa de nuestras vidas. Además, agradecemos a nuestras familias, por su cariño, su apoyo incondicional, su paciencia y su ayuda en todo momento.

Dedicamos este proyecto a todos aquellos estudiantes que, como nosotros, buscan siempre superarse y aprender de cada experiencia para mejorar y crecer profesionalmente. Este proyecto fue un gran desafío y estamos orgullosos de haberlo logrado con la ayuda y apoyo de nuestros seres queridos.

ÍNDICE DE CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

RESUMEN

Este proyecto tiene como objetivo analizar los datos recopilados por el INEC de las empresas ecuatorianas y su uso de las TIC durante los años 2012 al 2015. El proyecto utiliza técnicas de Big data y Machine Learning para modelar y predecir el uso de internet por las empresas, así como la inversión en TIC que han realizado. A través de un dashboard en PowerBI, se analizarán los datos recopilados y se obtendrá una comparación adecuada entre los años estudiados.

El objetivo general del proyecto es analizar los resultados de las encuestas realizadas en Ecuador sobre la utilización de TIC en las empresas. Además, se busca contextualizar la información, implementar modelos de aprendizaje supervisado para predecir el uso de internet y la inversión en TIC, y analizar los datos obtenidos mediante un dashboard en PowerBI.

Este proyecto podría ayudar a entender cómo la falta de inversión en TIC afecta a la competitividad de las empresas y a establecer políticas que fomenten una mejor implementación de las TIC en las empresas ecuatorianas.

ABSTRACT

This project aims to analyze the data collected by the INEC of Ecuadorian companies and their use of ICT during the years 2012 to 2015. The project uses Big data and Machine Learning techniques to model and predict the internet usage by the companies, as well as the ICT investment they have made. Through a PowerBI dashboard, the collected data will be analyzed and an adequate comparison between the studied years will be obtained.

The general objective of the project is to analyze the results of the surveys carried out in Ecuador regarding the use of ICT in companies. In addition, the project aims to contextualize the information, implement supervised learning models to predict the use of the internet and the ICT investment, and analyze the collected data through a PowerBI dashboard.

This project could help understand how the lack of ICT investment affects the competitiveness of companies and establish policies that encourage better implementation of ICT in Ecuadorian companies.

1. INTRODUCCIÓN

Actualmente, es evidente que las Tecnologías de la Información y la Comunicación (TIC) se han convertido en uno de los activos más valiosos dentro de los entornos empresariales. Sin embargo, medir el impacto de las TIC en una organización resulta desafiante debido a los altos costos asociados con su implementación inicial. Esta percepción de altos costos ha llevado a considerar las TIC como un gasto en lugar de una inversión.

Los costos que inciden en la implementación de las TIC dentro de una organización se destinan a la compra, desarrollo, ejecución de proyectos y equipos tecnológicos, además se debe considerar los valores del presupuesto destinados a las capacitaciones del personal que labora en el departamento de las TIC.

Según Costa et al. (2019) con datos del Instituto Nacional de Estadísticas y Censos (INEC), el sector empresarial a nivel nacional es responsable de producir el 75% del trabajo. Sin embargo, es preocupante observar que el 75% de estas empresas no utiliza adecuadamente las TIC en sus procesos diarios. Lo que llevan a la siguiente pregunta de investigación: ¿Cómo incide la falta de inversión en las TIC y la utilización de estas dentro de la competitividad de las empresas en el Ecuador?

Por tanto, surge la necesidad de analizar cómo la falta de inversión en TIC afecta la competitividad de las empresas. Las TIC ofrecen numerosas ventajas, como la mejora de la eficiencia operativa, la toma de decisiones informada, la ampliación del alcance de mercado y la promoción de la innovación. Sin embargo, si las empresas no reconocen el valor estratégico de las TIC y no invierten adecuadamente en ellas, es probable que se queden rezagadas en un entorno empresarial cada vez más competitivo y globalizado.

Ante estos antecedentes el presente proyecto realizara el análisis de los datos obtenidos en el INEC sobre las empresas que utilizan las TIC, mediante una representación gráfica, para establecer una comparación adecuada entre los años 2012 y 2015.

1.1. Antecedentes

Existen algunos estudios relacionados con el tema que dan a conocer empresas las cuales utilizan las TIC, para mejorar sus ingresos económicos y así por alcanzar a las metas deseadas.

Carrillo (2020), analiza el Impacto de las Tecnologías de Información en las pequeñas y medianas empresas del sector servicios en el Distrito Metropolitano de Quito (p.12), mediante investigación descriptiva pretende identificar el uso de TIC en las pequeñas y medianas organizaciones en el sector de servicios de reparación independiente. La muestra seleccionada no probabilística, con un enfoque cuantitativo para el tratamiento de las variables. Dando a conocer que es importante que las empresas manejen las TIC por varias situaciones, al momento de buscar proveedores, dar a conocer a la empresa por medio del marketing digital, con la utilización de las firmas electrónicas, etc., es considerable pensar en una inversión para la implementación de las TIC dentro de las empresas.

Molina y López (2021), realizan un estudio del uso de las Tics como medio de reactivación económica en las MiPymes de alojamiento del cantón Guayaquil para tiempos post-COVID, en el cual se identifican los factores adecuados para la implementación y uso de las TIC las cuales fueron muy importantes en la pandemia, porque ayudaron a que algunas empresas no cierren sus puertas y tengan grandes pérdidas.

1.2. Planteamiento del problema

A nivel mundial, se observa que los países desarrollados como: Estados Unidos, Suecia, Finlandia, Singapur y Dinamarca, son líderes en la adopción y utilización de las (TIC. Estos países han demostrado cómo las TIC pueden brindar numerosos beneficios a las empresas. Sin

embargo, a pesar de estas ventajas, también se ha evidenciado un retroceso en su aprovechamiento debido al desconocimiento y la falta de habilidades en su uso y manejo.

En Ecuador, se ha observado una deficiente utilización de TIC, lo cual limita el aprovechamiento de los recursos disponibles. Esto se debe a diversos factores, como la ineficiencia de la estructura, los entornos políticos y regulatorios, así como las condiciones del mercado. Como consecuencia, se dificulta el desarrollo adecuado de las empresas en general.

Ante esta problemática el presente proyecto plantea un análisis sobre la falta de una visualización clara y completa de las tendencias de inversión en TIC por parte de las empresas ecuatorianas durante el periodo comprendido entre 2012 y 2015. Aunque existen datos obtenidos a través de encuestas realizadas por el INEC, se requiere realizar un análisis más detallado y la creación de gráficos que permitan comprender la evolución de esta inversión a lo largo del tiempo y su impacto en el desarrollo empresarial del país.

1.3. Justificación

En esta sección se presenta la justificación teórica, social y técnica que dan apertura y justifican el análisis a realizar como parte del proyecto en cuestión.

1.3.1 Justificación teórica

Con la presente investigación se lograrán obtener los conceptos, historias y teorías adecuadas sobre las TIC. Además, se busca recopilar información sobre las metodologías más utilizadas, como CRISP-DM, y comprender su proceso de aplicación en el análisis de datos. El propósito es despejar cualquier duda que pueda surgir al considerar la viabilidad de aplicar el modelo CRISP-DM en una empresa.

1.3.2 Justificación Social

Es importante tener en cuenta que el uso de las TIC se ha vuelto cada vez más relevante en nuestra sociedad. Estas tecnologías son vitales para el desarrollo y la competitividad de las empresas en un entorno globalizado y digitalizado.

El presente proyecto presentará un análisis que podrá orientar a las empresas en la capacitación de sus empleados en el uso de las TIC, lo que, brindará a los participantes la oportunidad de adquirir conocimientos y habilidades necesarios para utilizar eficientemente estas tecnologías en su entorno laboral.

1.3.3 Justificación Técnica

Esto permitirá conocer la manera de trabajar dentro de la empresa con el uso de las tecnologías de informática y comunicación, dando una presentación factible al proyecto, ya que se cuentan con todos los recursos necesarios para aplicar nuevos modelos de ejecución o manejo de las TIC.

1.4. Objetivos

1.4.1 General

Analizar los resultados de las encuestas realizadas en el Ecuador, sobre la utilización de las TIC para aplicar Big data y modelado de Machine Learning, utilizando la información del INEC.

1.4.2 Específicos

- Contextualizar la información relevante sobre el tema a ser investigado, para la obtención de los datos adecuados que ayuden al desarrollo del presente estudio.
- Analizar los datos obtenidos del INEC sobre las empresas que utilizan TIC, mediante un dashboard elaborado en PowerBI, para conocer el estado de inversiones, uso de equipo y personal utilizado en los años 2012 al 2015.
- Implementar modelos de aprendizaje supervisado para predecir si las empresas necesitan internet, con la herramienta Jupyter notebook.

- Desplegar un modelo de regresión lineal que permita al usuario ingresar los datos y predecir si la empresa requiere contratar personal especialista en TIC.

1.5. Alcance

El estudio de la información resultante de las encuestas del INEC realizadas a las empresas en Ecuador para conocer si utilizan las TIC dentro de sus establecimientos, ayudará a la comprensión de la importancia de contar con una inversión para implementar un modelo de las TIC, a nivel nacional. Es importante conocer sobre el uso de los TIC, sus características para ayudar a las empresas que se dedican a la manufacturación, el comercio y la minería y que son servicios las cuales fueron encuestadas por el INEC a tomar la mejor decisión permitiendo el incremento de los beneficios de las empresas del País en el presente y futuro.

Con la finalización del proyecto se presentará un análisis de los datos obtenidos en el INEC sobre las empresas que utilizan las TIC, mediante una representación gráfica, para establecer una comparación adecuada entre los años 2012 y 2015.

1.6. Marco referencial

En este apartado se delimitará el estado actual de la investigación por medio de la fundamentación legal.

1.6.1 Fundamentación legal

En relación a las Tecnologías de la Información y de las Comunicaciones (TIC), la Constitución del Ecuador, señala en el capítulo segundo, derechos del buen vivir, en las secciones tercera y cuarta, aspectos relativos a las TIC que se debe tener presentes no solo desde el punto de vista del ciudadano, sino también de empresa (Constitución Política de la República Del Ecuador, 2008).

Sección Tercera, Comunicación e Información, Artículo 16: todas las personas en forma individual o colectiva tienen derecho a:

Numeral dos, el acceso universal a las tecnologías de información y comunicación.

Sección cuarta, Cultura y ciencia:

Artículo 25, Las personas tienen derecho a gozar de los beneficios y aplicaciones del progreso científico y de los saberes ancestrales.

En la **Sección quinta**, Acción de hábeas data, se sigue:

Artículo 92, Toda persona, por sus propios derechos o como representante legitimado para el efecto, tendrá derecho a conocer de la existencia y a acceder a los documentos, datos genéticos, bancos o archivos de datos personales e informes que sobre sí misma, o sobre sus bienes, consten en entidades públicas o privadas, en soporte material o electrónico. Asimismo, tendrá derecho a conocer el uso que se haga de ellos, su finalidad, el origen y destino de información personal y el tiempo de vigencia del archivo o banco de datos. Las personas responsables de los bancos o archivos de datos personales podrán difundir la información archivada con autorización de su titular o de la ley.

En la **Sección octava**, Ciencia, tecnología, innovación y saberes ancestrales, se dispone:

Artículo 385, el sistema nacional de ciencia, tecnología, innovación y saberes ancestrales, en el marco del respeto al ambiente, la naturaleza, la vida, las culturas y la soberanía, tendrá como finalidad:

Generar, adaptar y difundir conocimientos científicos y tecnológicos.

Recuperar, fortalecer y potenciar los saberes ancestrales.

Desarrollar tecnologías e innovaciones que impulsen la producción nacional, eleven la eficiencia y productividad, mejoren la calidad de vida y contribuyan a la realización del buen vivir.

Finalmente, en lo que se refiere a la constitución el **Capítulo sexto**, **Derechos de libertad**, en el artículo 66 garantiza:

Numeral 11. El derecho a guardar reserva sobre sus convicciones. Nadie podrá ser obligado a declarar sobre las mismas. En ningún caso se podrá exigir o utilizar sin autorización del titular o de sus legítimos representantes, la información personal o de terceros sobre sus creencias religiosas, filiación o pensamiento político; ni sobre datos referentes a su salud y vida sexual, salvo por necesidades de atención médica.

La **"Ley del Sistema Nacional de Registro de Datos Públicos"**, entra en vigencia desde el 31 de marzo de 2010 y garantiza la seguridad jurídica, organiza, regula, sistematiza e interconecta la información, así como la eficacia y la eficiencia de su manejo, su publicidad, transparencia, acceso e implementación de nuevas tecnologías (Asamblea Nacional, 2010).

Para finalizar este artículo, se menciona la "**Ley de Comercio Electrónico, Firmas Electrónicas y Mensajes de Datos"**. Esta Ley regula los mensajes de datos, la firma electrónica, los servicios de certificación, la contratación electrónica y telemática, la prestación de servicios electrónicos, a través de redes de información, incluido el comercio electrónico y la protección a los usuarios de estos sistemas (Nacional et al., 2002).

Se puede evidenciar que todo lo relacionado anteriormente de las TIC, no son ejecutadas de la manera correcta, se puede decir que falta muchísimo por realizar dentro de las diversas empresas, teniendo en cuenta las faltas o carencias, entonces queda en la actualidad, el cumplimiento de las diversas participaciones y leyes en donde interviene las participaciones activas para poderlas mejorar.

Es fundamental tener presente, que todos los días, este tema se presenta a diario porque se tratan de asuntos transversales a la actividad humana y el espíritu que se posee para regular

los contenidos, esto se basa en conformidad a las maduraciones de las sociedades en torno a lo que significa para el propio ser las TIC, en donde estas garantizan los accesos, a las protecciones de los consumidores, usuarios, empresas, en base a las responsabilidades, en las creaciones de métodos y medios de los comercios o tecnologías la que garantice la utilización de la información que estén orientados a impedir daños en sus activos intangibles o tangibles.

Por ende, el uso de estas protegen a los personas e instituciones contra los fraudes informáticos o los espionajes industriales, para la prevención de los daños físicos de la infraestructura crítica en los casos del sistema de los tiempos reales, en donde se cuidan la integridad y salud de las personas frente al sistema médico crítico, como puede ser aquel en donde se emplea herramientas con tecnología radiante en donde se incorpora los software y la electrónica, en donde se propicia los diversos tratamientos responsables de información, evitar los malos usos de las TIC y también el respeto de la privacidad, teniendo como ejemplos el crimen o delito electrónico.

2 MARCO TEÓRICO

En el presente capitulo se detalla la recopilación de conceptos, basado en información que ayudará a sustentar el proyecto. Se ha seleccionado varios conceptos basados en el tema de estudio, las herramientas tecnológicas a usar y se los detalla a continuación:

2.1. Datos

Son las representaciones de las variables que pueden ser cuantitativas o cualitativas que indican los valores que se les va asignando a los objetos y se representan por medio de las secuencias de los números, letras o símbolos (Etecé, 2022) .

Por otra parte, el dato entorno de la informática, son las expresiones generales que describen diversas características de las entidades sobre las que operan, estos son aplicaciones o programas que poseen las funciones de los procesamientos de datos, porque cada lenguaje de programaciones adquiere conjuntos con los datos a partir de los que se trabajan. Todas las informaciones que salen y entran a los ordenadores lo van haciendo de manera de datos.

Se puede decir que dentro de estos archivos van existiendo datos que son los paquetes más diminutos de otros datos denominados registros, estos se van reuniendo por las diferentes características las cuales resulten set similares o iguales.

2.2. Dataset

Conjuntos de datos, que son ordenados bajo los sistemas de almacenamientos que otorgan los diferentes lineamientos que son principales para las búsquedas o directorios de las diversas informaciones que se quieren trabajar (Caceres, 2023).

Los contenidos de las tablas dentro de las bases de los datos poseen diversas columnas, en donde se va a ir acumulando los registros. Estas se pueden denominar la categoría de los datos, y, por ende, las columnas, serán las posibles variables que las conforman. Estas uniones entre la columna y fila, es lo que se va a llamar *Dataset*.

2.3. Calidad de datos

Se refieren a los grados en que los datos cumplen con los estándares de precisión, completitud, consistencia, integridad, actualidad y relevancia para un propósito determinado. Mejorar la calidad de los datos implica implementar medidas de control de disposición, como la validación, la normalización, la eliminación de duplicados y la gestión de la eficacia de los datos. Estas acciones buscan garantizar que los datos sean confiables, coherentes y pertinentes, lo que a su vez mejora la toma de decisiones, el análisis y los resultados obtenidos en diferentes contextos y aplicaciones (Deloitte, 2021).

2.4. Recolección de datos

Estrategias mediante las cuales las organizaciones recaban y analizan información proveniente de diversas fuentes con el objetivo de obtener una visión integral, abordar preguntas cruciales, evaluar sus logros y anticipar tendencias futuras (Santos, 2022).

2.5. Limpieza de Datos

Denominada como "*Data Cleansing* o *Data Scrubbing*", engloban diversos procedimientos, predestinados para el mejoramiento de la calidad de los datos. Teniendo en cuenta que existen diversas prácticas y herramientas para la eliminación del problema de los conjuntos de datos (Datascientest, 2023).

Este proceso se basa en la corrección o eliminación de registros inexactos en las bases de datos o en los *dataset*. Los conjuntos de datos deben ser coherentes y libres de errores. Esta parte es esencial para la utilización y la explotación de los datos. Sin una limpieza de datos adecuada es probable que los resultados del análisis sean distorsionados e incorrectos.

Además, permite transformaciones de datos uniformes y consistentes que involucran tipos de formato apropiados, lo que se puede lograr mediante la integración de múltiples fuentes de datos. Dado que el resultado de cualquier análisis de datos depende en grandes

medidas de la calidad y consistencia de los datos, se debe mencionar que la preparación de datos es un paso crítico en cualquier análisis de datos.

2.6. Transformación de Datos

Son los procesos de convertir datos sin procesar de un formato a otro para que puedan ser utilizados por un sistema o aplicación de destino. Este proceso implica varios pasos, como el filtrado de datos según reglas predefinidas y la combinación de diferentes campos para obtener una vista unificada. Las herramientas de transformación de datos juegan un papel clave para hacer que este proceso sea eficiente y flexible (Sudor, 2020).

2.7. Machine Learning

Comprende un conjunto de algoritmos de aprendizaje automático que trabajan con grandes cantidades de datos y utilizarlos para la identificación de los patrones. Hiff (2021) hace referencia a dos distintas definiciones para la contextualización del aprendizaje automático, donde manifiesta que los "algoritmos de *machine learning* son el alma que mueven los procesos de aprendizaje. Gracias a ellos podemos obtener la información que necesitamos para tomar decisiones o predecir el comportamiento de los datos" (p.23).

En el ámbito de la inteligencia artificial y ciencia de datos, *Machine Learning* tiene como objetivo principal adaptar los datos a modelos que puedan ser interpretados y utilizados en diversos campos especializados. El aprendizaje automático intenta comprender cómo se organizan los datos y encontrar patrones importantes para desarrollar modelos que permitan a los profesionales de la industria usarlos de manera efectiva en diferentes sectores de trabajo (Mancilla, 2022).

2.8. Análisis predictivo

El análisis predictivo implica el examen y la interpretación de conjuntos de datos utilizando métodos estadísticos y algoritmos para descubrir patrones y predecir procesos o fenómenos.

Por ejemplo, con la analítica predictiva es posible estimar los comportamientos futuros de los clientes, el volumen de ventas de las empresas o las tendencias en un sector del mercado. El método se basa en analizar datos históricos e identificar relaciones causales o correlaciones entre variables para poder realizar predicciones y anticipar posibles escenarios futuros.

Utilizando herramientas y modelos de análisis predictivo, se pueden tomar decisiones estratégicas más informadas para optimizar los resultados y el rendimiento en campos tan diversos como los negocios, las finanzas o la ciencia (Cali, 2022).

Uno de los algoritmos de *machine learning* más utilizados para el análisis predictivo es la regresión lineal. El análisis de la regresión lineal se utiliza para predecir el valor de una variable según el valor de otra. La variable que desea predecir se denomina variable dependiente. La variable que está utilizando para predecir el valor de la otra variable se denomina variable independiente. Estos modelos analizan las relaciones existentes entre las variables dependientes y los conjuntos de variables independientes. Estas relaciones se expresan como las ecuaciones que predicen las variables de las respuestas como las funciones lineales de los diversos parámetros, los que se ajustan para que las medidas para ajustarse en optimización (IBM, 2022).

2.9. Capacidades analíticas

Las implicaciones de las capacidades de los análisis de los datos, la identificación del patrón, el establecimiento de las extracciones y las conexiones, para la determinación de la conclusión fundamentada a partir de las informaciones disponibles que permiten la toma de las decisiones que son las más convenientes y así el logro de los resultados que sean superiores.

Estos van implicando las capacidades de los análisis de los datos, en donde se identifica el patrón, de los establecimientos de la conexión y la extracción de la conclusión que es

fundamentada a partir de las informaciones disponibles. En los desarrollos de las capacidades analíticas, en donde son capaces para la realización de los diversos análisis exhaustivos y los objetivos de las situaciones de forma compleja, las evaluaciones múltiples la perspectiva y la opción de adquirir la decisión informada y estratégica (Perez, 2019).

2.10. Gráficos estadísticos

Los gráficos estadísticos son herramientas para visualizar datos y simplificar información compleja de una manera fácil de entender. Su función es brindar información de formas claras y precisas para que los usuarios puedan comparar y comprender la evolución de las diversas variables. Los gráficos estadísticos proporcionan una representación visual que simplifica la interpretación de los datos y facilitan los análisis de patrones y tendencias (Arteaga et al., 2021).

Utilizan principalmente elementos gráficos como barras, líneas, puntos y cortes para representar datos de una manera visualmente atractiva y accesible. Los gráficos estadísticos ayudan a resumir y organizar grandes conjuntos de datos al presentar la información visualmente para que el usuario pueda identificar fácilmente las diferencias, las similitudes y los cambios en los valores.

2.11. Visualización de datos

Según (Aprendeconalf, 2023), estos implican a las presentaciones de las informaciones utilizando objetos como mapas, gráficos y otros formatos visuales para facilitar que la mente humana comprenda y absorba información relevante.

A menudo, este término se utiliza de manera intercambiable con otros conceptos, como gráficos de información, visualización de información y gráficos estadísticos, todos ellos relacionados con la presentación visual de datos para facilitar su análisis y comprensión. El propósito principal de la visualización de datos es ayudar a identificar el patrón, tendencia y valor atípico en conjuntos de datos extensos.

2.12. Evaluación de modelos

Las evaluaciones de modelos son los tipos de procesos que son fundamentales en el campo de los aprendizajes automáticos, que se encarga de realizar una medición de la capacidad de un sistema de tipo *machine learning* que tenga la facultad de generalizar a nuevos datos. De esta manera este proceso de evaluación se efectúa mediante la utilización de conjuntos de datos que sean considerados de prueba y que adicionalmente no hayan sido utilizados para entrenar el modelo. Por lo cual la aplicación de la evaluación permite aumentar su precisión, ajustar la tasa de error e incrementar la capacidad de realizar múltiples tipos de predicciones eficientes y correctas en contexto de situaciones reales (Arias et al., 2019).

2.13. Herramientas

A continuación, se presenta la Tabla I con las herramientas a utilizar en el proyecto con su respectiva Justificación ya que estos materiales son indispensables que nos ayudan hacer un análisis más rápido

Tabla I. *Herramientas a utilizar*

Herramienta	Justificación
Python	Se utilizará para el análisis de los datos obtenidos creando y gestionando de forma rápida los datos, ofreciendo una manipulación y representación adecuada para datos complejos.
Numpy	Esto nos permite la realización de cálculos lógicos y matemáticos de alta complejidad y sobre los cuadros y matrices.
Pandas	Se lo utiliza por la facilidad de uso gracias a su sintaxis clara y concisa. Y contiene una gran funcionalidad y eficiencia puede comunicarse con otras.

Matplotlib	La razón de utilizar el paquete matplotlib es un beneficio de extensa biblioteca de funciones para generar gráficos 2d y 3d, ofreciendo la funcionalidad de generar dos interfaces de programación en pocas palabras, dos estilos distintos para crear y personalizar las gráficas un estilo denominado Matlab luego se hereda a la interfaz de este software.
Scikit-learn	Ayuda a proporcionar el acceso a adaptaciones poderosos de muchos algoritmos comunes, proporcionando una API propia y estandarizada.
Jupyter Notebook	Con este programa ayudara a establecer una interfaz web de código abierto, permitiendo la inserción de vídeo, audio, texto, imágenes, y también ayuda a la realización de código a través del navegador en múltiples lenguajes.

Nota. Elaboración propia (2023).

3 METODOLOGÍA

Se utilizará este enfoque, porque, se completarán para realizar la preparación de la tabla de datos, que luego será gráfica, esto se realiza a través de la adquisición de datos utilizando diversas técnicas, para interpretar los datos recopilados y sacar conclusiones sobre los datos de investigación adquiridos.

3.1 Método de Investigación

La modalidad de este estudio que se llevará a cabo es de, un enfoque mixto, esto quiere decir que se realizara la investigación cualitativa y cuantitativa las cuales permitirán recaudar información de primera mano o directa.

3.1.1 Investigación Cuantitativa

Se utiliza el presente enfoque, porque los datos que se obtendrán en el Instituto Nacional de Estadísticas y censo serán clasificados, ordenados y tabulados, serán obtenidos en la ML.

3.1.2 Investigación Analítica

Se le puede utilizar el método analítico ya que permitirá la descripción correspondiente sobre los datos obtenidos para su análisis final.

3.2 Metodología de Análisis de Datos

Para la realización del presente proyecto se realiza un análisis de datos para la exploración, limpieza, transformación y modelado de conjuntos de datos con el objetivo de descubrir información útil para la elaboración del proyecto.

A continuación, en la Tabla II es de una comparación de tres metodologías.

Tabla II. *Comparación de algunos métodos*

Descripción	Knowledge Discovery in Database	Sample, Explore, Modify Model and Access	Cross-Industry Standard Process for Data Mining

Enfoque	Orientado a la identificación de patrones más favorables.	Orientación al desarrollo del proceso de MD o minería de datos.	Orientación a los objetivos empresariales.
Uso	Productos enfocados en patrones de datos	Ligado a productos SAS	Metodología abierta y gratuita
Metodología	Metodología de patrones arquitectónicos orientados a datos.	Metodología aún no definida	Metodología de gestión de proyectos
Complejidad	Más complejo de implementar que los otros dos, tiene una cantidad considerable de fases a desarrollar.	Simple y bastante ágil sus fases están más implementadas a desarrollo ágil.	Es menos complejos de entender y aplicar, cuentas con una curva de adaptabilidad muy amplia para cualquier analista de datos.
Nro. de fases de desarrollo	9	6	6
Siglas	KDD	SEMMA	CRISP-DM

Relevancia actual	Baja	Media	Alta

Nota. Elaboración propia (2023).

En el presente proyecto se utilizará la metodología *Cross-Industry Standard Process for Data Mining* (CRISP-DM), por lo que se mostrará la información obtenida de una manera de fácil comprensión, permitiendo crear modelos de aprendizaje automático adaptados a las necesidades específicas del cliente, proporcionando una representación del ciclo de vida del proyecto con su debido análisis de datos.

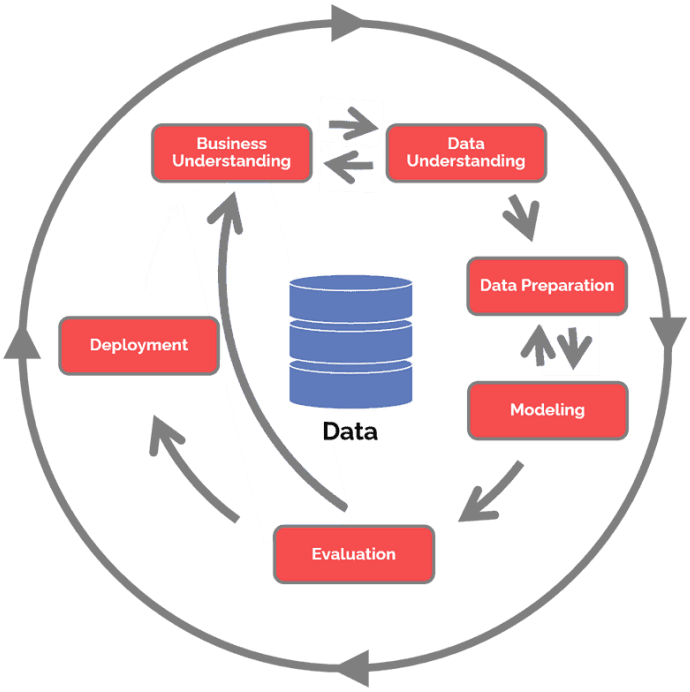


Figura 1. Modelo CRISP-DM

Nota. Información obtenida de (Hotz, 2023).

Para un mejor entendimiento de la metodología *Cross-Industry Standard Process for Data Mining,* se presenta una explicación.

3.2.1 Descripción del modelo CRISP-DM

Para realizar la respectiva aplicación del modelo de acuerdo a Galán (2019) se debe realizar los siguientes pasos:

Comprensión del negocio (Objetivos y requerimientos desde una perspectiva no técnica)

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)
- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio).
- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

Comprensión de los datos (Familiarizarse con los datos teniendo presente los objetivos del negocio)

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

Preparación de los datos (Obtener dataset) Selección de los datos

- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

Modelado (Aplicar las técnicas de minería de datos a los dataset)

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

Evaluación (De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio)

- Evaluación de resultados
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones

Despliegue (Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización)

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final
- Revisión del proyecto

Con la información obtenida anteriormente se puede visualizar cuales son los pasos para utilizar el modelo CRISP-DM, el cual permitirá establecer un parámetro adecuado para seguir la secuencia de su utilización, obteniendo buenos resultados y sin pérdidas de tiempo al momento de llevar a su ejecución.

Cada uno de sus aspectos son fundamentales al momento de realizar la presente investigación.

3.2.2 Metodología CRISP-DM de cómo se va a realizar la investigación

Comprensión del negocio

El Instituto Nacional de Estadística y Censos (INEC) presenta los principales resultados, sobre el estado de las TIC en el país, mediante la identificación del valor de la inversión, características y uso del internet, comercio electrónico y uso de equipos tecnológicos obtenidos de los sectores de manufactura, minería, comercio y servicios a partir de las encuestas industriales de los años 2012 al 2015. Convirtiéndose en una fuente adecuada para solucionar un problema mediante la aplicación de reglas técnicas que permitan transformar los

conjuntos de datos. De acuerdo con los objetivos planteados anteriormente, el presente estudio consiste en identificar los factores que influyen en la toma de decisiones, verificando la utilidad de la información tomada de la base de datos del INEC para el bienestar de la investigación.

Comprensión de datos

Se recopila los datos establecidos para poder realizar la investigación sobre las TIC en el entorno empresarial del Ecuador, datos obtenidos por el INEC.

Preparación de los datos

- Establecer el universo de los datos con los que vamos a trabajar.
- Construiremos un dataset, apto para ser usados en los modelos de minería de datos.
- Realizaremos tareas de limpieza de datos, mediante la transformación, supresión y formateo de datos.
- Se consideran aspectos relevantes como son:

  Año investigado, sector económico, , tamaño de la empresa, provincia de establecimiento, inversiones, equipos utilizados, acceso a internet, personal.

Modelado

En toda esta etapa se deben utilizar las técnicas y librerías, pues en esta indagación se encuentran:

- Librería NumPy
- Librería Panda
- Librería Matplotlib
- Seleccionar técnicas de modelado más adecuadas para nuestro juego de datos y nuestros objetivos.
- Fijar una estrategia de verificación de la calidad del modelo.
- Construiremos un modelo a partir de la aplicación de las técnicas seleccionadas sobre la base de datos.

- Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriores establecidos.

Evaluación del modelado

Revisar que los resultados obtenidos en base a los objetivos planteados, en todo el proceso de minería de datos que nos ha llevado hasta este punto.

Despliegue

Realizar seguimiento y mantenimiento de la parte más operativa del estudio.

Al final en esta última etapa se va a colocar el modelo en presente, el modelo debe ser utilizado en un ambiente activo, tal como se logra medir la eficiencia del modelo y que logra estar en constante actualización en caso de recordar algún ajuste. En el futuro siendo traicionero. En caso de optar va a depender de las situaciones que se presenten, en la implementación para el entorno del sector manufacturero, minería y comercio electrónico, es fundamental establecer que el modelo cumple con todos los requisitos para satisfacer un problema Business, en que el grado tiene la posibilidad de tomar decisiones en tiempo real y evitar perdida de dinero o recursos e incluso mejorar el sistema para garantizar que la organización tenga una comprensión eficiente, y otro elemento que también se debe tener en cuenta es monitorear el modelo para obtener ciertas actualizaciones y mejoras en función de los nuevos avances en el área de las ciencias de datos que vayan surgiendo.

4. IMPLEMENTACIÓN Y RESULTADOS

La implementación de la metodología CRISP-DM se la realizara paulatinamente aplicando cada una de sus fases, utilizamos el dataset conseguido de la base de datos del INEC.

4.1. Fase 1 – 2

El módulo de TIC´S del formulario ENESEM (Encuesta Estructural Empresarial), nos brindara la información necesaria para realizar un análisis del contenido. Los datos que contiene el formulario son otorgados por las empresas que pertenecen a diversos sectores empresariales del Ecuador tanto públicas como privadas y contiene varios módulos siendo objeto de estudio el módulo de TIC'S, en esta fase de la metodología unificaremos la comprensión del negocio y comprensión de los datos.

En los periodos de referencia (años 2012 al 2015) tenemos un muestreo de 6080 empresas registradas en el Directorio de Empresas y Establecimientos (DIEE). La data obtenida del módulo mencionado cuenta con alrededor de 68 columnas que corresponden a variables de tipo cualitativo y cuantitativo después de realizar un análisis se procede con el primer descarte de columnas (variables) por tratarse de datos de categorización o redundantes en la información por tanto se consideran datos irrelevantes para ser analizados.

A continuación, en la Tabla III**Tabla *II*** tenemos la lista de varias descartadas por no tener relevancia para el análisis a realizar.

Tabla III. *Lista de variables descartadas*

Campo	Descripción
ciiu_seccion	Código de la actividad - Sección de la CIIU
ciiu_division	Código de la actividad - División de la CIIU
ciiu_clase	Código de la actividad - Clase de la CIIU
tic36_otros	Utilizo otros tipos dispositivos tecnológicos, cuantos
tic4_personal_total_comp	Personal que utilizó computadoras en su rutina normal

tic41_personal_comp_m	Personal que utilizó computadoras – mujeres
tic42_personal_comp_h	Personal que utilizó computadoras – hombre
tic7c_otros_1412	Tubo otros tipos de conexión a internet
tic9_transacciones	Realizó transacciones comerciales a través de internet
tic91_total_compras	Compró productos o servicios por internet
tic92_total_ventas	Vendió productos o servicios por internet
tic10_interaccion_adm_pub	Interacciono con la administración pública por internet
tic101_obtener_info_publicas	Obtuvo información a través de las páginas web de las Administraciones Públicas
tic102_impreso_publicos	Consiguió impresos o formularios de las páginas web de las Administraciones Públicas
tic103_devolver_impresos	Uso internet para devolver impresos finalizados
tic104_gestion_electronica	Uso internet para realizar gestión electrónica completa, sin necesidad de ningún trámite adicional en papel
tic105_declaracion_impuestos	Uso internet para declarar impuestos
tic106_tramites_iess	Uso internet para realizar trámites vinculados con el IEES
tic107_portal_compras_publicas	Uso internet para acceder al Portal de Compras Públicas con el fin de participar en una licitación
tic111_gestion_cliente	Negocio electrónico - Gestión con los clientes
tic112_control_pedidos	Negocio electrónico - Control y seguimiento de pedidos
tic113_gestion_inventarios	Negocio electrónico - Gestión de la cadena de suministros, logística, control de inventarios
tic114_gestion_finanzas	Negocio electrónico - Gestión de finanzas y presupuestos
tic115_gestion_rrhh	Negocio electrónico - Gestión de los recursos humanos
tic116_soporte_ventas	Negocio electrónico - Servicio y soporte de ventas

tic117_id	Negocio electrónico - Investigación y desarrollo
tic118_gestion_conocimiento	Negocio electrónico - Gestión del conocimiento
tic124_mensajeria_especializada	Medios de comunicación - Mensajería especializada
tic125_fax	Medios de comunicación – Fax
tic126_call_center	Medios de comunicación - Call center
tic127_otros_1434	Medios de comunicación - Otros
tic128_ninguno	Medios de comunicación - Ninguno
tic131_sistemas_operativos	Software Libre - Sistemas Operativos (LINUX)
tic132_navegador_internet	Software Libre - Navegadores de Internet (Mozilla Firefox, Chrome, Opera, Safari)
tic133_aplicaciones_ofimaticas	Software Libre - Aplicaciones ofimáticas (Open Office)
tic134_erp_crm	Software Libre - Aplicaciones de código abierto para el procesamiento automático de información del tipo ERP
tic135_seguridad	Software Libre - Otras, como software de seguridad (Open SSL, SSH), plataformas de aprendizaje (Moodle)
tic17_firma_digital	La empresa utilizo firma digital en comunicaciones
k	Número de empresas

Nota. Elaboración propia (2023).

Después de entender la data que tenemos y definir las variables que vamos a aplicar los modelos ML procedemos con el primer descarte obteniendo una nueva data con 28 variables que usaremos, para la aplicación de modelos de ML de aprendizaje supervisado como también la creación de un dashboard en Power BI que nos mostrara información sobre el uso de Internet, uso de equipos tecnológicos, personal que conoce de TIC'S e inversiones realizadas por las empresas.

A continuación, en la Tabla IV**Tabla *III*Tabla *II*** tenemos la lista de varias depurada con las cuales se trabajará en la siguiente fase.

Tabla IV. *Lista de variables depuradas*

Campo	Descripción
año	Año de investigación
id_diee	Código de Identificación Directorio de Empresa
sector_economico	Sector Económico
tam_empresa	Tamaño de empresa
Provincia	Código de provincia
tic1_inversion	Hizo inversiones la empresa TIC
tic2_valor_inversion	Valor invertido en TIC
tic31_computadoras	Computadoras que tiene la empresa
tic32_pda	PDA que tiene la empresa
tic33_smartphone	Smartphone que tiene la empresa
tic34_notebooks	Notebooks que tiene la empresa
tic35_tablets	Tablet que tiene la empresa
tic5_conexion_internet	Dispone la empresa conexión a internet
tic6_personal_total_int	Personal ocupado que utilizó internet
tic61_personal_int_m	Trabajadoras mujeres que utilizó internet
tic62_personal_int_h	Trabajadoras hombres que utilizó internet
tic7a_bandaanchafija	Conexión a internet utilizado - Banda Ancha Fija
tic7b_bandaanchamovil	Conexión a internet utilizado - Banda Ancha Móvil
tic8_web	Su empresa tiene sitio o página web
tic121_telefonos_celulares	Medios de comunicación - Teléfonos celulares
tic122_correo_electronico	Medios de comunicación - Correo electrónico

tic123_redes_sociales	Medios de comunicación - Redes Sociales
tic14_especialistas_tic	Tiene contratado personal especialista en TIC
tic141_especialistas_tic_m	Personal especialista en TIC - mujeres
tic142_especialistas_tic_h	Personal especialista en TIC - hombres
tic151_personal_conoc_tic	Número total de personal con conocimientos en TIC
tic151_personal_conoc_tic_m	Personal con conocimientos en TIC - mujeres
tic152_personal_conoc_tic_h	Personal con conocimientos en TIC – hombres
tic16_intranet	La empresa conto con intranet

Nota. Elaboración propia (2023).

Una vez que definimos las variables de nuestra data realizamos una evaluación en el cuadernillo de Jupyter donde leemos el archivo de Excel e imprimimos para visualizar la muestra de los datos de esta manera verificamos si no tenemos algún error en la data, no debemos olvidar que para poder trabajar con DataFrame, listas, arreglos, gráficos y para poder dividir, modelar, entrenar y evaluar los modelos; debemos importar las librerías necesarias en nuestro caso debemos importar las librerías:

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- LogisticRegression
- GaussianNB
- LinearRegression
- train_test_split
- minmax_scale
- metrics

- classification_report
- confusion_matrix
- StandardScaler
- accuracy_score,
- f1_score
- precision_score

Cuando ya importamos nuestras librerías podemos trabajar con nuestro dataset

```
tesis_df = pd.read_excel('data_tesis.xlsx')
tesis_df
```

	año	id_diee	Sect_Econ	tam_empresa	provincia	tic2_valor_inversion	tic31_com
0	2012	13582534017	SERVICIOS	PEQUEÑA EMPRESA	AZUAY	0.0	
1	2013	13582673017	SERVICIOS	PEQUEÑA EMPRESA	AZUAY	3634.0	
2	2014	13582673017	SERVICIOS	PEQUEÑA EMPRESA	AZUAY	0.0	
3	2015	13582673017	4	2	101	3064.0	
4	2013	13582772015	MANUFACTURA	PEQUEÑA EMPRESA	AZUAY	0.0	
...	...	...	...	...	...	...	
15066	2015	46786977223	3	3	322	0.0	
15067	2015	46789417059	1	3	105	0.0	
15068	2015	46826132183	4	2	118	746.0	
15069	2015	46873591188	1	4	118	29000.0	
15070	2015	47008897090	1	5	209	94365.0	

15071 rows × 25 columns

Figura 2. Lectura e impresión del archivo de Excel (Dataset)

Con el método .info() visualizamos un resumen informativo de nuestro DataFrame que incluye el nombre de cada columna, el número de filas y columnas, que tipo de datos contienen las columnas, los datos no nulos, como también la cantidad de memoria usada por el DataFrame, la información se muestra para tener una visión de la calidad y la estructura de los datos.

```
tesis_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15071 entries, 0 to 15070
Data columns (total 29 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   año                          15071 non-null  int64
 1   id_diee                      15071 non-null  int64
 2   Sect_Econ                    15071 non-null  object
 3   tam_empresa                  15071 non-null  object
 4   provincia                    15071 non-null  object
 5   tic1_inversion               15071 non-null  object
 6   tic2_valor_inversion         15071 non-null  float64
 7   tic31_computadoras           15071 non-null  int64
 8   tic32_pda                    15071 non-null  int64
 9   tic33_smartphone             15071 non-null  int64
 10  tic34_notebooks              15071 non-null  int64
 11  tic35_tablets                15071 non-null  int64
 12  tic5_conexion_internet       15071 non-null  object
 13  tic6_personal_total_int      14960 non-null  float64
 14  tic61_personal_int_m         10832 non-null  float64
 15  tic62_personal_int_h         10832 non-null  float64
 16  tic7a_bandaanchafija         14960 non-null  object
 17  tic7b_bandaanchamovil        14960 non-null  object
 18  tic8_web                     14960 non-null  object
 19  tic121_telefonos_celulares   15071 non-null  object
 20  tic122_correo_electronico    14960 non-null  object
 21  tic123_redes_sociales        14960 non-null  object
 22  tic14_especialistas_tic      15071 non-null  object
 23  tic141_especialistas_tic_m   8961 non-null   float64
 24  tic142_especialistas_tic_h   8961 non-null   float64
 25  tic151_personal_conoc_tic    15071 non-null  int64
 26  tic151_personal_conoc_tic_m  15071 non-null  int64
 27  tic152_personal_conoc_tic_h  15071 non-null  int64
 28  tic16_intranet               15071 non-null  object
dtypes: float64(6), int64(10), object(13)
memory usage: 3.3+ MB
```

Figura 3. Información del DataFrame

El método .describe() nos presenta un resumen estadístico de cada variable (columna) de nuestro DataFrame, esta visualización nos proporciona la facilidad de detectar valores atípicos o inusualmente altos o bajos.

```
tesis_df.describe()
```

	año	id_diee	tic2_valor_inversion	tic31_computadoras	tic32_pda
count	15071.000000	1.507100e+04	1.507100e+04	15071.000000	15071.000000
mean	2013.407339	1.562120e+10	5.095299e+04	48.223476	1.671356
std	1.104505	6.365339e+09	7.557917e+05	244.412819	24.939544
min	2012.000000	1.358253e+10	0.000000e+00	0.000000	0.000000
25%	2012.000000	1.370831e+10	0.000000e+00	5.000000	0.000000
50%	2013.000000	1.382739e+10	0.000000e+00	12.000000	0.000000
75%	2014.000000	1.473857e+10	6.570500e+03	30.000000	0.000000
max	2015.000000	4.700890e+10	5.400916e+07	8500.000000	1403.000000

Figura 4. Método .describe()

Con la ayuda de la librería seaborn y la línea de código sns.heatmap(tesis_df.isna()), podemos obtener una gráfica que nos muestra la ubicación de los datos nulos, ayudando a tener una apreciación más clara del estado de nuestros datos.

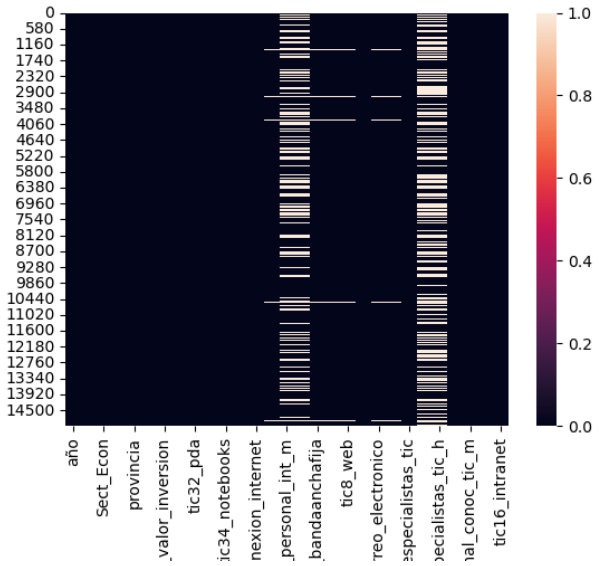


Figura 5. Datos nulos

Dentro de la librería seaborn tenemos varios gráficos que nos facilita la evaluación de nuestros datos de forma visual de esta manera vamos comprobando como están compuestos nuestros registros.

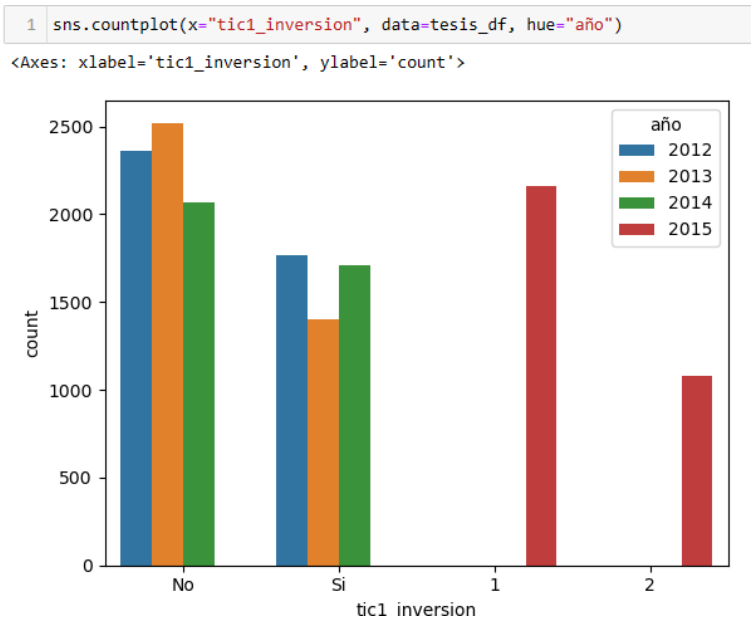


Figura 6. Inversiones de las empresa por año

```
1 sns.countplot(x="Sect_Econ", data=tesis_df)
```

```
<Axes: xlabel='Sect_Econ', ylabel='count'>
```

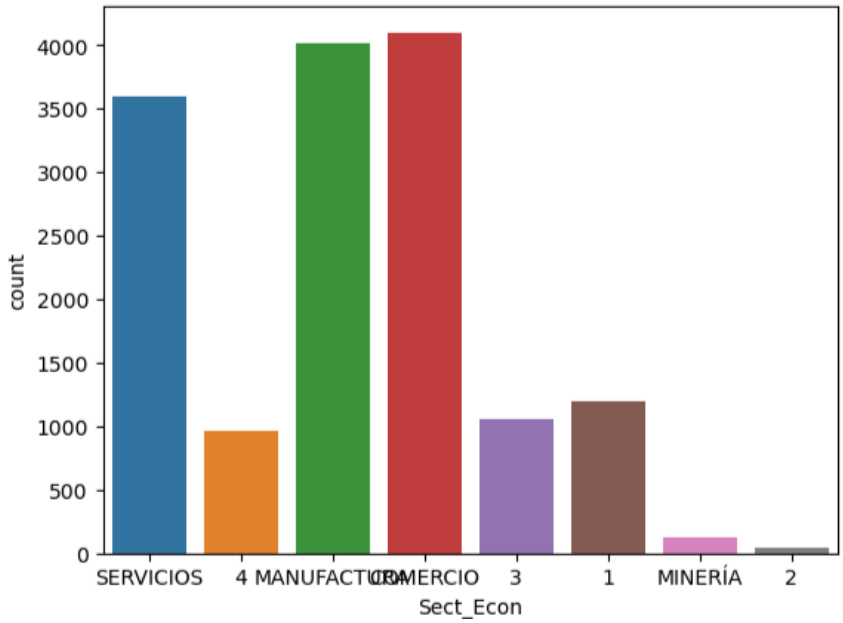


Figura 7. Cantidad de empresas por sector económico

4.2. Fase 3

Para iniciar con la limpieza de datos se realiza una nueva evaluación de las variables y el necesario cambio de nombres de las mismas, después de hacer un análisis de los registros y encontrar columnas con datos nulos que no se pueden normalizar o completar, eliminamos dichas columnas que son:

'id_diee','tic62_personal_int_h',

'tic61_personal_int_m',

'tic141_especialistas_tic_m',

'tic142_especialistas_tic_h'

Las columnas eliminadas tienen una gran cantidad de datos nulos como podemos observar en la siguiente grafica aplicando el código tesis_df.isna().sum().sort_values().

```
# Conteo de datos nulos
tesis_df.isna().sum().sort_values()
```

```
año                              0
tic151_personal_conoc_tic_m      0
tic151_personal_conoc_tic        0
tic14_especialistas_tic          0
tic121_telefonos_celulares       0
tic152_personal_conoc_tic_h      0
tic5_conexion_internet           0
tic35_tablets                    0
tic34_notebooks                  0
tic16_intranet                   0
tic32_pda                        0
id_diee                          0
Sect_Econ                        0
tam_empresa                      0
tic33_smartphone                 0
provincia                        0
tic1_inversion                   0
tic2_valor_inversion             0
tic31_computadoras               0
tic7a_bandaanchafija           111
tic7b_bandaanchamovil          111
tic8_web                       111
tic122_correo_electronico      111
tic123_redes_sociales          111
tic6_personal_total_int        111
tic62_personal_int_h          4239
tic61_personal_int_m          4239
tic141_especialistas_tic_m    6110
tic142_especialistas_tic_h    6110
dtype: int64
```

Figura 8. Datos nulos

Una vez eliminado los datos encontramos la necesidad de realizar los cambios de nombres de las variables que serán objeto de nuestro estudio, mediante la línea de código tesis_df=tesis_df.rename(columns={'Sect_Econ':'Sector_Económico',

'tam_empresa':'Tamaño_Empresa', 'provincia':'Provincia',

'tic1_inversion':'Inversion_Tic','tic2_valor_inversion':'Valor_Inversion',

'tic31_computadoras':'Computador_cant','tic32_pda':'Pda_cant',

'tic33_smartphone':'Celular_cant','tic34_notebooks':'Notebooks_cant',

'tic35_tablets':'Tablet_cant','tic5_conexion_internet':'Conexion_Internet',

'tic6_personal_total_int':'Personal_q_uso_Internet',

'tic7a_bandaanchafija':'Internet_ba_Fija',

'tic7b_bandaanchamovil':'Internet_ba_Movil',

'tic121_telefonos_celulares':'Uso_Celular',

'tic122_correo_electronico':'Uso_Email',

'tic123_redes_sociales':'Uso_R_Sociales',

'tic14_especialistas_tic':'Especialistas_tic',

'tic151_personal_conoc_tic':'Personal_Conocimiento_tic',

'tic151_personal_conoc_tic_m':'Mujeres_Conocimiento_tic',

'tic152_personal_conoc_tic_h':'Hombres_Conocimiento_tic',

'tic16_intranet':'Intranet','tic8_web':'Pag_Web'}).

En el análisis de nuestros datos se define que también es necesario unificar el tipo de datos de las columnas, al encontrar datos tipo objeto y de tipo entero en la misma columna en las variables, 'Sector_Económico',

'Tamaño_Empresa',

'Provincia',

'Inversion_Tic',

'Conexion_Internet',

'Internet_ba_Fija',

'Internet_ba_Movil',

'Pag_Web',

'Uso_Celular',

'Uso_Email',

'Uso_R_Sociales',

'Especialistas_tic',

'Intranet'

Esta unificación de datos se realiza con la función .replace que tiene la siguiente estructura: “nombre_df”[“columna”] = “nombre_df” ['columna'].replace({“v_1”: “v_2”, }),

según sea necesario como por ejemplo en la variable 'Inversion_Tic' que su código es tesis_df['Inversion_Tic'] = tesis_df['Inversion_Tic'].replace({2: 0, 'Si': 1, 'No': 0}). Teniendo que reemplazamos el valor 2 por 0 y las palabras si por 1 y no por 0, obteniendo como resultado datos enteros de 0 y 1 que corresponde a falso y verdadero respectivamente.

Después de reemplazar datos para unificarlos, completamos los datos de los registros faltantes ya que en el análisis de la data se observa que dicho registros faltantes se puede complementar en base al dato de otra columna, ejemplo si no tiene internet por ende no tiene conexión de banda ancha fija, para realizar esta actividad creamos una función y su código es:

```
def sin_internet(columnas): # columnas Conexion_Internet y Internet_ba_Fija
    ba_fija = columnas[0]
    internet = columnas[1]
    if pd.isnull(ba_fija):
        if internet == 0:
            return 0
    else:
        return ba_fija
# invocar la función sin_internet
tesis_df["Internet_ba_Fija"]=tesis_df[[
"Internet_ba_Fija","Conexion_Internet"]].apply(sin_internet, axis=1)
```

Para corroborar que todo el proceso de limpieza de datos se realizado correctamente vamos a mostrar unas imágenes en las cuales podemos observar uniformidad en el tipo de dato y que ya no existen datos nulos.

```
1 tesis_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15071 entries, 0 to 15070
Data columns (total 24 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   año                         15071 non-null  int64
 1   Sector_Económico            15071 non-null  object
 2   Tamaño_Empresa              15071 non-null  object
 3   Provincia                   15071 non-null  object
 4   Inversion_Tic               15071 non-null  int64
 5   Valor_Inversion             15071 non-null  float64
 6   Computador_cant             15071 non-null  int64
 7   Pda_cant                    15071 non-null  int64
 8   Celular_cant                15071 non-null  int64
 9   Notebooks_cant              15071 non-null  int64
 10  Tablet_cant                 15071 non-null  int64
 11  Conexion_Internet           15071 non-null  int64
 12  Personal_q_uso_Internet     15071 non-null  int64
 13  Internet_ba_Fija            15071 non-null  int64
 14  Internet_ba_Movil           15071 non-null  int64
 15  Pag_Web                     15071 non-null  int64
 16  Uso_Celular                 15071 non-null  int64
 17  Uso_Email                   15071 non-null  int64
 18  Uso_R_Sociales              15071 non-null  int64
 19  Especialistas_tic           15071 non-null  int64
 20  Personal_Conocimiento_tic   15071 non-null  int64
 21  Mujeres_Conocimiento_tic    15071 non-null  int64
 22  Hombres_Conocimiento_tic    15071 non-null  int64
 23  Intranet                    15071 non-null  int64
dtypes: float64(1), int64(20), object(3)
memory usage: 2.8+ MB
```

Figura 9. Información detallada de la data

```
1 sns.set(font_scale=0.6)
2 sns.countplot(x="Tamaño_Empresa", data=tesis_df)
```

```
<Axes: xlabel='Tamaño_Empresa', ylabel='count'>
```

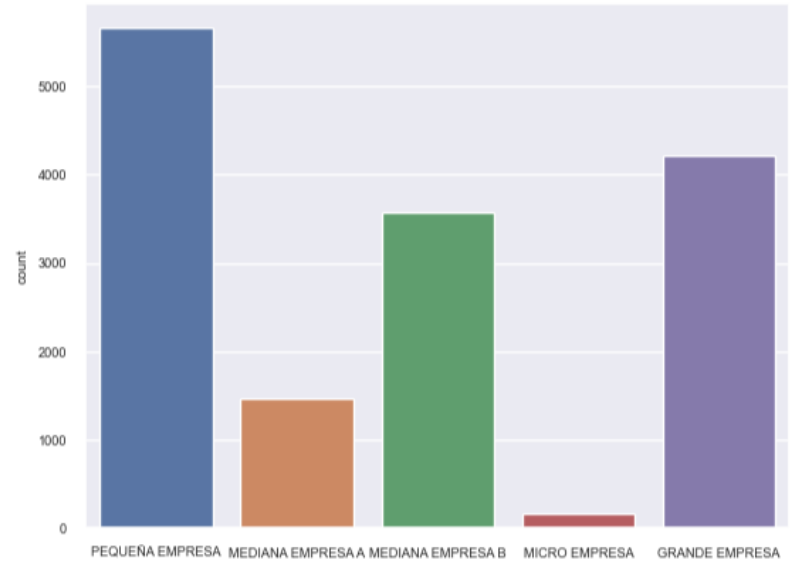


Figura 10. Variable tamaño de empresa

```
sns.countplot(x="Provincia", data=tesis_df)
plt.xticks(rotation=90)
plt.show()
```

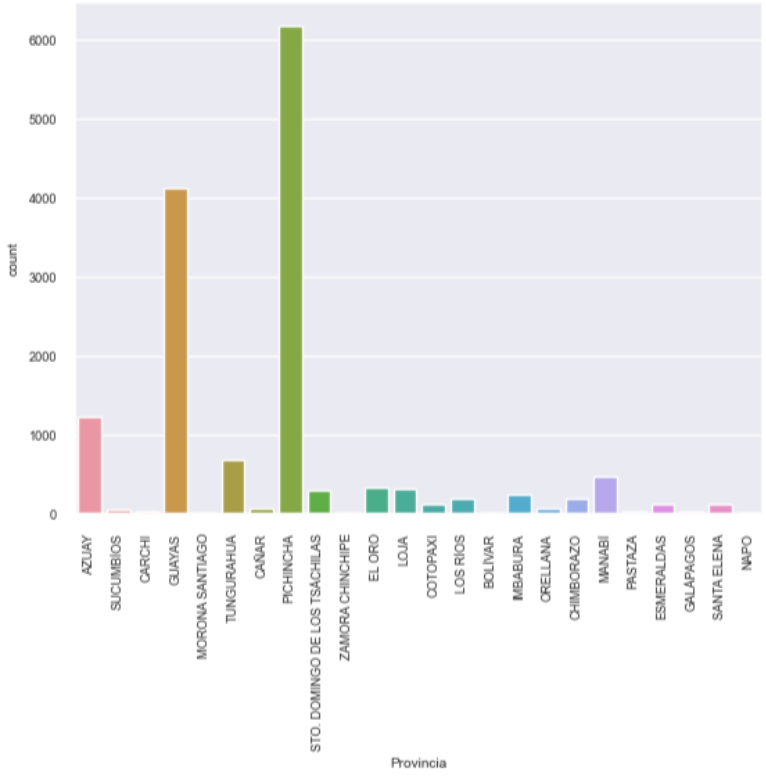


Figura 11. Variable provincias

```
# Grafico para verificacion de datos nulos
sns.heatmap(tesis_df.isna())
```

```
<Axes: >
```

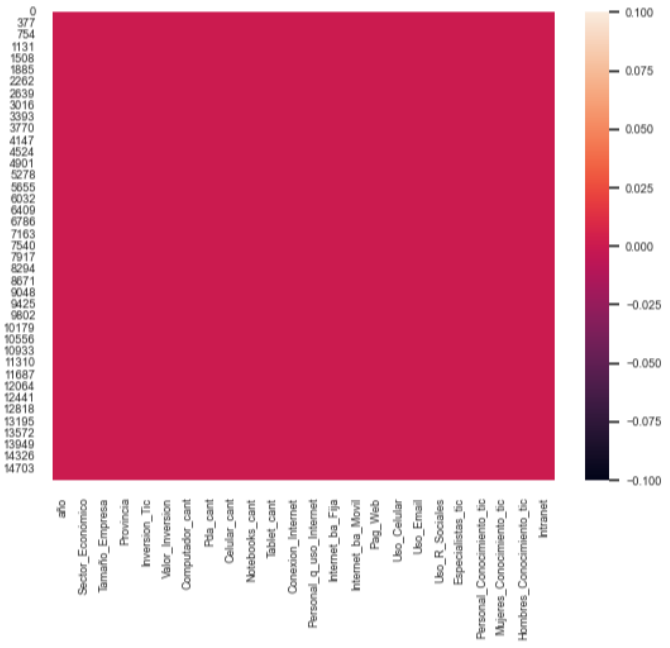


Figura 12. Verificación de datos nulos

Como última parte de esta fase de limpieza, después de limpiar, construir, integrar y formatear datos; guardamos la data resultante en formato Excel ya que nos servirá de base para la creación del dashboard en PowerBI; utilizamos el código: tesis_df.to_excel('data_BI.xlsx', index=False)

4.3. Fase 4

El modelado parte con la selección de la técnica que en este caso usamos algoritmos de predicción con entrenamiento supervisado para los modelos de Machine Learning, por otra parte, usamos la herramienta PowerBI para el análisis de data y para crear un dashboard de las etiquetas categóricas que no se pudieron modelar en ML.

En primera instancia declaramos las variables etiqueta y las variables características, para continuar con la división de la data y terminar con el entrenamiento del modelo.

Nuestro primer modelo es un algoritmo de regresión logística que tiene como objetivo predecir el resultado de la variable categórica 'Conexion_Internet', donde el resultado será si la empresa tiene o no conexión a internet.

Declaramos la etiquetas (Y), y las características (X)

```
X = tesis_df[['Computador_cant', 'Pda_cant',
      'Celular_cant', 'Notebooks_cant', 'Tablet_cant',
      'Pag_Web', 'Uso_Celular', 'Uso_Email', 'Uso_R_Sociales',
      'Especialistas_tic', 'Personal_Conocimiento_tic','Intranet']]
y = tesis_df['Conexion_Internet']
```

Continuamos con la división de datos para el entrenamiento del modelo, lo recomendable es una división de 70/30 u 80/20 para la aplicación usaremos 70/30 y tenemos el siguiente código: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state = 42); el entrenamiento del modelo lo conseguimos creando el modelo y entrenándolo de la siguiente manera:

lrm = LogisticRegression()

lrm.fit(X_train, y_train)

De esta manera cumplimos con la fase 4 y demos aplicar para para los siguientes modelos que son algoritmo bayesiano y regresión lineal, donde cambia la creación del modelo y el entrenamiento.

Regresión lineal, lrm = LinearRegression()

lrm.fit(X_train, y_train)

Bayesiano , gnb = GaussianNB()

gnb.fit(X_train, y_train)

4.4. Fase 5

Finalmente llegamos a la evaluación del modelo, gracias a las librerías de sklearn tenemos varias métricas que nos dan la posibilidad de medir los resultados del modelo como por ejemplo el coeficiente de determinación, acurrancy, matriz de confusión, reporte de clasificación entre otros, como podemos evidenciar a continuación en los gráficos de los resultados de nuestro modelo ejemplo Regresión Logística(Internet).

```
# para calcular el score de test
score = lrm.score(X_test, y_test)
print(score)
```

```
0.9854046881910659
```

Figura 13. Resultado score

En la figura anterior tenemos el resultado del 98.54% de valores clasificados correctamente y por otra parte en la siguiente figura tenemos el resultado de R2 que también es del 98.20% lo cual le convierte en un modelo confiable para ponerlo en producción.

```
# Resultado del entrenamiento
print("Coeficiente de determinacion R2", lrm.score(X_train,y_train))
```

```
Coeficiente de determinacion R2 0.9820836098208361
```

Figura 14. R2

La función de predicción nos ayuda a visualizar los resultado que arroja el modelo después de su entrenamiento.

```
# Realizar predicciones utilizando el set de prueba
y_pred = lrm.predict(X_test)
y_pred
```

```
array([1, 1, 1, ..., 1, 1, 1], dtype=int64)
```

Figura 15. Función predicción

Tenemos el reporte de clasificación donde podemos visualizar el porcentaje y medidas de precisión de nuestro modelo.

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.76	0.95	0.84	189
1	1.00	0.99	0.99	4333
accuracy			0.99	4522
macro avg	0.88	0.97	0.92	4522
weighted avg	0.99	0.99	0.99	4522

Figura 16. Reporte de clasificación

Dentro de la evaluación también es un indicador bastante confiable la matriz de confusión donde nos indica los positivos y negativos correctos como también falsos negativos y positivos.

```
confusion_matrix(y_test, y_pred)
```

```
array([[ 179,   10],
       [  56, 4277]], dtype=int64)
```

Figura 17. Matriz de confusión

Los resultados mostrados en la imagen nos dicen que la tasa de precisión es ((179+4277)/4522) corresponde al 98.54% y la tasa de error ((56+10)/4522) dando el 1.46%.

Una vez evaluado el modelo debemos guardarlo para posterior despliegue para guardar el modelo usamos la librería pickle.

5. VISUALIZACIÓN Y ANÁLISIS DE DATOS

La información que presentamos en este apartado nos ayuda a observar de mejor manera los resultados de los modelos de Machine Learning y Bigdata, este tipo de información está orientada a la compresión de los resultados y nos ayuda a tomar decisiones.

5.1. Visualización de datos

Después de realizar varias pruebas con diferentes tipos de algoritmos tanto de regresión lineal, logística, bayesianos, K-NN y SVM aplicando a diferentes etiquetas se obtiene resultado con % muy bajos e incluso negativos por tal razón se define trabajar con los modelos que mejor resultado nos dio como podemos observar en la siguiente ilustración.

Ister resultado de algoritmos

Modelo / Variable	R. Lineal		R. Logistica	R. Polynomial	Bayesiano	Arbol DD	k-NN	SVM
	score	R2						
V. Inversion	17.00%	51.24%		-5.33%		-1.20%		-0.01%
V. Inversion (Normalizado)	25.00%	51.00%				19.00%		
Cant. Computadores	38.00%	35.00%				-75.00%		4.00%
Personal conoce TIC	13.99%	37.07%				-30.00%	2.00%	
Inversion			67.00%		58.00%	-40.00%	4.00%	
Especialistas			75.00%		75.00%	-2.00%	19.00%	
Internet			98.54%		16.00%	39.00%	45.00%	-7.00%
Intranet			68.00%		60.00%	-44.00%	4.00%	
Pag. Web			65.87%		56.00%	-32.00%	12.00%	-11.00%

Figura 18. Resultados pruebas de algoritmos

En el despliegue de los modelos evaluamos varias opciones iniciando por la más sencilla que es poner en producción el modelo en un servidor en entorno local, para este fin utilizamos un editor de código que es Visual Studio Code y el lenguaje de programación Python, dentro del mismo importamos la librería streamlit que nos permite crear aplicaciones web interactivas.

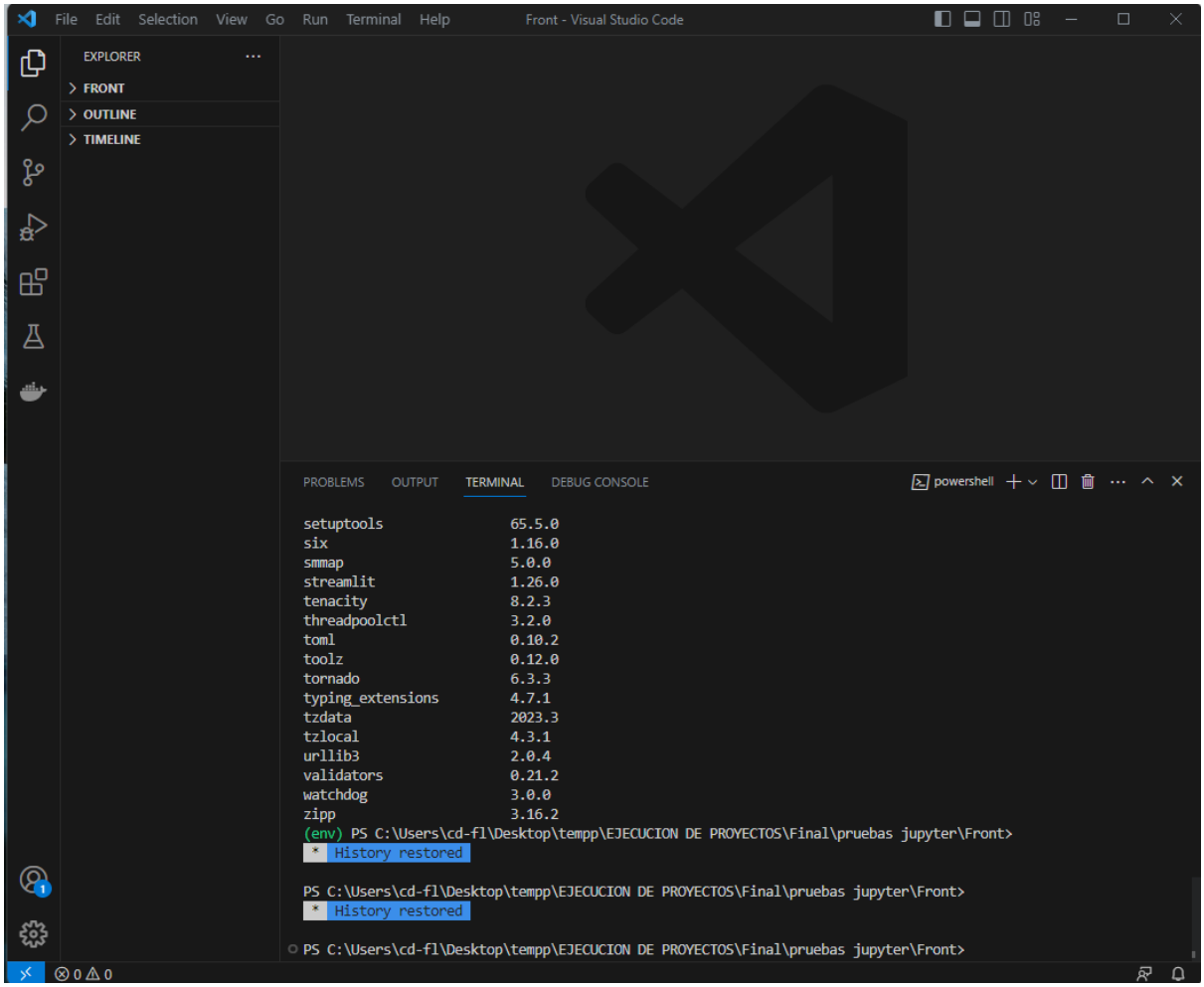


Figura 19. Entorno de Visual Studio Code

Con Python tenemos la facilidad de implementar servidores virtuales para realizar las pruebas necesarias de nuestros modelos antes de ponerlos en producción, la manera de crear y usar estos entornos es con los siguientes pasos:

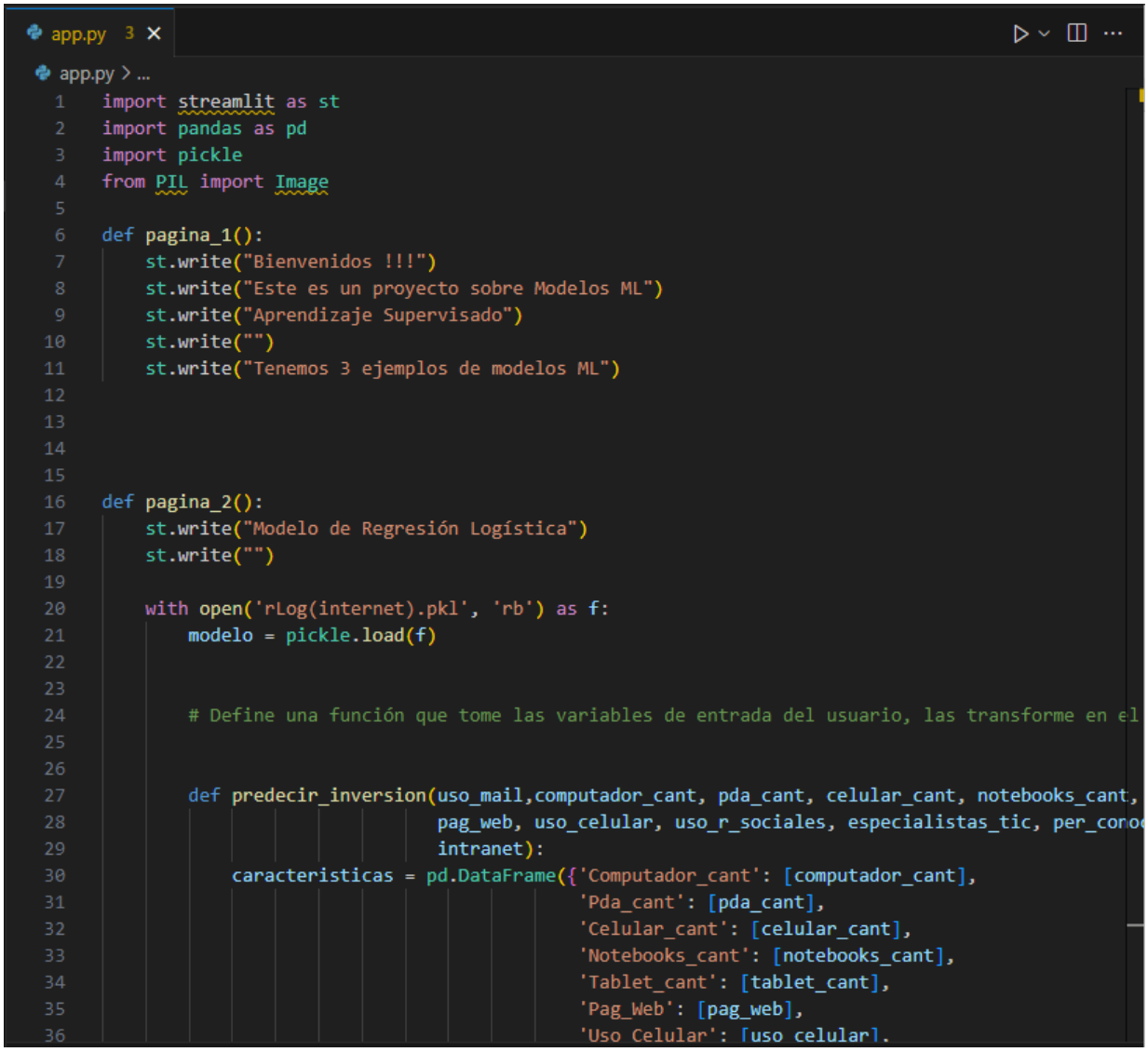
Abrir cmd: dirigirnos al directorio raíz de nuestra app

Crear entorno virtual: python -m venv env

Ingresamos al entorno: .\env\Scripts\activate

Instalamos streamlit: pip install streamlit

Después de eso paso nos dirigimos a Visual Studio Code, para desarrollar nuestra interfaz que consumirá nuestro modelo .pkl.



```python
import streamlit as st
import pandas as pd
import pickle
from PIL import Image

def pagina_1():
    st.write("Bienvenidos !!!")
    st.write("Este es un proyecto sobre Modelos ML")
    st.write("Aprendizaje Supervisado")
    st.write("")
    st.write("Tenemos 3 ejemplos de modelos ML")




def pagina_2():
    st.write("Modelo de Regresión Logística")
    st.write("")

    with open('rLog(internet).pkl', 'rb') as f:
        modelo = pickle.load(f)


        # Define una función que tome las variables de entrada del usuario, las transforme en el


        def predecir_inversion(uso_mail,computador_cant, pda_cant, celular_cant, notebooks_cant,
                               pag_web, uso_celular, uso_r_sociales, especialistas_tic, per_coroc
                               intranet):
            caracteristicas = pd.DataFrame({'Computador_cant': [computador_cant],
                                            'Pda_cant': [pda_cant],
                                            'Celular_cant': [celular_cant],
                                            'Notebooks_cant': [notebooks_cant],
                                            'Tablet_cant': [tablet_cant],
                                            'Pag_Web': [pag_web],
                                            'Uso Celular': [uso celular],
```

Figura 20. Desarrollo del app

Una vez desarrollada app ponemos a prueba y desplegamos el modelo en el entorno local con los siguientes comandos

.\env\Scripts\activate

streamlit run app.py

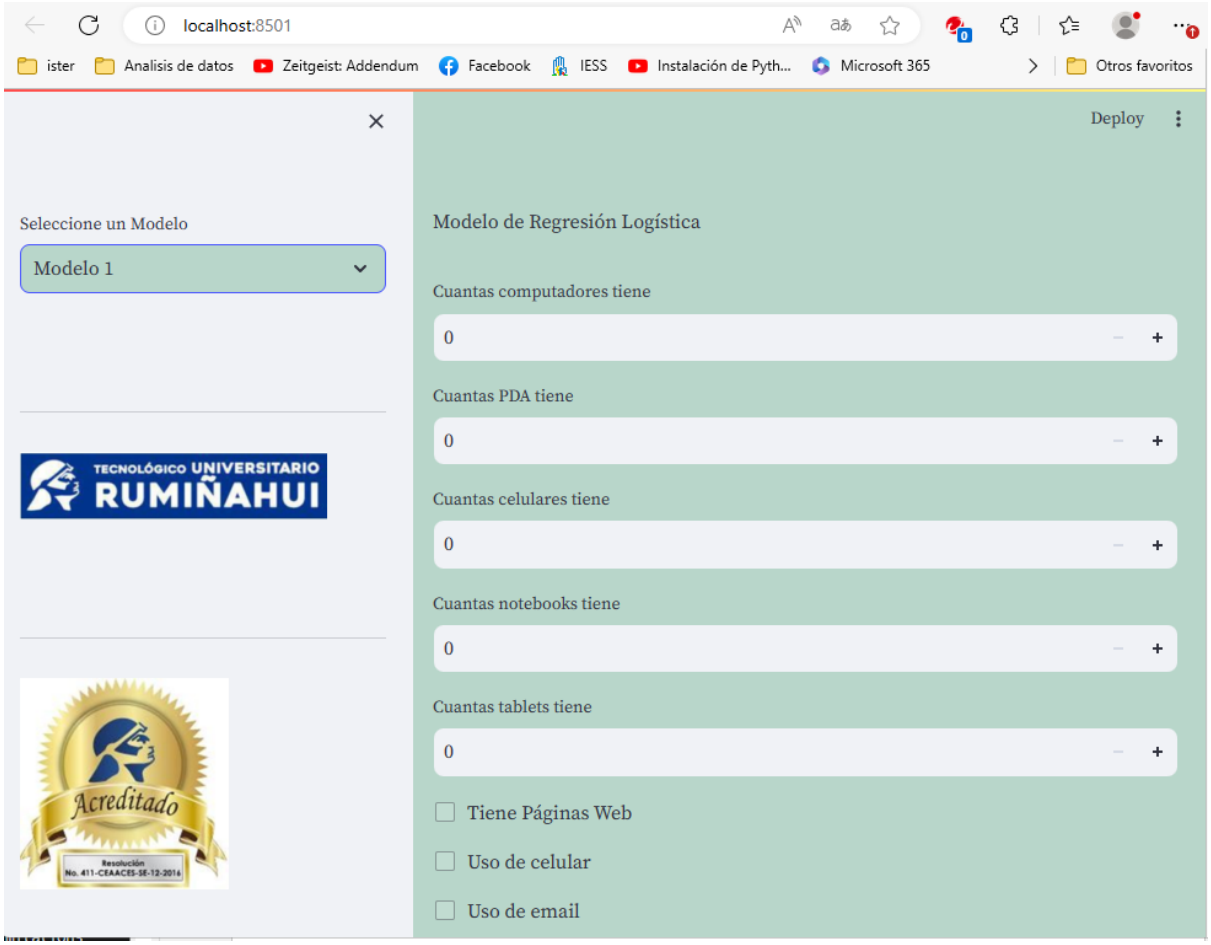


Figura 21. Despliegue entorno local

Como podemos mirar está corriendo el localhost:8501 aquí ya podemos ingresar los datos para calcular el resultado del modelo y poder evaluar; una vez realizado las pruebas se toma la decisión de subir la app a la nube para facilidad del usuario, se realiza investiga las opciones de subir a la nube de GitHub, a un VPS de Hostinger y a la nube de Google Cloud siendo este ultimo la mejor opción ya que es más fácil y hay manera de realizar de forma gratuita. Por otro lado, GitHub no permite correr aplicaciones web de python y Hostinger también se considera una muy buena opción, pero es de pago.

En Google Cloud, necesitamos iniciar sesión con una cuenta de Gmail o a su vez crear una cuenta nueva, creamos un nuevo proyecto y a la par instalamos el Cloud SDK en nuestro computador para enlazar nuestra cuenta y poder subir nuestra app después de este realizar varios pasos para completar este proceso, nos genera un url para poder acceder a la app y

utilizar los modelos desplegados que en este caso subimos 3 modelos de Machine Learn, nuestro link es app · Streamlit (despliegue-398320.uw.r.appspot.com).

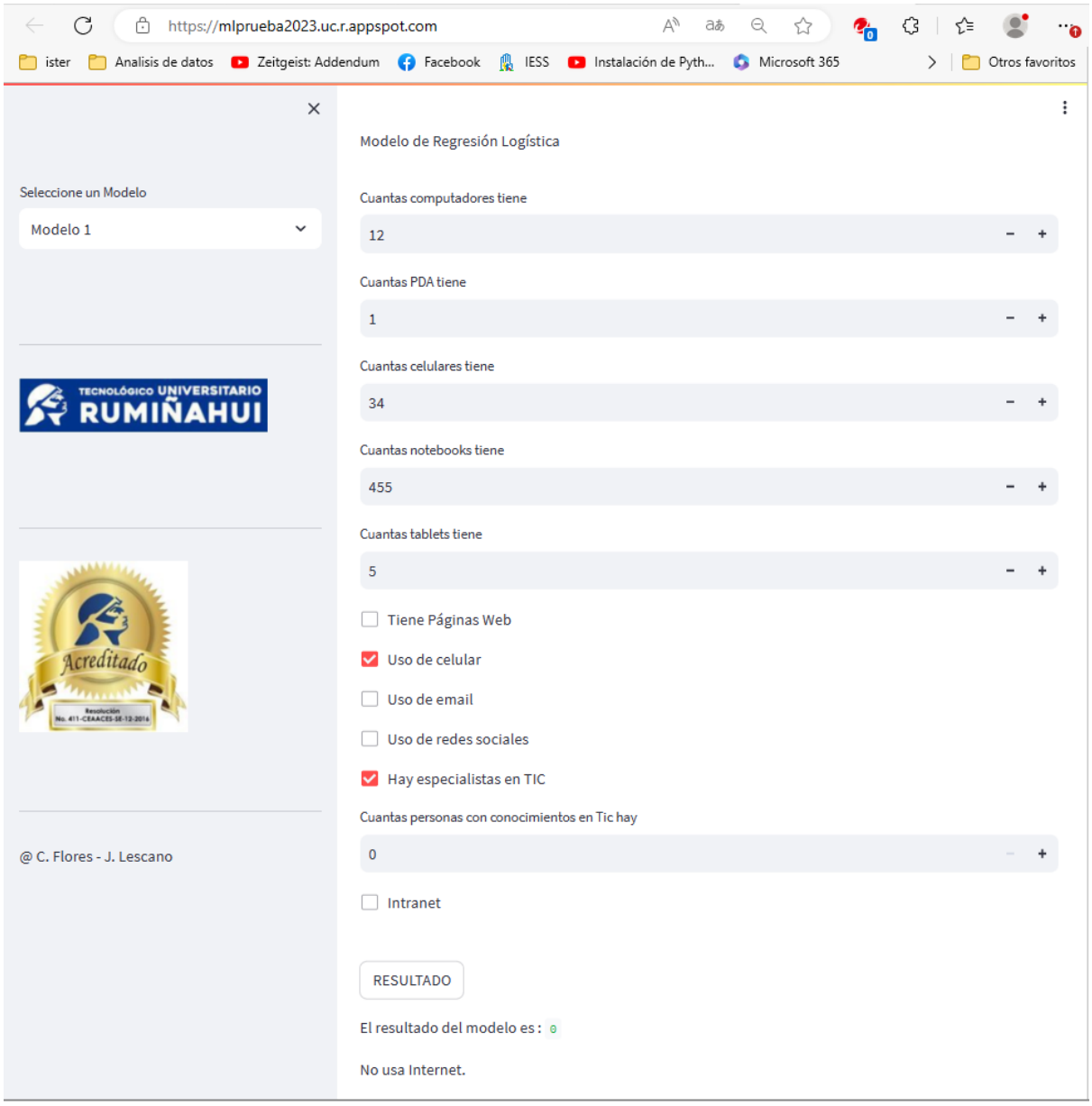


Figura 22. Modelo regresión logística

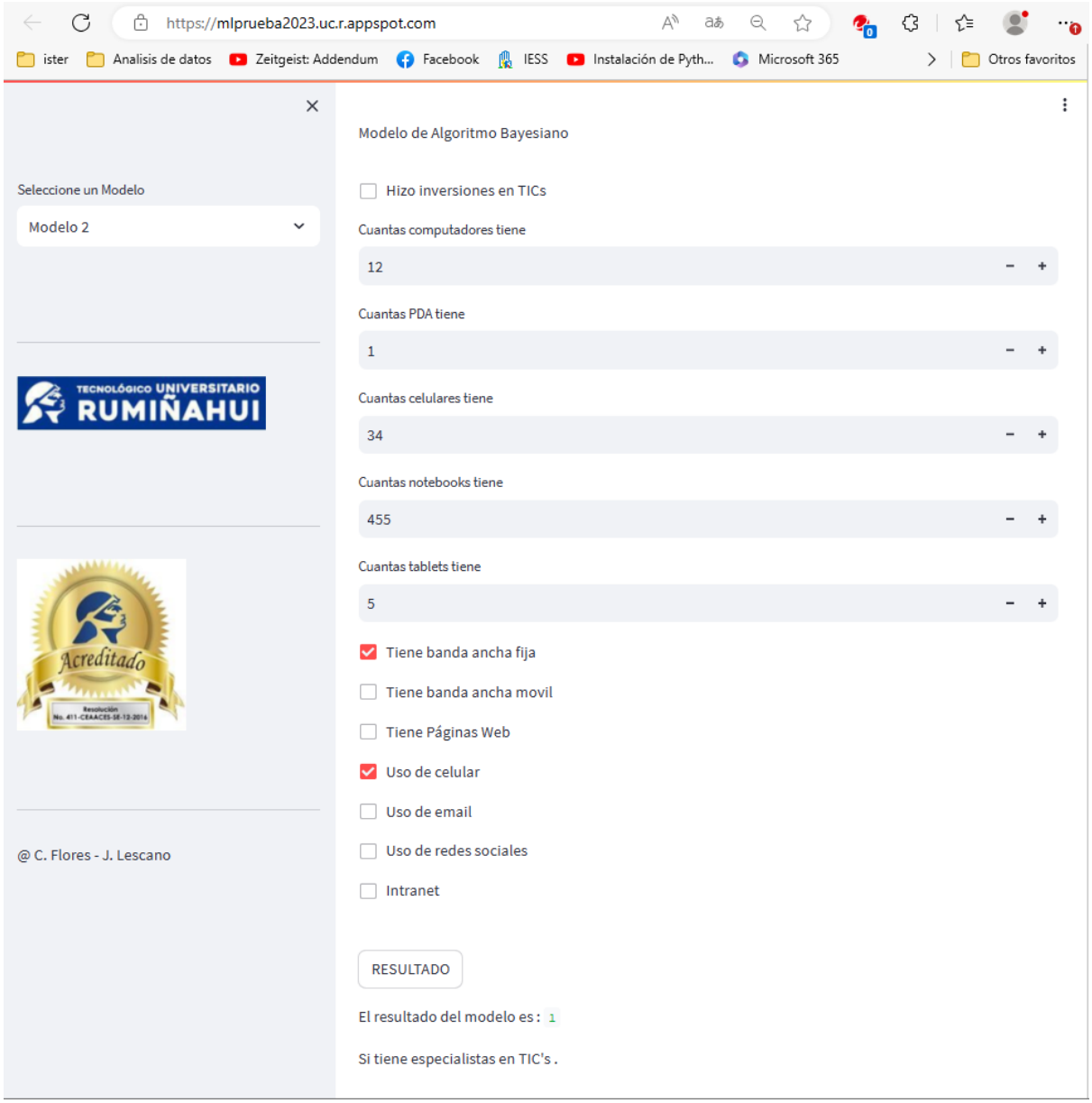


Figura 23. Algoritmo bayesiano

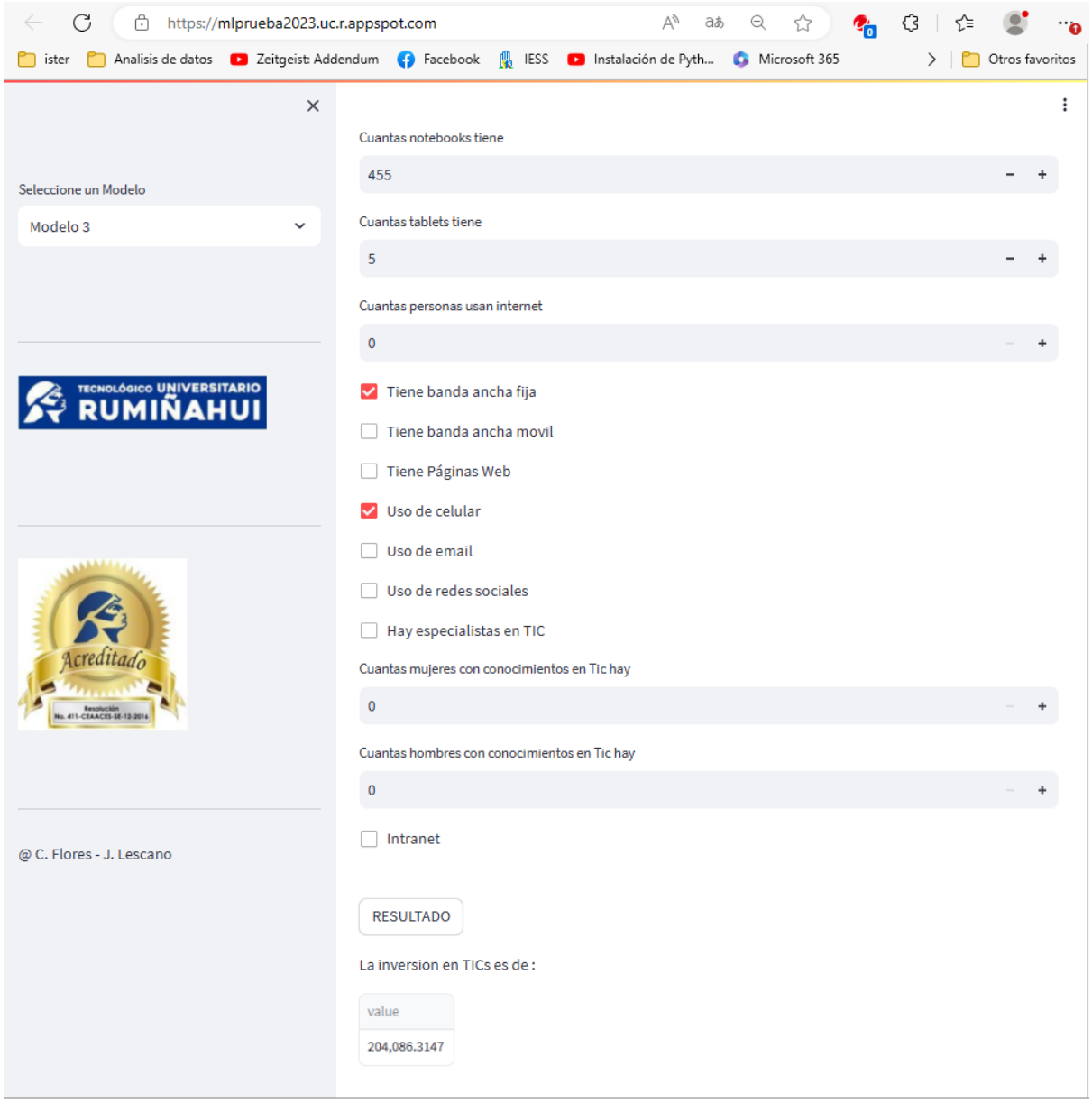


Figura 24. Modelo regresión lineal

Las etiquetas como año, provincia, sector económico, tamaño de empresa se utilizaron para hacer análisis con PowerBI, mediante la creación de dashboard para la presentación de la información de manera gráfica y dinámica, estos informes nos ayudan a comprender el estado actual del negocio por medio de gráficos de barras, de pasteles e incluso mapas.

Tenemos un análisis global que presenta información de la inversiones que realizaron las empresa por provincias y segmentado por años desde el 2012 al 2015, cada ventana tiene

hipervínculos que nos lleva a otros informes, en el mapa se evidencia que las empresas que hicieron mayor inversión en tics pertenecen a las provincias de Pichincha y Guayas.

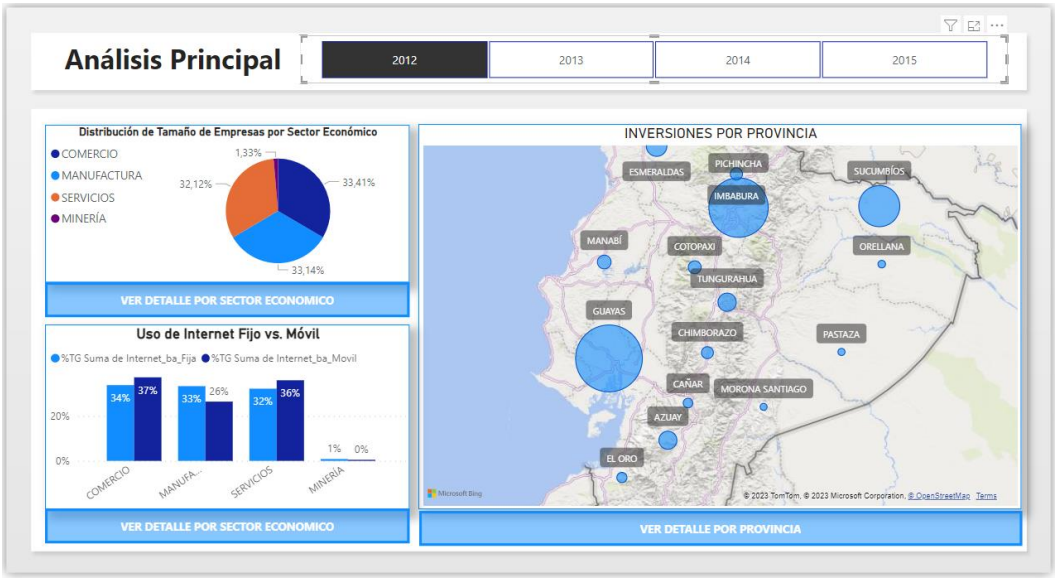


Figura 25. Analisis global

En el análisis de inversión por provincia ratificamos que las empresa que más invierten en tics son de Pichincha y Guayas con el 89.31% con referencia la inversión total de los 4 años que corresponde a $685,772.599 millones de dólares de un total de $767,912.523 millones de dólares y las empresa de la provincia de Bolívar las que menos invierten tics con apenas $26.092 dólares

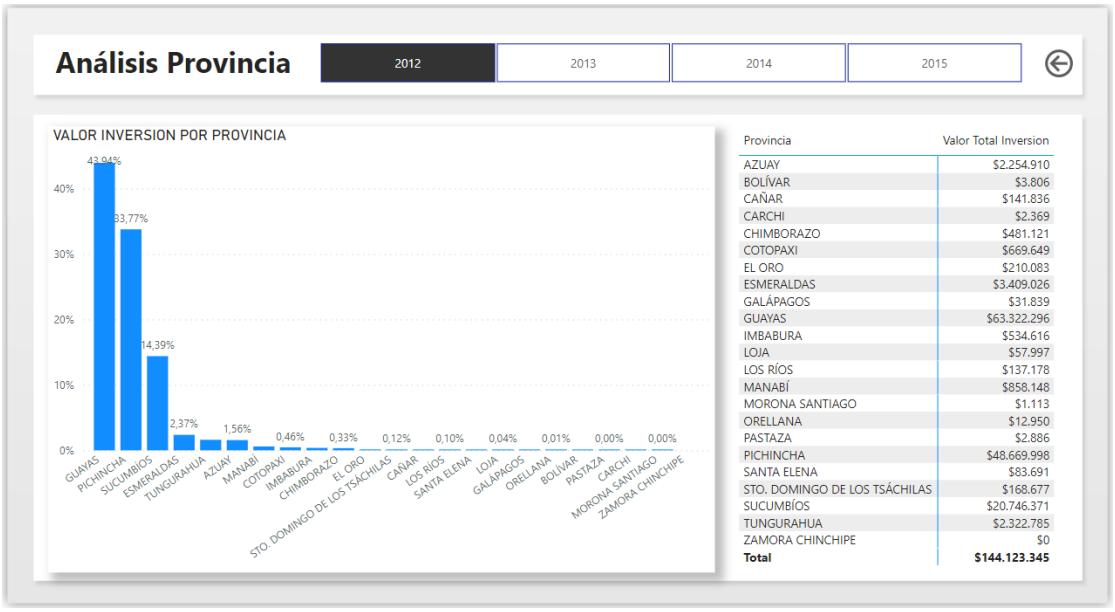


Figura 26. Inversiones por provincias

Teniendo la variable hombres y mujeres que tienen conocimiento en TIC'S se hizo un análisis para determinar los porcentajes que y la distribución por el tipo de sector económico y ubicación geográfica de las empresa.

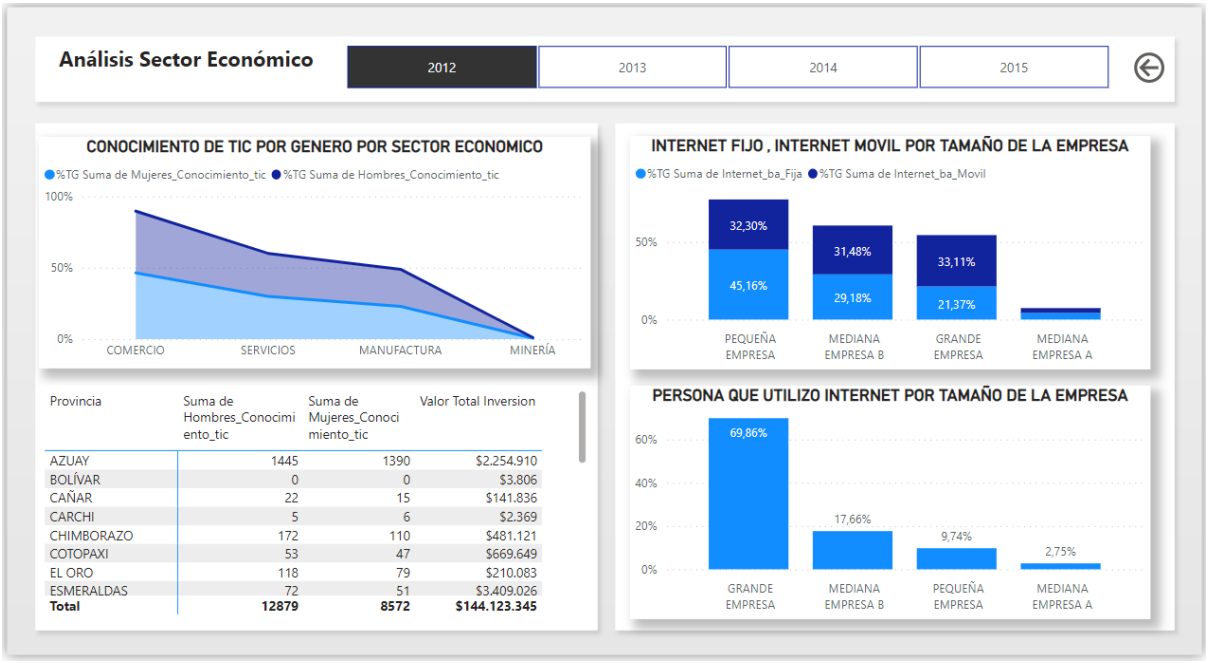


Figura 27. Análisis de personal por sector económico empresarial

Otra variable interesante es el uso de internet para lo cual usamos las variables que definen el tipo de internet que usaron las empresa por su tamaño haciendo una relación para

saber la cantidad de empresas que usan internet de banda ancha fija o móvil, como también tenemos la distribución por provincias.

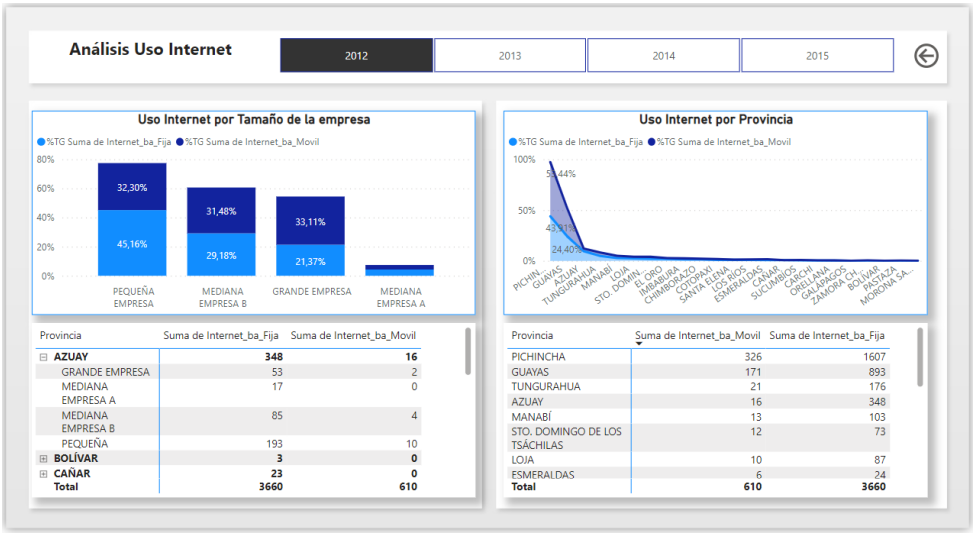


Figura 28. Tipo de internet por tamaño de empresa

5.2. Discusión y análisis de resultados

El análisis de datos realizado en este proyecto presenta varios resultados que muchas pruebas que se hicieron en cada etapa de desarrollo, pruebas datas, de modelos, de reportes, de despliegue, de servidores; que en el proceso de implementación se fueron descartando y puliendo.

Iniciamos con la elección de un entorno de desarrollo y un lenguaje de programación, al tener una gran cantidad de opciones, investigamos y decidimos usar la herramienta Jupyter Notebook con lengua de programación Python por ser una herramienta potente y un lenguaje de programación de uso general que cuentan con una extensa lista de librerías que nos facilita el trabajo a la hora de desarrollar nuestro proyecto, debemos resaltar que cuenta con el respaldo de una gran comunidad para desarrollo y soporte de igual forma son compatibles con la gran mayoría de sistemas operativos.

La búsqueda de la data correcta fue una de las etapas más complejas del desarrollo del proyecto ya que muchas son incompletas o tiene información redundante o de poca valides, después de elegir la data la parte de comprender la data y comprender el negocio requeriré de

un arduo trabajo, ya que existe una gran cantidad de información; la limpieza y depuración de la data es una fase indispensable para dar continuidad al desarrollo.

En la aplicación de modelos se requiere realizar muchas pruebas ya que al existir un extenso número de algoritmos debemos encontrar el que más se apegue al cumplimiento de nuestros objetivos, cuando ya tenemos definido los modelos viene la etapa de despliegue que también requiere de investigación para elegir uno acorde a nuestras expectativas.

La aplicación de Bigdata a nuestras variables categóricas que no se pudo aplicar ningún modelo de aprendizaje supervisado, fue una etapa complementaria ya que si por una parte se busca predecir o clasificar los datos por medio de modelos que nos muestran las posibilidades futuras con Bigdata podemos conocer el estado actual del negocio nos brinda información que nos ayuda a entender de mejor manera al negocio.

Los resultados obtenido del desarrollo de este proyecto en el caso de la aplicación los modelos nos dieron a entender que depende mucho del tipo de modelo que se elija y de los datos que se definan como etiquetas y características de igual forma el planteamiento de los objetivos no obliga a buscar los mejores resultados de los modelos como podemos evidencia en la **Figura *18*** que muestra 34 respuestas de las pruebas realizadas.

Resaltando que el modelo que más se apegó al objetivo propuesto es el algoritmo de regresión logística con la etiqueta Uso de Internet, el 98.54% de precisión del modelo nos indica que cuando despleguemos el modelo y el usuario ingrese los datos el resultado será confiable, el segundo modelo es con un algoritmo bayesiano con un 75% de precisión nos indica que tenemos un buen porcentaje aplicado al resultado donde el modelo nos indica si la empresa tiene o no un especialista en TIC’S por último el algoritmo de regresión logística intenta predecir el valor que una empresa invierte en materia de TIC’S, el % influye ya que existe una gran línea de valores entre el más alto y el más bajo.

Con la herramienta PowerBI hicimos un análisis de los datos de las variables inversión, provincia, año, sector económico, tamaño de la empresa, tipo de internet que usan las empresas así como el personal que tiene conocimiento en TIC'S segregado por género, como también la distribución por ubicación geográfica; creando varios dashboard interactivos que nos muestra la información numérica, porcentajes y segmentado por periodos de los años 2012 al 2015, obteniendo resultados interesantes como las empresas que mayor inversión hacen están en las provincias de Pichincha y Guayas, que en las personas que tienen conocimiento en Tics predomina el género masculino con 60% versus el 40%, el tipo de internet que más usan las empresas es el internet de banda ancha fija.

6. CONCLUSIONES

Las encuestas que realiza el INEC contribuyen con información muy relevante que nos ayuda a conocer el estado e importancia que dan las empresa al tema de las Tics, aun cuando es de conocimiento general que no todas las empresas son encuestadas, existe una gran cantidad de registros en la base de datos.

Los dashboard que se crean con PowerBI son muy eficientes y nos permiten conocer el estado actual del negocio, se puede realizar varias relaciones entre las variables y mostrar información muy atractiva y segmentada, este tipo de información se considera para la toma de decisiones.

Con la implementación de modelos de aprendizaje supervisado se puede predecir o clasificar la información, es muy importante analizar las variables que se tienen en la data para de esta forma obtener buenos resultados en la precisión del modelo.

Muchas empresa muestran un porcentaje muy pequeño en inversiones en Tics esto puede ser por un factor de desconocimiento o porque sus procesos o giro de negocio aún son artesanales y no tecnifican sus instalaciones.

7. RECOMENDACIONES

La principal recomendación es elegir muy bien la data para la implementación de modelos de aprendizaje supervisado ya que de esto depende si el resultado es confiable o no, después de la elección de la data tomarse el tiempo suficiente para la comprensión de los datos y del negocio.

La elección de las herramientas y lenguaje de programación es muy importante ya que existen varias opciones en nuestro caso elegimos PowerBI, Jupyter y Python estos últimos por poseer gran cantidad de librerías y están respaldados por una extensa comunidad que brindan soporte, además se debe destacar que es un lenguaje universal y multiplataforma.

El despliegue de modelos es una parte bastante laboriosa ya que hay varias opciones y hay que investigar el alcance de cada una por ejemplo GitHub requiere conocimientos medios del manejo de la plataforma para poder subir aplicaciones, lamentablemente no es compatible con aplicaciones .py, razón por la cual optamos por subir nuestra aplicación al servidor de Google Cloud que es de fácil acceso se puede configurar para obtener un periodo sin.

Con la herramienta PowerBI debemos tener en cuenta que, al ser una aplicación de pago destinada principalmente para organizaciones, podemos tener limitantes dependiendo del tipo de licencias y configuraciones de la organización, por ejemplo, limitaciones para publicar los dashboard creados.

8. BIBLIOGRAFÍA

Aprendeconalf. (2023). Visualización de datos. *Aprendeconalf*. https://aprendeconalf.es/docencia/python/manual/pandas/#:~:text=Pandas es una librería de,NumPy pero con nuevas funcionalidades.

Arteaga, P., Jiménez, M., & Batanero, C. (2021). Variables que caracterizan los gráficos estadísticos y las tareas relacionadas con ellos en los libros de texto de educación secundaria en Costa Rica. *Avances de Investigación En Educación Matemática*, *20*, 125–140. https://doi.org/10.35763/aiem20.4001

Asamblea Nacional. (2010). *Ley Del Sistema Nacional De Registro De Datos Públicos*.

Caceres, D. (2023). Datasets: Qué son y cómo acceder a ellos. *Openwebinars*, 3.

Cali, C. (2022). *Modelo predictivos de las ventas de productos de primera necesidad en el sector comercial*.

Carrillo, V. (2020). Impacto de las Tecnologías de Información en las pequeñas y medianas empresas del sector servicios en el Distrito Metropolitano de Quito. *UIDE*, *21*(1), 1–9. https://repositorio.uide.edu.ec/bitstream/37000/4368/1/T-UIDE-1368.pdf

Constitución Política de la República Del Ecuador, A. N. C. (2008). *La constitución*. 54. http://pdba.georgetown.edu/Parties/Ecuador/Leyes/constitucion.pdf

Costa, P., Armijos, V., Loaiza, F., & Aguirre, G. (2019). Inversión en TICS en las empresas del Ecuador para el fortalecimiento de la gestión empresarial Periodo de análisis 2012-2015. *Revista Espacios*, *39*(47), 5–11. http://www.revistaespacios.com/a18v39n47/a18v39n47p05.pdf

Datascientest. (2023). Limpieza de datos. *Datascientest*. https://datascientest.com/es/datacleaning-limpieza-de-datos-definicion-tecnicas-importancia-en-data-science

Deloitte. (2021). *Calidad de Datos en la era del Big Data Automatización y priorización como*

habilitadores clave.

Etecé, E. (2022). Datos. *Editorial Etecé*.

Galán, V. (2019). Crisp-Dm a Un Proyecto De Minería. *BIBLIOTECA de La Universidad Carlos III de Madrid*, 120. https://e-archivo.uc3m.es/handle/10016/22198

Hotz, N. (2023). What is CRISP DM. *Datascience*.

IBM. (2022). Regresión lineal. *Ibm*.

Mancilla, E. (2022). Algoritmo de machine Learning. *Invgate*. https://blog.invgate.com/es/machine-learning

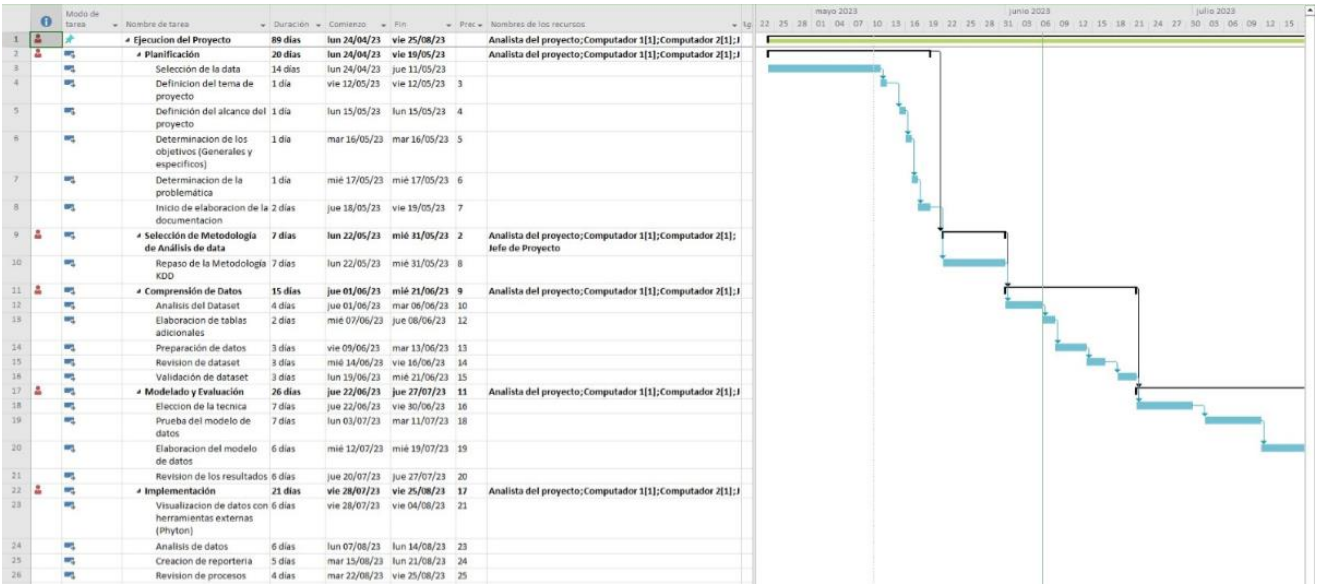
Molina Salazar, R., & López Morales, A. (2021). El uso de las TICs como medio de reactivación económica en las MiPymes de alojamiento del cantón Guayaquil para tiempos post-COVID. *Universidad Católica de Santiago de Guayaquil*, 135. http://201.159.223.180/bitstream/3317/17354/1/T-UCSG-PRE-ECO-ADM-594.pdf

Nacional, C., Electronico, L. E. Y. D. E. C., Datos, E. Y. M. D. E., Preliminar, T., & Generales, P. (2002). *Ley de comercio electronico, firmas y mensajes de datos*. 1–17. https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2012/11/Ley-de-Comercio-Electronico-Firmas-y-Mensajes-de-Datos.pdf?fbclid=IwAR2PhfFJMvEU4S0R_nYNE2--YV9mjaGvZ-eTb0efkBpKn5QEgmnrIwJeGMA

Perez, A. (2019). Capacidades analíticas. *Business School*. https://www.obsbusiness.school/blog/capacidad-analitica-sus-ventajas-en-el-desarrollo-de-proyectos

9. ANEXOS. -

A continuación, se presenta la imagen de la administración del tiempo del Proyecto, este cronograma de actividades está realizado en la herramienta Project de Microsoft



	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin	Pred	Nombres de los recursos
1		**Ejecucion del Proyecto**	**89 días**	**lun 24/04/23**	**vie 25/08/23**		**Analista del proyecto;Computador 1[1];Computador 2[1];J**
2		**Planificación**	**20 días**	**lun 24/04/23**	**vie 19/05/23**		**Analista del proyecto;Computador 1[1];Computador 2[1];J**
3		Selección de la data	14 días	lun 24/04/23	jue 11/05/23		
4		Definicion del tema de proyecto	1 día	vie 12/05/23	vie 12/05/23	3	
5		Definición del alcance del proyecto	1 día	lun 15/05/23	lun 15/05/23	4	
6		Determinacion de los objetivos (Generales y especificos)	1 día	mar 16/05/23	mar 16/05/23	5	
7		Determinacion de la problemática	1 día	mié 17/05/23	mié 17/05/23	6	
8		Inicio de elaboracion de la documentacion	2 días	jue 18/05/23	vie 19/05/23	7	
9		**Selección de Metodología de Análisis de data**	**7 días**	**lun 22/05/23**	**mié 31/05/23**	2	**Analista del proyecto;Computador 1[1];Computador 2[1];Jefe de Proyecto**
10		Repaso de la Metodología KDD	7 días	lun 22/05/23	mié 31/05/23	8	
11		**Comprensión de Datos**	**15 días**	**jue 01/06/23**	**mié 21/06/23**	9	**Analista del proyecto;Computador 1[1];Computador 2[1];J**
12		Analisis del Dataset	4 días	jue 01/06/23	mar 06/06/23	10	
13		Elaboracion de tablas adicionales	2 días	mié 07/06/23	jue 08/06/23	12	
14		Preparación de datos	3 días	vie 09/06/23	mar 13/06/23	13	
15		Revision de dataset	3 días	mié 14/06/23	vie 16/06/23	14	
16		Validación de dataset	3 días	lun 19/06/23	mié 21/06/23	15	
17		**Modelado y Evaluación**	**26 días**	**jue 22/06/23**	**jue 27/07/23**	11	**Analista del proyecto;Computador 1[1];Computador 2[1];J**
18		Eleccion de la tecnica	7 días	jue 22/06/23	vie 30/06/23	16	
19		Prueba del modelo de datos	7 días	lun 03/07/23	mar 11/07/23	18	
20		Elaboracion del modelo de datos	6 días	mié 12/07/23	mié 19/07/23	19	
21		Revision de los resultados	6 días	jue 20/07/23	jue 27/07/23	20	
22		**Implementación**	**21 días**	**vie 28/07/23**	**vie 25/08/23**	17	**Analista del proyecto;Computador 1[1];Computador 2[1];J**
23		Visualizacion de datos con herramientas externas (Phyton)	6 días	vie 28/07/23	vie 04/08/23	21	
24		Analisis de datos	6 días	lun 07/08/23	lun 14/08/23	23	
25		Creacion de reporteria	5 días	mar 15/08/23	lun 21/08/23	24	
26		Revision de procesos	4 días	mar 22/08/23	vie 25/08/23	25	

Figura 29. Cronograma de actividades

Link

Repositorio GitHub

Este es un enlace de GitHub con los archivos utilizados en el proyecto.

Dann041428/ml2023 (github.com)

Enlace Dashboard

Este enlace contiene la publicación del dashboard creado en PowerBI

Analisis - Power BI

Enlace Despliegue

Este enlace contiene el despliegue de los modelos en Google Cloud

app · Streamlit (despliegue-398320.uw.r.appspot.com)