

**ASIGNATURA**: Extracción de Conocimiento en Base de Datos

**ALUMNO**: Alexis Arath Alatorre Delgadillo

**PROFESOR**: María Eugenia Guerrero Chan

PORTAFOLIO

**Nayarit, Santiago, agosto de 2024**

INDICE

[**UNIDAD 1** 2](#_Toc169002896)

[**Resumen\_U1\_ECBD** 2](#_Toc169002897)

[**Ejercicios frecuencias** 5](#_Toc169002898)

[**Practica ejercicios\_U1\_ECBD** 7](#_Toc169002899)

[**EXAMEN** 12](#_Toc169002900)

[**UNIDAD 2** 13](#_Toc169002901)

[**Instalación del lenguaje R** 13](#_Toc169002902)

[**Instalación de Rstudio** 19](#_Toc169002903)

[**Mapa conceptual** 23](#_Toc169002904)

[**EXAMEN** 25](#_Toc169002905)

# **UNIDAD 1**

## **Resumen\_U1\_ECBD**

Instrucciones:

En el mundo actual, impulsado por la revolución digital y la creciente importancia de los datos, el análisis de datos se ha convertido en una herramienta esencial para las organizaciones de todos los sectores. Los datos son considerados como el nuevo combustible que impulsa la innovación, la toma de decisiones informadas y la ventaja competitiva. Sin embargo, la mera recopilación de datos no es suficiente; es necesario tener la capacidad de analizarlos y extraer información valiosa de ellos. Es aquí donde entran en juego las metodologías para el análisis de datos. Estas metodologías son enfoques estructurados y sistemáticos que proporcionan un marco de trabajo para guiar a los analistas de datos a través de las diferentes etapas del proceso de análisis, desde la recopilación y preparación de los datos hasta la comunicación de los resultados y hallazgos. Estas metodologías ayudan a garantizar la calidad, consistencia y eficiencia del análisis, permitiendo a las organizaciones aprovechar al máximo el valor de sus datos y obtener información valiosa que pueda ser utilizada para respaldar la toma de decisiones, resolver problemas, identificar oportunidades y generar nuevos conocimientos. Las metodologías para el análisis de datos son fundamentales en un mundo cada vez más orientado a los datos, donde la capacidad de extraer información significativa de conjuntos de datos masivos y complejos se ha convertido en una habilidad crítica. Estas metodologías no solo proporcionan un enfoque estructurado, sino que también fomentan la colaboración entre los diferentes actores involucrados en el proceso de análisis de datos, asegurando que los esfuerzos estén alineados con los objetivos estratégicos de la organización.

1. Título: Metodologías para el análisis de datos
2. Contenido (Introducción y Desarrollo):
3. **¿Qué es un Proyecto de Análisis de Datos?** Un proyecto de análisis de datos es un esfuerzo organizado y estructurado que implica la recopilación, procesamiento, análisis e interpretación de datos con el fin de obtener información valiosa y conocimientos que puedan ser utilizados para respaldar la toma de decisiones, resolver problemas o generar nuevos conocimientos.
4. **Características de un Proyecto de Análisis de Datos:**

* Tiene un objetivo claro y definido.
* Involucra la recopilación, procesamiento y análisis de datos.
* Utiliza técnicas y herramientas de análisis de datos.
* Requiere habilidades y conocimientos en análisis de datos, estadística y gestión de proyectos.
* Produce resultados e información valiosa para la toma de decisiones.

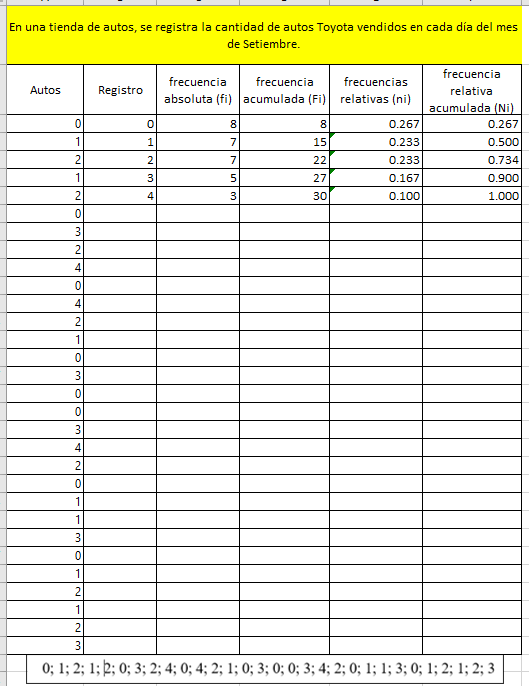
1. **¿Qué es una Metodología para el Análisis de Datos?** Una metodología para el análisis de datos es un enfoque sistemático y estructurado que proporciona un conjunto de principios, prácticas y técnicas para guiar el proceso de análisis de datos desde el inicio hasta el final. Estas metodologías establecen una serie de pasos o fases que deben seguirse para garantizar un análisis de datos eficiente, consistente y de alta calidad.
2. **Principales Metodologías o Métodos de Análisis de Datos:**

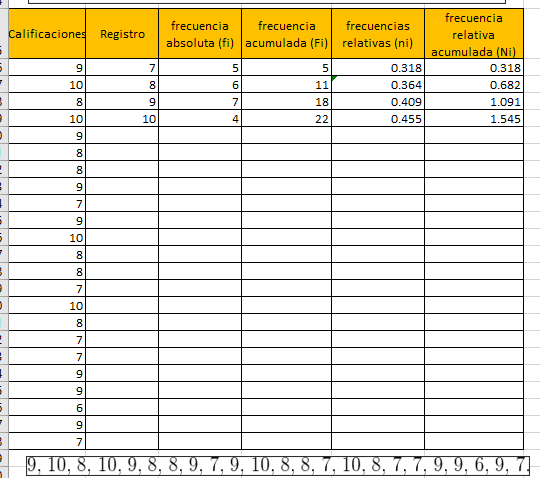
* **CRISP-DM (Cross-Industry Standard Process for Data Mining):** Es una metodología ampliamente utilizada que consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Se aplica en proyectos de minería de datos y análisis predictivo.
* **SEMMA (Sample, Explore, Modify, Model, Assess):** Desarrollada por SAS Institute, esta metodología se centra en el ciclo de vida del modelado de datos. Se utiliza comúnmente en proyectos de minería de datos y análisis predictivo.
* **KDD (Knowledge Discovery in Databases):** Es un proceso iterativo que consiste en seleccionar, preprocesar, transformar y minar los datos para descubrir patrones y conocimientos útiles. Se aplica en proyectos de minería de datos y descubrimiento de conocimiento.
* **Metodología Ágil para Análisis de Datos:** Inspirada en los principios de desarrollo ágil de software, esta metodología promueve la colaboración estrecha entre analistas y partes interesadas, entregando valor de manera iterativa y adaptándose a los cambios. Se utiliza en proyectos de análisis de datos donde los requisitos son dinámicos.
* **Metodología Lean para Análisis de Datos:** Basada en los principios Lean, esta metodología se centra en eliminar el desperdicio y maximizar el valor entregado al cliente. Se aplica en proyectos de análisis de datos donde la eficiencia y la reducción de costos son críticas.

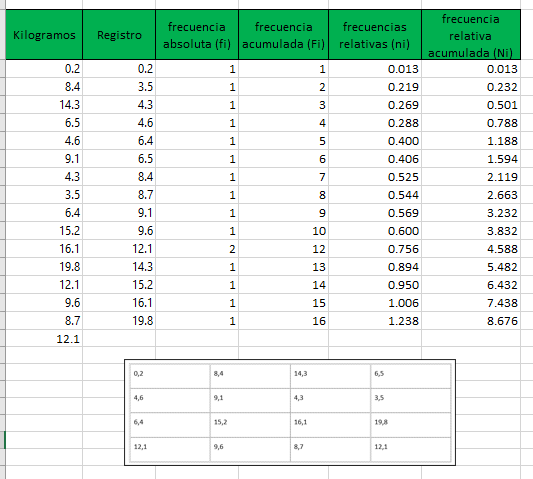
1. Bibliografía:

* Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.
* Rollins, J. (2015). Data Mining: A Practical Guide for Business and Industry. Apress.
* Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2017). Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. Wiley.
* Wills, G. (2014). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

## **Ejercicios frecuencias**







## **Practica ejercicios\_U1\_ECBD**

|  |  |
| --- | --- |
| **Instrumento** | *Práctica de ejercicios* |

|  |  |  |
| --- | --- | --- |
| **Alumno**: Alexis Arath Alatorre Delgadillo | | **Fecha: 21/05/2024** |
| **Carrera:** Ingeniería en Desarrollo y Gestión de Software | | **Grupo: IDGS91** |
| **Asignatura:** Extracción de Conocimiento en Bases de Datos. | **Unidad temática:** I. Introducción al análisis de datos | |
| **Profesor:** MGTI. María Eugenia Guerrero Chan | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Características** | **Casos de aplicación** | **Lenguajes y herramientas** |
| **Inteligencia Artificial** | * Razonamiento y toma de decisiones basados en datos. * Capacidad de aprendizaje y adaptación. * Procesamiento de lenguaje natural. * Reconocimiento de patrones. | * Asistentes virtuales (Siri, Alexa, etc.) * Sistemas de recomendación (Netflix, Amazon, etc.) * Vehículos autónomos | * Python (con bibliotecas como TensorFlow, PyTorch, Scikit-learn) * R * Java (con bibliotecas como Deeplearning4j, Weka) * Prolog (para sistemas expertos) * Herramientas de desarrollo de IA (Google Cloud AI, IBM Watson, etc.) |
| **Big Data** | * Grandes volúmenes de datos (estructurados y no estructurados) * Variedad de datos (numéricos, textuales, multimedia, etc.) * Velocidad de generación y procesamiento de datos. * Veracidad de los datos. | * Análisis de datos en tiempo real (redes sociales, sensores, etc.) * Análisis de patrones de comportamiento (comercio electrónico, banca, etc.) * Optimización de operaciones (cadenas de suministro, logística, etc.) * Investigación científica (genómica, astrofísica, etc.) | * Apache Hadoop * Apache Spark * NoSQL (MongoDB, Cassandra, HBase) * SQL (para bases de datos relacionales) * Herramientas de visualización (Tableau, Power BI, etc.) |
| **Machine Learning** | * Aprendizaje a partir de datos. * Detección de patrones y relaciones complejas. * Mejora continua del rendimiento con más datos. * Capacidad de generalización. | * Reconocimiento de imágenes y voz * Detección de fraudes * Predicción de demanda y precios * Personalización de contenido y recomendaciones | * Python (con bibliotecas como Scikit-learn, TensorFlow, Keras) * R * MATLAB * Java (con bibliotecas como Weka, Deeplearning4j) * Herramientas de desarrollo de ML (Google Cloud AI, Amazon SageMaker, etc.) |
| **Data Mining** | * Descubrimiento de patrones y conocimiento a partir de grandes volúmenes de datos. * Integración y limpieza de datos. * Selección y transformación de características relevantes. * Interpretación de resultados. | * Análisis de ventas y marketing * Detección de fraudes y riesgos * Optimización de procesos empresariales * Investigación científica y médica | * Python (con bibliotecas como Pandas, NumPy, Scikit-learn) * R * SQL (para consultas y transformaciones de datos) * Herramientas de minería de datos (Weka, RapidMiner, Orange) * Herramientas de visualización (Tableau, Power BI, etc.) |

Con base a la información que pusiste en la tabla **selecciona** y **justifica** con tus propias palabras una herramienta de Inteligencia Artificial, Machine Learning, Data Mining y una de Big Data como apoyo para la toma de decisiones.

La justificación es de manera individual, es decir; una para Inteligencia Artificial, Machine Learning, Data Mining y una de Big Data (cantidad de palabras entre 250 y 300 o más si así lo requiere).

**II.-Procedimientos y resultados:** (Poner aquí la estructura y orden de la información)

**Inteligencia Artificial**

Como herramienta de Inteligencia Artificial para respaldar la toma de decisiones, Google Cloud AI destaca por su versatilidad y potencia. Esta plataforma brinda un amplio abanico de servicios de IA, incluyendo reconocimiento de voz, visión computarizada, procesamiento de lenguaje natural y aprendizaje automático. Una de sus principales fortalezas radica en su capacidad para procesar y analizar grandes volúmenes de datos de manera eficiente, permitiendo obtener información valiosa y tomar decisiones más fundamentadas. Además, Google Cloud AI cuenta con modelos preentrenados y APIs robustas que facilitan la integración de la IA en diversas aplicaciones y sistemas. Esta herramienta resulta particularmente útil para empresas que buscan aprovechar las tecnologías de IA sin tener que construir y entrenar sus propios modelos desde cero, lo que puede ahorrar tiempo y recursos.

**Big Data**

Apache Spark se destaca como una herramienta sobresaliente para el manejo de Big Data y la toma de decisiones. Esta plataforma de código abierto está diseñada para procesar grandes volúmenes de datos de manera rápida y eficiente. Spark ofrece una arquitectura de procesamiento en memoria que lo hace significativamente más veloz que Apache Hadoop para ciertas operaciones. Además, proporciona una API unificada que permite trabajar con diversos formatos de datos, como estructurados, semi-estructurados y no estructurados. Spark también cuenta con bibliotecas específicas para machine learning, procesamiento de datos en tiempo real y análisis de grafos, lo que lo convierte en una herramienta versátil para diferentes casos de uso. La capacidad de Spark para procesar y analizar grandes conjuntos de datos de manera rápida y eficiente es crucial para la toma de decisiones informadas en entornos de Big Data.

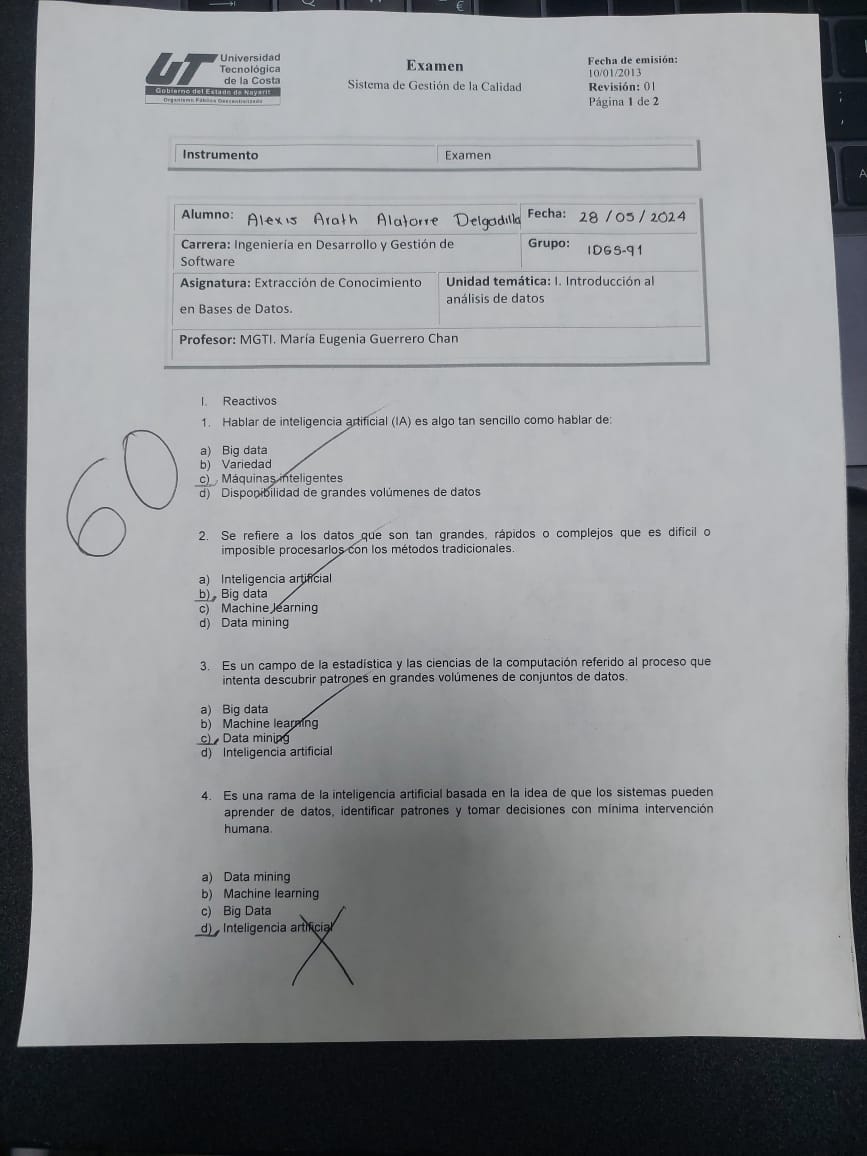
**Machine Learning**

TensorFlow se destaca como una herramienta sobresaliente para el desarrollo de soluciones de Machine Learning y la toma de decisiones. Desarrollada por Google, TensorFlow es una biblioteca de código abierto para machine learning y computación numérica. Una de las principales ventajas de TensorFlow es su capacidad para construir y entrenar modelos de aprendizaje profundo de alta complejidad, lo que la convierte en una herramienta poderosa para tareas como reconocimiento de imágenes, procesamiento de lenguaje natural y predicción de series de tiempo. Además, TensorFlow cuenta con una amplia comunidad de desarrolladores y una gran cantidad de recursos y documentación disponibles, lo que facilita su adopción y uso. La escalabilidad y flexibilidad de TensorFlow también la hacen adecuada para proyectos de machine learning a gran escala, lo que es crucial para la toma de decisiones basadas en datos en entornos empresariales.

**Data Mining**

En el ámbito del Data Mining, RapidMiner se destaca como una herramienta sobresaliente. Esta plataforma de código abierto ofrece una amplia gama de funcionalidades para el procesamiento de datos, la visualización y el modelado predictivo. RapidMiner cuenta con una interfaz gráfica de usuario intuitiva que facilita el diseño y ejecución de flujos de trabajo de minería de datos, lo que la convierte en una opción accesible incluso para usuarios sin experiencia en programación. Además, RapidMiner tiene una extensa biblioteca de operadores predefinidos para tareas como preprocesamiento de datos, selección de características, clasificación, agrupamiento y regresión. Esta herramienta también admite la integración con diferentes fuentes de datos y formatos de archivos, lo que la hace versátil para diferentes escenarios de minería de datos. La capacidad de RapidMiner para automatizar el proceso de descubrimiento de conocimiento a partir de datos es particularmente útil para la toma de decisiones basadas en información valiosa extraída de grandes conjuntos de datos.

## **EXAMEN**



# **UNIDAD 2**

## **Instalación del lenguaje R**

|  |  |
| --- | --- |
| **Instrumento** | *Práctica de ejercicios* |

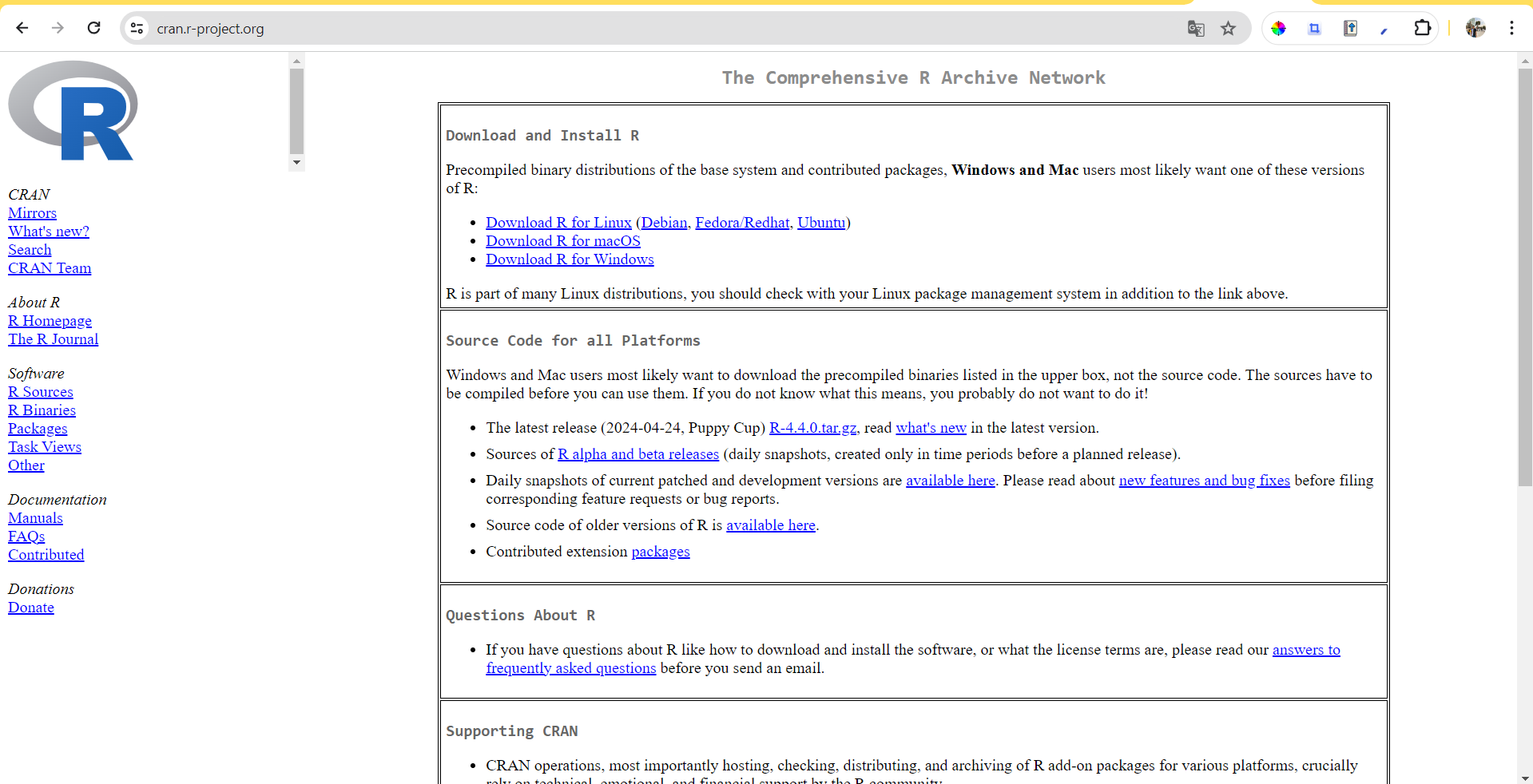
|  |  |  |
| --- | --- | --- |
| **Alumno**: Alexis Arath Alatorre Delgadillo | | **Fecha: 30/05/2024** |
| **Carrera: IDGS** | | **Grupo: IDGS91** |
| **Asignatura:** Extracción de Conocimiento en Base de Datos | **Unidad temática: Unidad** | |
| **Profesor: MGTI. María Eugenia Guerrero Chan** | | |

I.- Ejercicios a resolver:

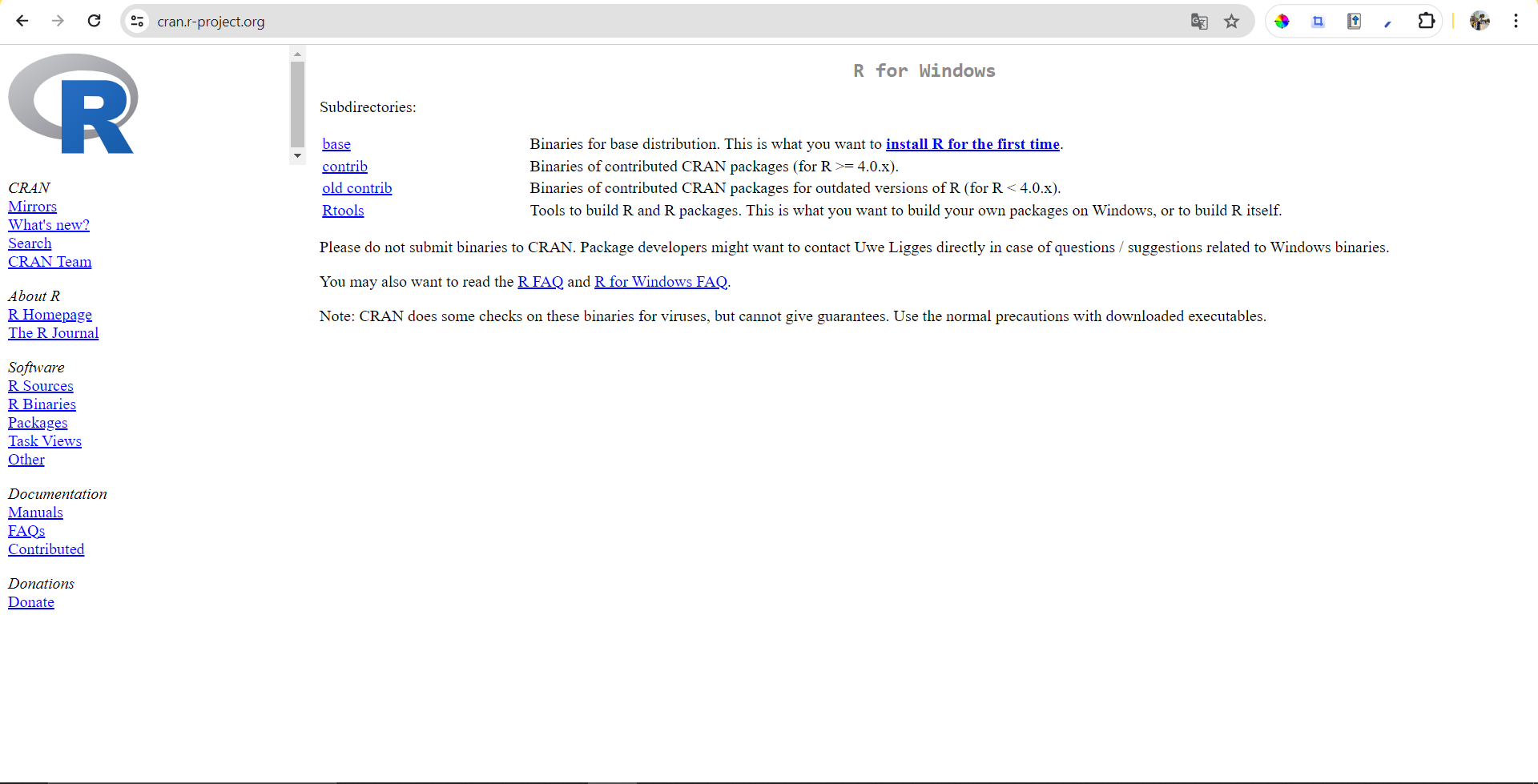
II.-Procedimientos y resultados:

**INSTALACIÓN DEL LENGUAJE R**

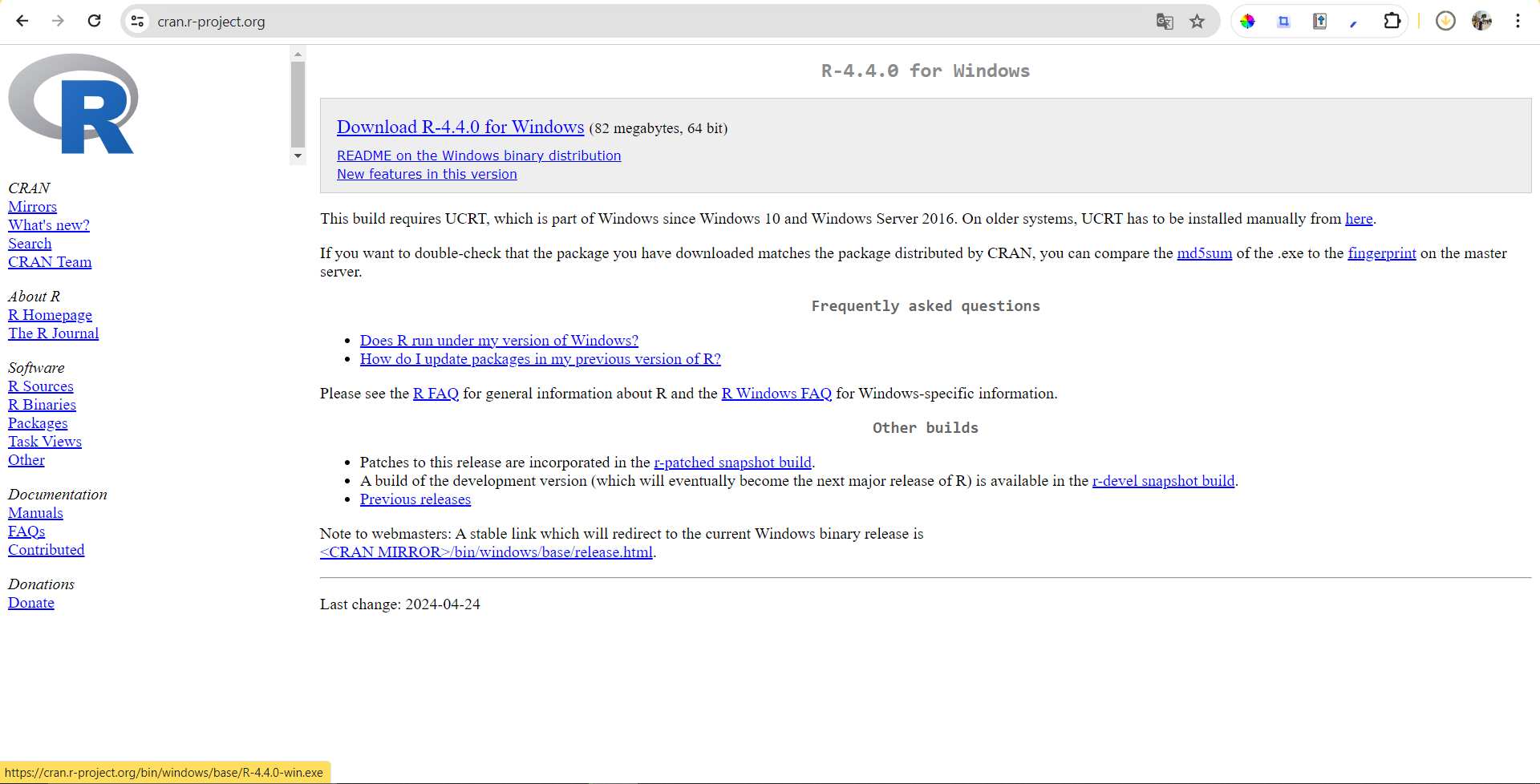
Para descargar R debes acceder a CRAN (Comprehensive R Archive Network) que es el repositorio central de software de R. cliquea en Download R for Windows



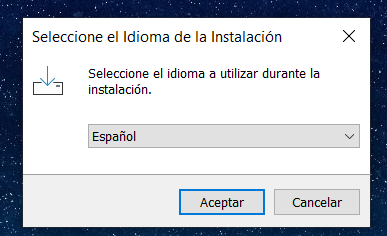
Y cambiará al siguiente apartado y cliqueamos en “Install R for the first time”



Luego otro clic en Download R-4.4.0 for Windows y en unos momentos empezará la descarga del instalador.



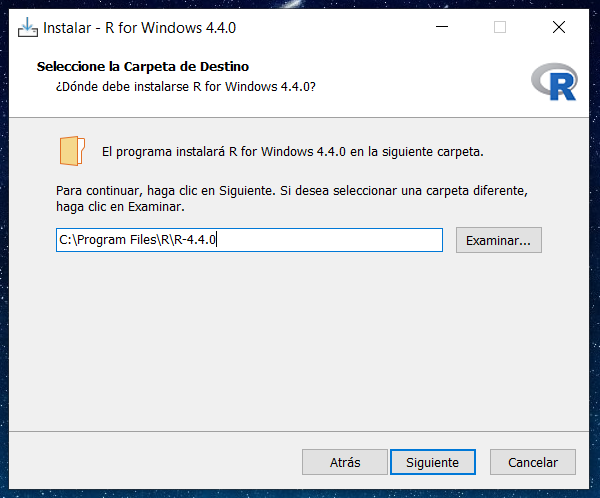
Una vez descargado, ejecutamos el archivo y seleccionamos el idioma preferido, luego “Aceptar”.



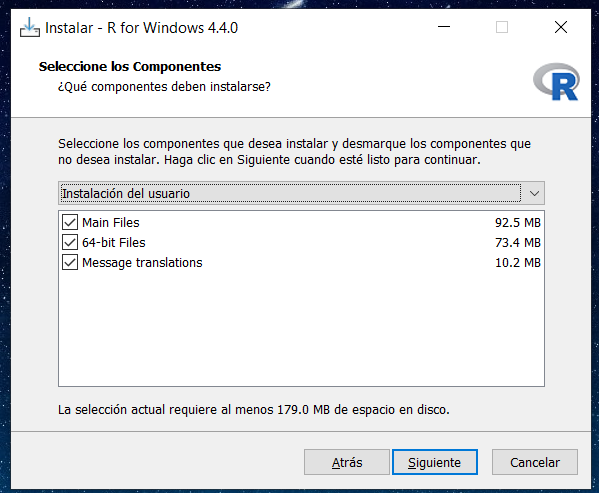
Después en “siguiente”.



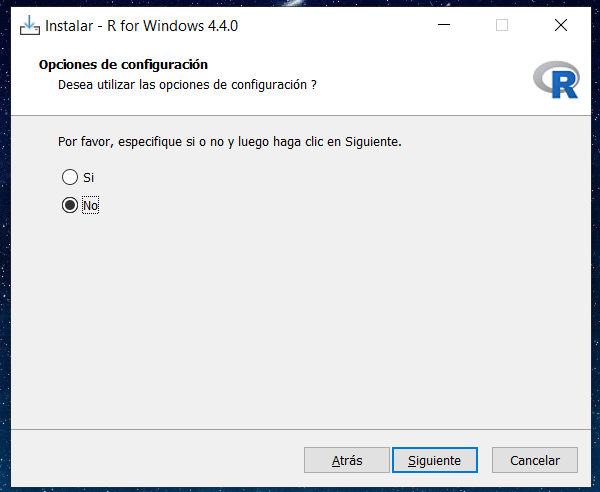
Ahora elegir en donde se va a instalar los archivos del archivo y cliquear en “siguiente”.



Nuevamente en siguiente para seguir, ya que comentará a instalarse unos componentes necesarios para el funcionamiento de archivo, y nuevamente “siguiente”.



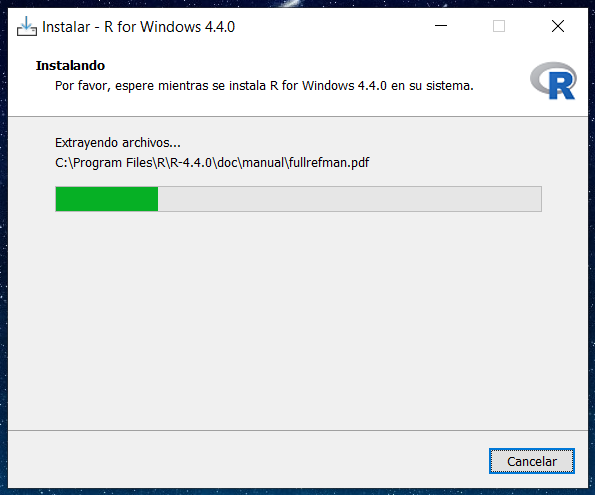
Nuevamente “siguiente”, elegimos que no la opción por defecto.



Este apartado tienes decir la ruta la cual se colocará el acceso directo del programa, “siguiente”.



En seguida comenzará a instalación del programa.

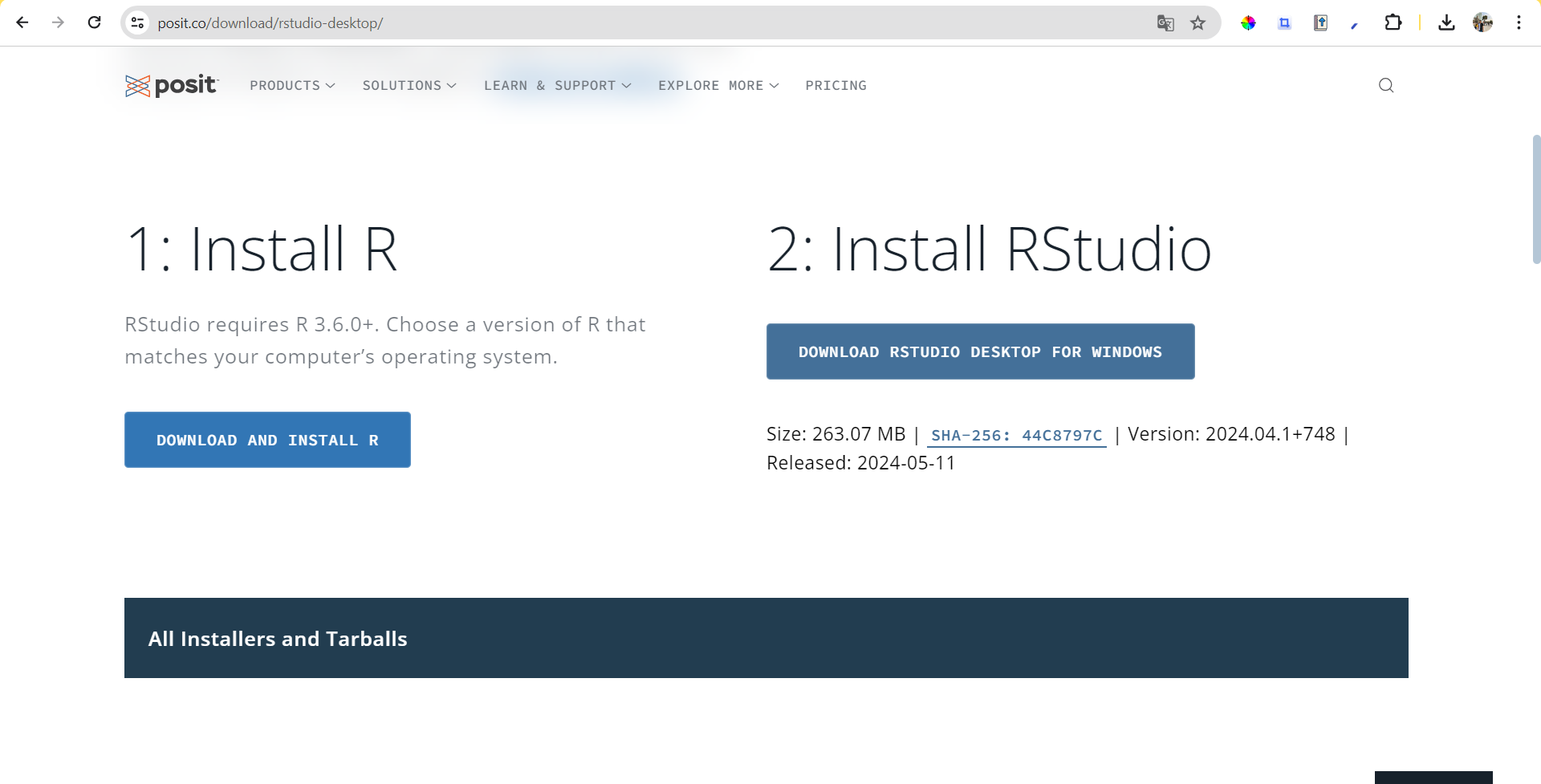


Una vez finalizado mostrará el siguiente mensaje lo cual la instalación a sido exitosa.

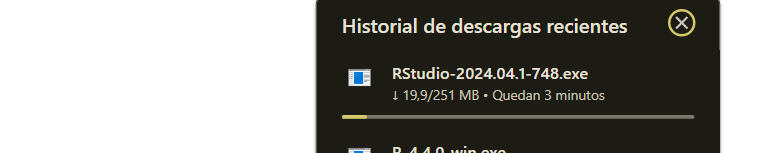


## **Instalación de Rstudio**

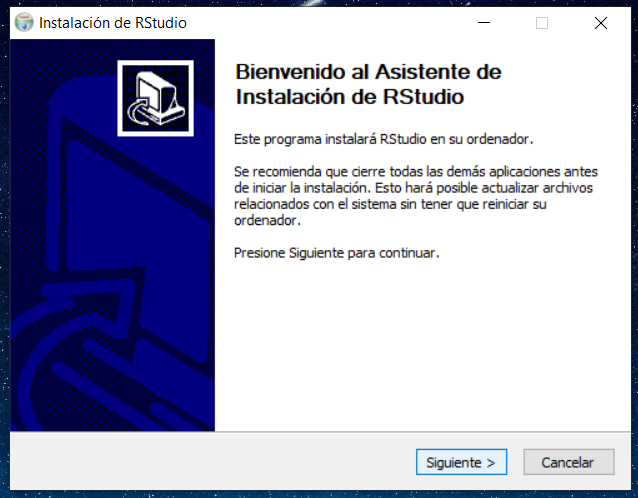
Accedemos a la siguiente pagina de **posit** la cual nos dice los pasos para la instalación del programa la cual es necesario tener instalado **R** para posterior a ello instalar **Rstudio,** cliqueamos en “DOWNLOAD RSTUDIO DESKTOP WINDOWS”, sí ya se tiene instalado R, de lo contrario primeramente instala R.



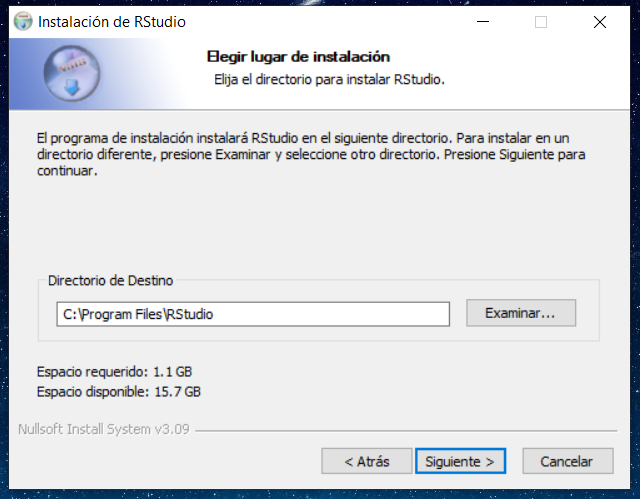
En seguida comenzará la descarga del programa Rstudio, una vez descargado cliqueamos en el programa.



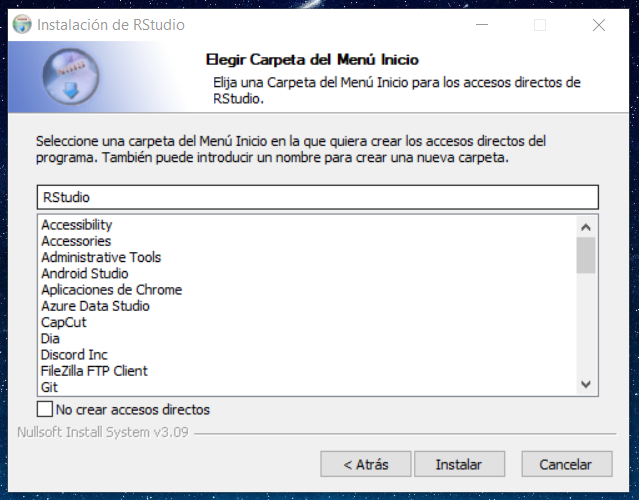
Y nos abrirá la siguiente ventana para la instalación de Rstudio



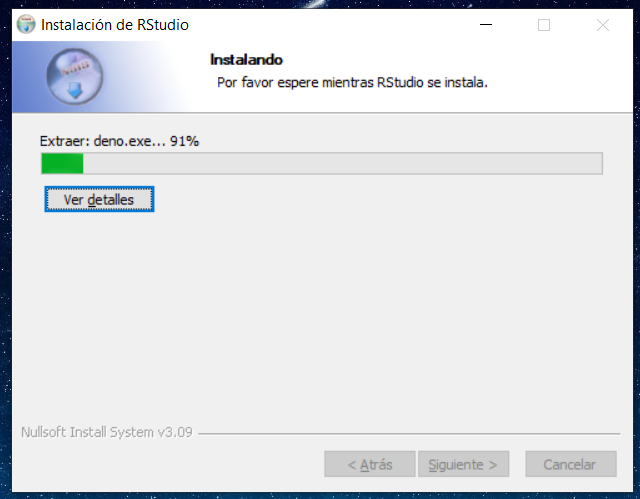
Ahora elegimos la ruta para guardar los archivos necesarios del programa y cliqueamos en “Siguiente”.



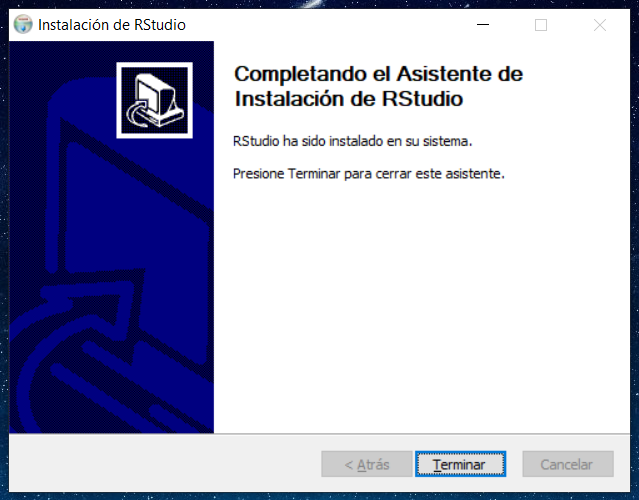
Y Cliqueamos en “finalizar”, ya que este apartado es para el acceso directo a Rstudio.



Y comenzará la instalación del programa



Y una vez instalado mostrará el siguiente mensaje, indicando que el programa a sido instalado correctamente.



**Bibliografía**

*The Comprehensive R Archive Network*. (2024). R-Project.org. https://cran.r-project.org/

‌*Posit*. (2024, May 7). Posit. https://posit.co/download/rstudio-desktop/

## **Mapa conceptual**

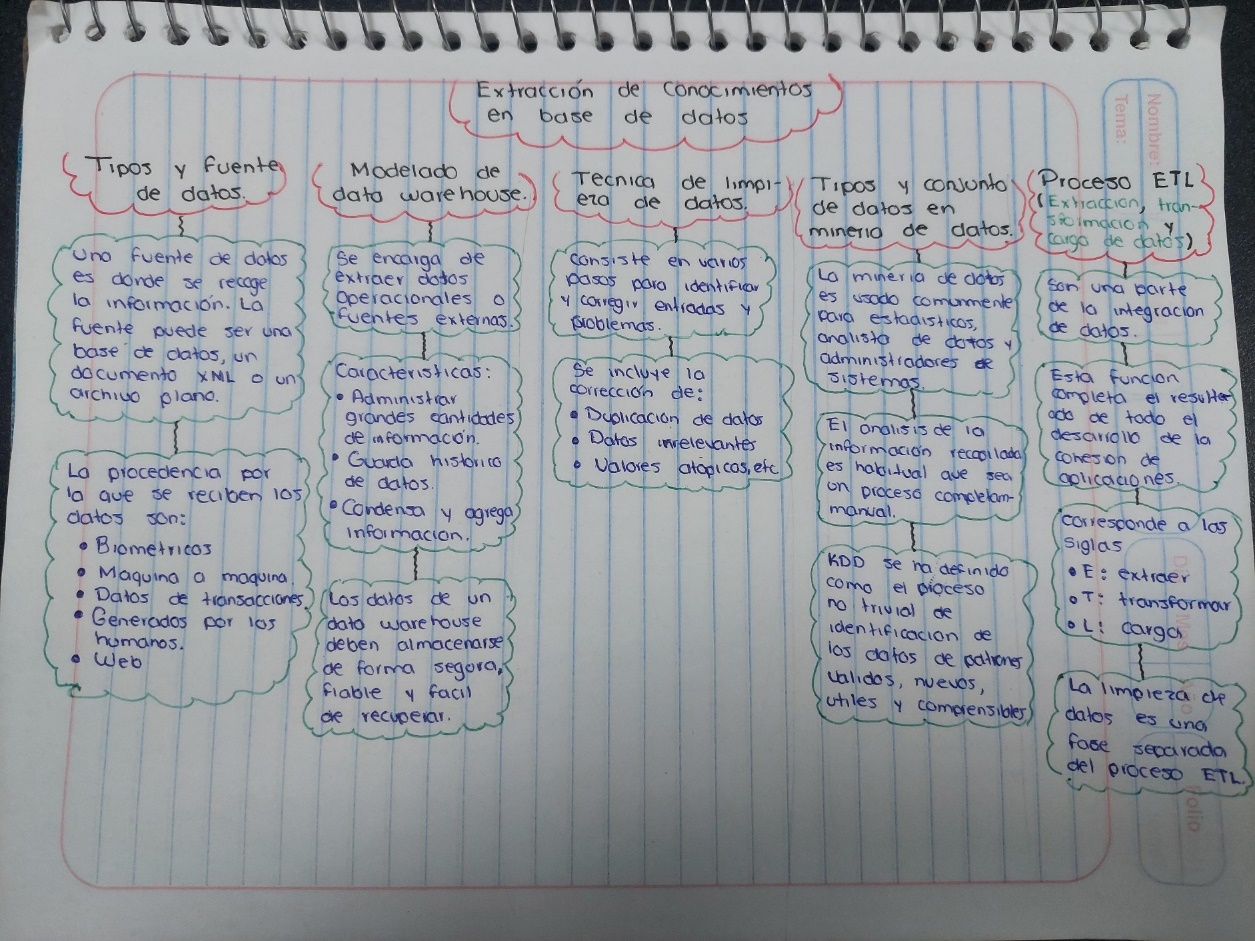
|  |  |
| --- | --- |
| **Instrumento** | *Mapa conceptual o Cuadro sinóptico* |

|  |  |  |
| --- | --- | --- |
| **Alumno**: Alexis Arath Alatorre Delgadillo | | **Fecha: 06/06/2024** |
| **Carrera: IDGS** | | **Grupo: IDGS91** |
| **Asignatura:** Extracción de conocimientos en base de datos | **Unidad temática: Unidad 2** | |
| **Profesor: MGTI: María Eugenia Guerrero Chan** | | |

I.-Título:

Extracción de conocimientos en base de datos

II.- Mapa conceptual o cuadro sinóptico:

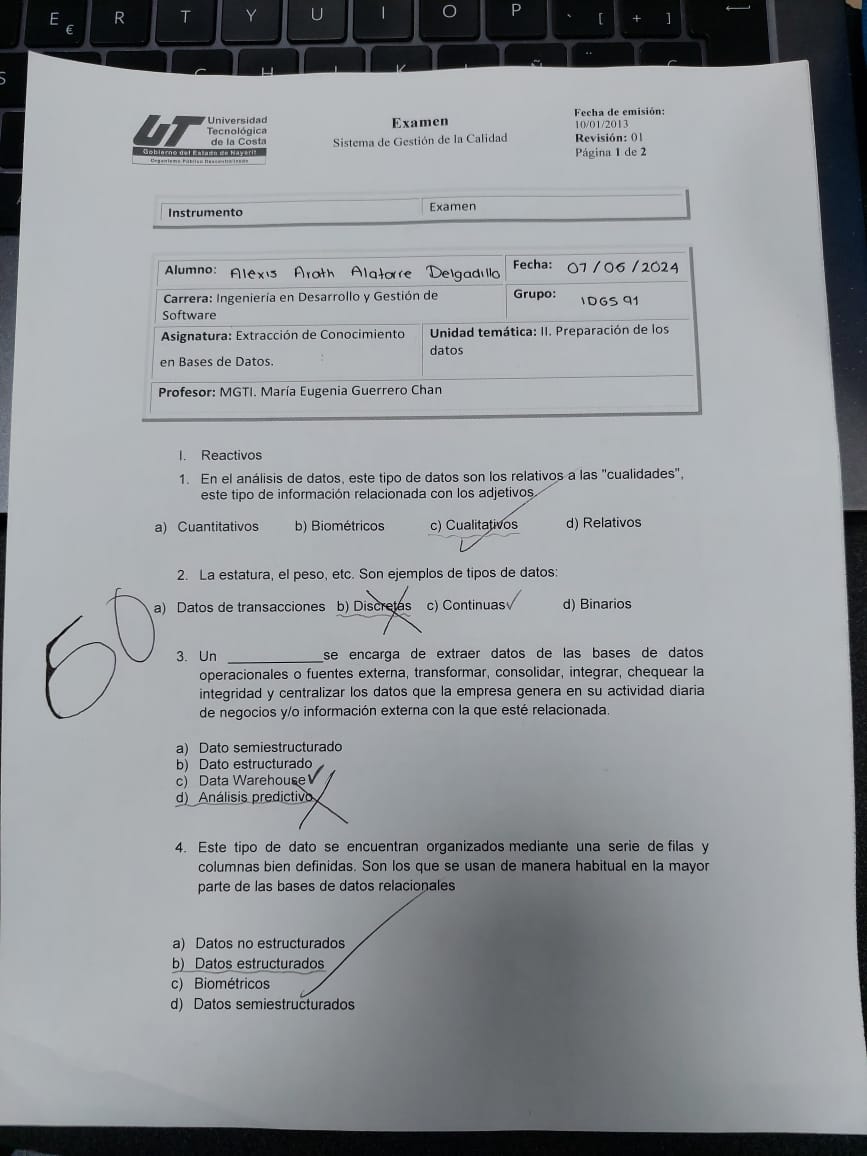


III.- Bibliografía

ECBD\_U2\_clase.pdf. (2019). *ECBD\_U2\_clase.pdf*. Google Docs. https://drive.google.com/file/d/1M\_tmEiuFlEDXEeg9xUXPPvTJRAN3EynV/view

‌

## **EXAMEN**



Bibliografía

*Inteligencia Artificial: Qué es y Por Qué Importa*. (2023). Sas.com. https://www.sas.com/es\_mx/insights/analytics/what-is-artificial-intelligence.html

*¿Qué es Big Data y por qué es importante? - Intel*. (2022). Intel. https://www.intel.la/content/www/xl/es/artificial-intelligence/analytics/what-is-big-data.html

‌*Aprendizaje automático: Qué es y por qué es importante*. (2023). Sas.com. https://www.sas.com/es\_mx/insights/analytics/machine-learning.html

‌*Minería de datos: Qué es y por qué es importante*. (2018). Sas.com. https://www.sas.com/es\_mx/insights/analytics/data-mining.html

‌