**Ingenieria en Desarrollo y Gestión de Software Multiplataforma**

**Universidad Tecnologica de la costa**

Portafolio de extracción de conocimiento en base de datos

Profesor: M.G.T.I Maria Eugenia Guerrera Chan

Grupo:IDGS91

Alumno: Irving Armando Martinez Torres

Indice

[Unidad 1- Introducción al analisis de datos. 2](#_Toc26946)

[Examen 2](#_Toc24118)

[Resumen 3](#_Toc25121)

[Practica de ejercicios 7](#_Toc10617)

[I.- Ejercicios a resolver: 7](#_Toc10203)

[II.-Procedimientos y resultados: 9](#_Toc1642)

[Ejercicios de frecuencias 12](#_Toc6349)

[Unidad 2- Preparación de datos 13](#_Toc26081)

[Examen 13](#_Toc1693)

[Mapa conceptual 14](#_Toc3697)

[Instalación de Lenguaje R y RStudio 14](#_Toc29452)

[I.- Ejercicios a resolver: 15](#_Toc25193)

[II.-Procedimientos y resultados: 15](#_Toc18548)

[Unidad 3. Análisis Supervisado 24](#_Toc5853)

[Examen 24](#_Toc4921)

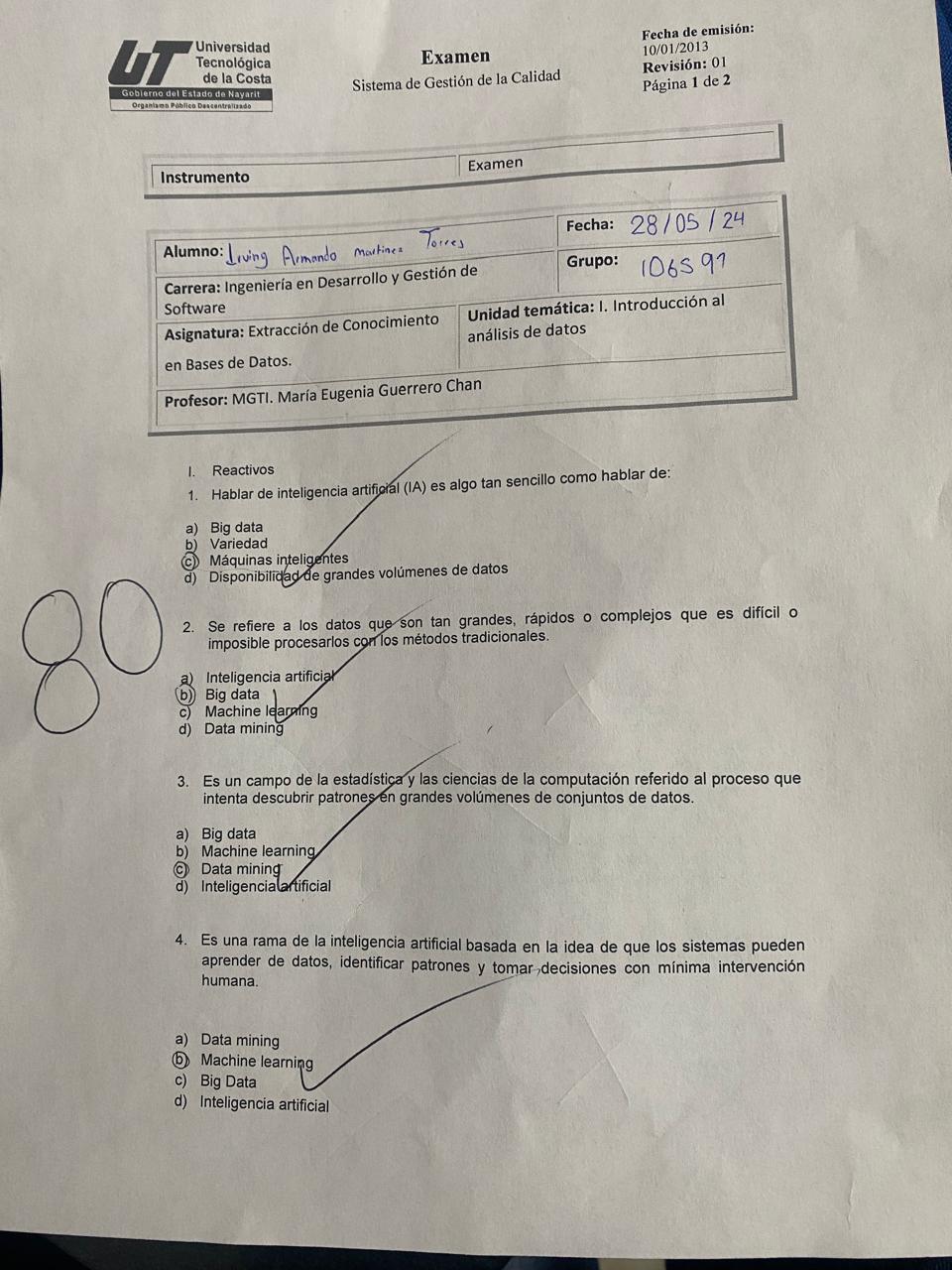
[Practica ejercicios1\_U3\_ECBD 25](#_Toc27602)

[Practica ejercicios2\_U3\_ECBD 27](#_Toc11183)

[Practica ejercicios3\_U3\_ECBD 29](#_Toc14093)

# Unidad 1- Introducción al analisis de datos.

## Examen



## Resumen

|  |  |
| --- | --- |
| **Instrumento:** | Resumen |

|  |  |  |
| --- | --- | --- |
| **Alumno**: Irving Armando Martinez Torres | | **Fecha: 02/05/25** |
| **Carrera: Ingeniera en gestión y desarrollo de software multiplataforma** | | **Grupo: IDGS91** |
| **Asignatura:** Extracción del Conocimiento en Base de Datos | **Unidad temática:** I. Introducción al análisis de datos | |
| **Profesor:** MGTI María Eugenia Guerrero Chan | | |

Instrucciones:

* Escribir una *Introducción* sobre las Metodologías para el análisis de datos (entre 200 y 250 palabras) y en *Desarrollo* la información que se está pidiendo que investigues.

1. Título: Metodologías para el análisis de datos
2. Contenido (Introducción y Desarrollo):
3. Que es un Proyecto de Análisis de Datos
4. Características de un Proyecto de Análisis de Datos.
5. Que es una Metodología para el análisis de datos.
6. Principales Metodologías o Métodos de análisis de datos (definición de cada uno de ellos y ejemplos donde se aplican).
7. Bibliografía:

Instrucciones:

**Introducción**

En el vasto campo de la ciencia de datos, las metodologías para el análisis de datos juegan un papel crucial, ya que guían a los analistas y científicos de datos en el proceso de transformar datos crudos en insights accionables y decisiones informadas. Estas metodologías no solo proporcionan un marco de trabajo estructurado para abordar problemas complejos, sino que también ayudan a asegurar la precisión y la eficacia de los resultados obtenidos. Entre las metodologías más prominentes se incluyen el análisis estadístico, el aprendizaje automático, la minería de datos, y la visualización de datos.

Cada una de estas metodologías tiene sus propias técnicas y herramientas específicas, elegidas en función de la naturaleza y complejidad del conjunto de datos y del objetivo del análisis. Por ejemplo, el análisis estadístico puede incluir métodos como la regresión y la prueba de hipótesis para identificar tendencias y patrones. Por otro lado, el aprendizaje automático aprovecha algoritmos que pueden aprender de los datos y hacer predicciones o clasificaciones. La minería de datos se enfoca en el proceso de descubrir patrones y relaciones ocultas en grandes volúmenes de datos, mientras que la visualización de datos ayuda a interpretar los resultados de manera intuitiva mediante gráficos y diagramas.

La elección de una metodología adecuada depende crucialmente de la claridad del objetivo analítico, la calidad y el volumen de los datos disponibles, y las capacidades técnicas del equipo de análisis. Este enfoque estructurado no solo maximiza la eficiencia del análisis, sino que también minimiza los riesgos de errores y malinterpretaciones, asegurando que las decisiones basadas en datos sean tan confiables como sea posible.

1. **Título: Metodologías para el análisis de datos**
2. Contenido (Introducción y Desarrollo):

**1. ¿Qué es un Proyecto de Análisis de Datos?**

Un proyecto de análisis de datos es una iniciativa estructurada donde se recopilan, procesan y analizan datos para extraer información útil, descubrir patrones, hacer predicciones, o informar decisiones. El objetivo es transformar los datos crudos en conocimientos que puedan ser utilizados para apoyar la toma de decisiones y resolver problemas específicos. Este tipo de proyecto se emplea en una variedad de campos como negocios, salud, tecnología, y más, donde los datos son un recurso valioso para la comprensión y mejora de procesos, productos y servicios.

**2. Características de un Proyecto de Análisis de Datos**

Objetivo Claro: Cada proyecto debe tener una meta bien definida que oriente todas las fases del análisis.

Recolección de Datos: Involucra la acumulación de datos pertinentes desde diversas fuentes, asegurando que sean relevantes y de alta calidad.

Análisis Riguroso: Utilización de técnicas estadísticas y computacionales para interpretar los datos recogidos.

Iteratividad: Los proyectos de análisis de datos a menudo requieren ajustes y repetición de análisis para refinar los resultados.

Interdisciplinariedad: Combina habilidades de varios dominios, como estadística, programación y conocimiento específico del sector.

Presentación de Resultados: Los resultados son comunicados a los stakeholders de manera comprensible, a menudo mediante visualizaciones.

**3. ¿Qué es una Metodología para el Análisis de Datos?**

Una metodología para el análisis de datos es un conjunto sistemático de principios y técnicas utilizadas para dirigir el proceso de recolección, análisis e interpretación de datos. Sirve como guía para asegurar que el análisis sea estructurado, eficiente y que los resultados sean válidos y aplicables. La metodología adecuada es esencial para manejar la complejidad de los datos y las preguntas de investigación, ayudando a los analistas a evitar errores y a maximizar la utilidad de los datos recogidos.

**4. Principales Metodologías o Métodos de Análisis de Datos**

Análisis Estadístico: Involucra técnicas como la regresión y análisis de varianza para explorar y modelar los datos. Ejemplo: en estudios de mercado, para evaluar las relaciones entre características del consumidor y sus preferencias de compra.

Aprendizaje Automático: Usa algoritmos que aprenden de los datos para hacer predicciones o clasificaciones. Ejemplo: en la detección de fraudes, donde los modelos predictivos identifican transacciones sospechosas.

Minería de Datos: Proceso de explorar grandes bases de datos para encontrar patrones ocultos. Ejemplo: en el retail, para descubrir patrones de compra y optimizar las estrategias de inventario.

Visualización de Datos: Técnicas para representar datos de manera gráfica, facilitando su interpretación. Ejemplo: dashboards en tiempo real para monitorear indicadores clave de desempeño en una empresa.Bibliografía:

**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, Inc.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers.

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Nussbaumer Knaflic, C. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

## Practica de ejercicios

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

|  |  |  |
| --- | --- | --- |
| **Alumno**: Irving Armando Martinez Torres | | **Fecha: 21 de mayo 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 | | |

**I.- Ejercicios a resolver:**

**Instrucciones:**

Observa la tabla siguiente e investiga la información que se pide y con base a la investigación llena la tabla.

**VALOR (50%)**

Tabla 1.- Llena la siguiente tabla con la información que se pide.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Características** | **Casos de aplicación** | **Lenguajes y herramientas** |
| Inteligencia Artificial | - Simula la inteligencia humana  - Automatiza tareas - Toma de decisiones autónomas  - Reconocimiento de patrones | - Asistentes virtuales (como Siri y Alexa)  - Diagnóstico médico  - Vehículos autónomos | - Python (TensorFlow, PyTorch)  - Lisp  - Prolog |
| Big Data | - Gestión de grandes volúmenes de datos  - Alta velocidad de procesamiento  - Diversidad de datos  - Veracidad de la información | - Análisis de tendencias del mercado  - Predicción de fallos en maquinaria  - Análisis de comportamiento de clientes | - Hadoop  - Spark  - NoSQL (MongoDB, Cassandra) |
| Machine Learning | - Algoritmos que aprenden de los datos - Modelos predictivos  - Automatización de tareas  - Mejora continua | - Recomendadores de productos (Amazon, Netflix)  - Detección de fraudes - Análisis de sentimientos | - Python (scikit-learn, Keras)  - R  - MATLABs |
| Data Mining | Recomendadores de productos (Amazon, Netflix) - Detección de fraudes - Análisis de sentimientos | - Marketing dirigido - Análisis de cesta de la compra - Segmentación de clientes | - SQL  - RapidMiner  - Weka |

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:**

**Inteligencia Artificial**

Herramienta Seleccionada: TensorFlow

**Justificación:** TensorFlow es una de las bibliotecas más utilizadas en el campo de la Inteligencia Artificial, desarrollada por Google Brain. Esta herramienta es extremadamente poderosa para el desarrollo y entrenamiento de modelos de aprendizaje profundo (deep learning). TensorFlow permite la creación de redes neuronales complejas y su optimización mediante técnicas avanzadas de machine learning. La capacidad de TensorFlow para manejar grandes volúmenes de datos y realizar operaciones matemáticas de manera eficiente la hace ideal para aplicaciones como el reconocimiento de imágenes, procesamiento de lenguaje natural y sistemas de recomendación. Además, su flexibilidad y escalabilidad permiten su implementación tanto en dispositivos móviles como en grandes clusters de servidores, lo que la convierte en una herramienta esencial para la toma de decisiones basada en datos masivos y complejos. La robustez de TensorFlow, junto con una gran comunidad de desarrolladores y abundante documentación, asegura que las organizaciones puedan construir y desplegar soluciones de IA efectivas y eficientes.

**Big Data**

**Herramienta Seleccionada:** Apache Hadoop

**Justificación:** Apache Hadoop es una de las plataformas más prominentes y ampliamente utilizadas para el manejo y análisis de grandes volúmenes de datos. Hadoop permite el procesamiento distribuido de conjuntos de datos extremadamente grandes a través de clusters de computadoras, utilizando un modelo de programación simple. La arquitectura de Hadoop, basada en HDFS (Hadoop Distributed File System) y el modelo de programación MapReduce, permite a las organizaciones almacenar y procesar datos a gran escala de manera eficiente y fiable. Hadoop es ideal para aplicaciones de análisis predictivo, detección de fraudes, análisis de tendencias del mercado y otros casos donde se requiere procesar grandes cantidades de datos rápidamente. Su capacidad de escalar horizontalmente y manejar fallos de hardware lo hace particularmente adecuado para entornos empresariales que necesitan una infraestructura de datos robusta y flexible. El ecosistema de Hadoop también incluye herramientas complementarias como Hive, Pig, y HBase, que amplían sus capacidades y facilitan el desarrollo de soluciones de Big Data complejas.

**Machine Learning**

**Herramienta Seleccionada:** scikit-learn

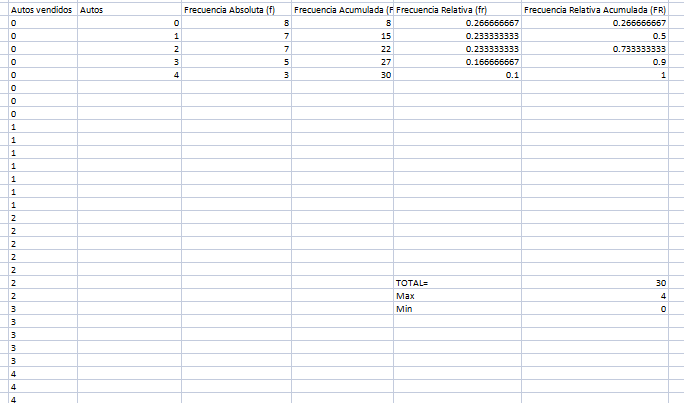
Justificación: scikit-learn es una biblioteca de machine learning de código abierto en Python que proporciona herramientas simples y eficientes para el análisis de datos y la minería de datos. Es conocida por su facilidad de uso y su integración con otros paquetes de la ciencia de datos en Python, como NumPy y pandas. scikit-learn ofrece una amplia gama de algoritmos de aprendizaje supervisado y no supervisado, que incluyen regresión, clasificación, clustering y reducción de dimensionalidad. Esta versatilidad hace que scikit-learn sea una excelente elección para desarrollar modelos predictivos y de análisis. Su diseño API coherente y su documentación exhaustiva facilitan el desarrollo rápido de prototipos y la implementación de soluciones de machine learning en producción. Para la toma de decisiones, scikit-learn permite a las organizaciones desarrollar modelos precisos y robustos que pueden ser utilizados para predecir tendencias, segmentar clientes, detectar anomalías y optimizar operaciones empresariales.

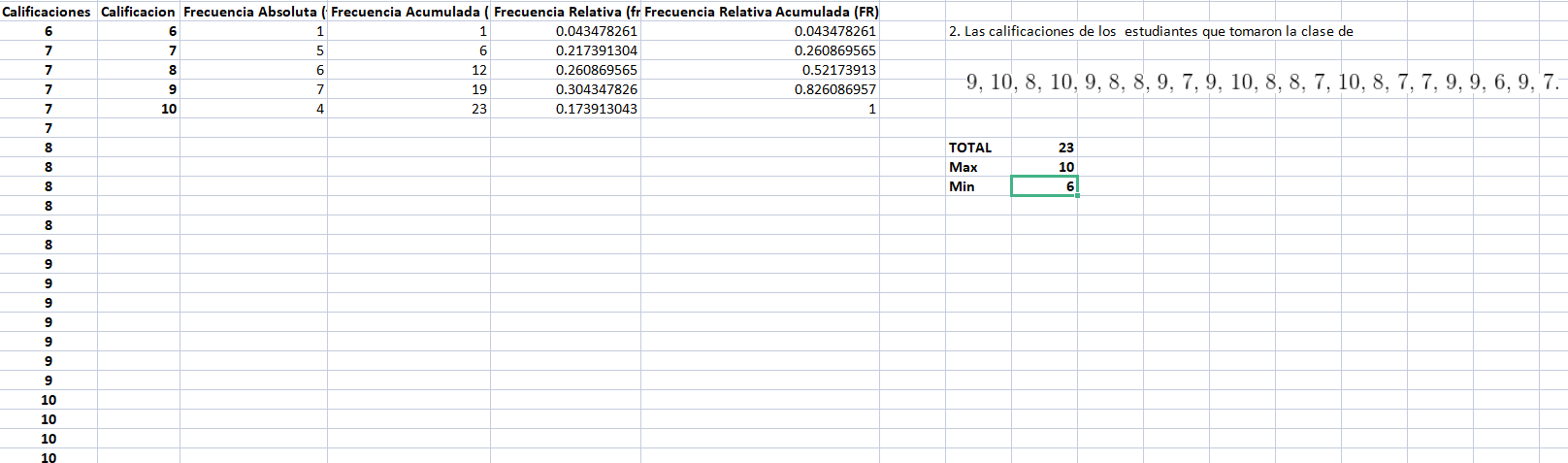
**Data Mining**

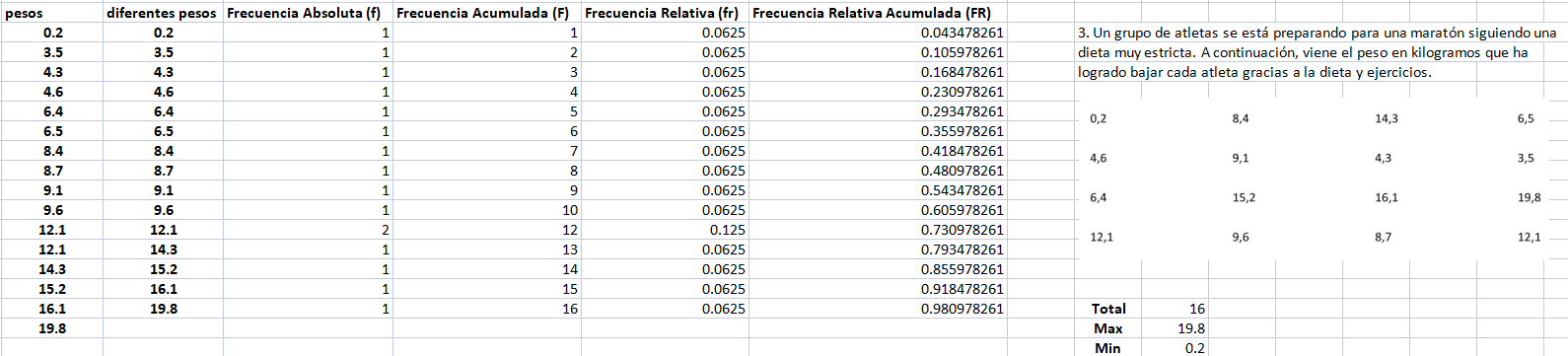
**Herramienta Seleccionada:** RapidMiner

**Justificación:** RapidMiner es una plataforma integral de data mining que permite a los usuarios realizar análisis de datos sin necesidad de habilidades avanzadas de programación. RapidMiner ofrece una interfaz gráfica intuitiva que facilita la construcción, evaluación y despliegue de modelos de minería de datos. Es especialmente útil para la preparación de datos, modelado predictivo, evaluación de modelos y la validación cruzada. RapidMiner soporta una amplia variedad de técnicas de minería de datos, incluyendo árboles de decisión, redes neuronales, máquinas de soporte vectorial y más. Su capacidad de integración con bases de datos y otras herramientas de análisis permite a las organizaciones extraer información valiosa y patrones ocultos de sus datos, lo cual es crucial para la toma de decisiones informadas. Con RapidMiner, las empresas pueden descubrir insights importantes que pueden ayudar en la segmentación de mercado, análisis de cesta de la compra, y en la identificación de oportunidades de negocio, mejorando así su competitividad y eficiencia operativa.

## Ejercicios de frecuencias

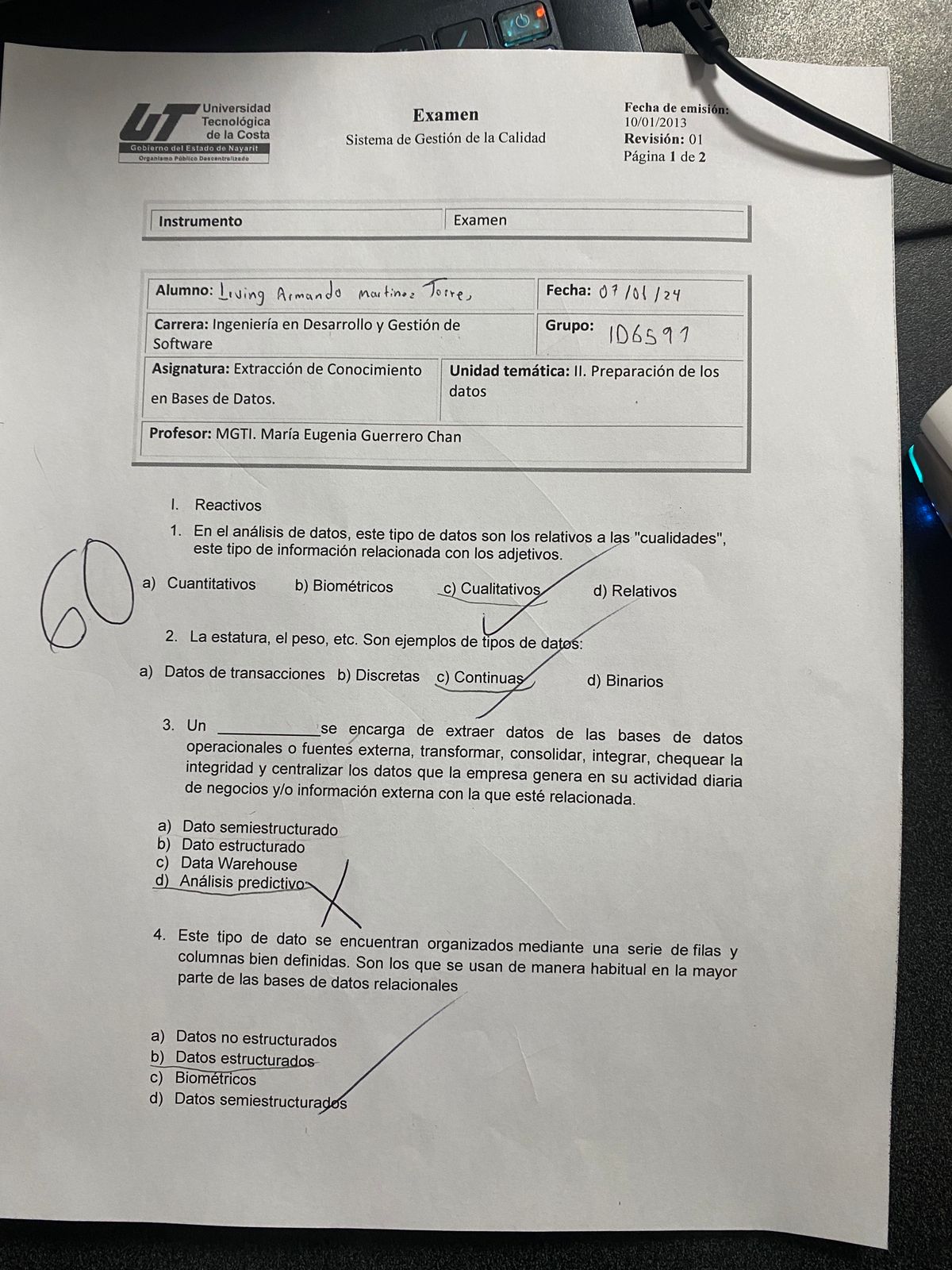




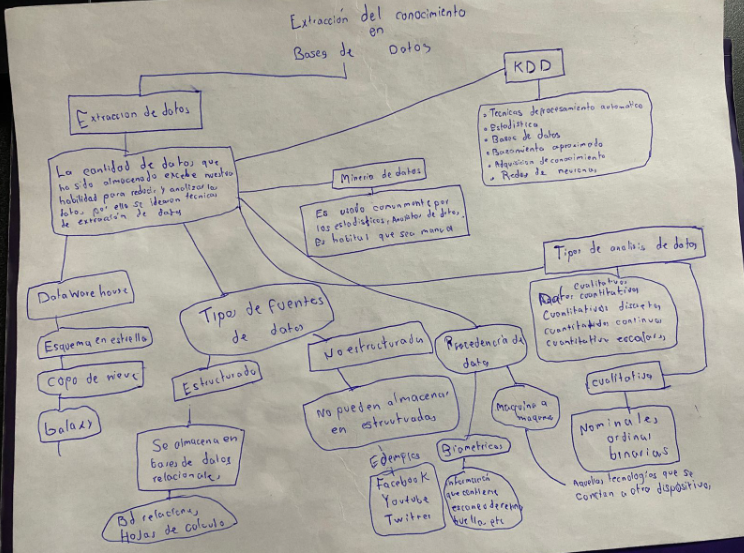


# Unidad 2- Preparación de datos

## Examen



## Mapa conceptual



## Instalación de Lenguaje R y RStudio

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

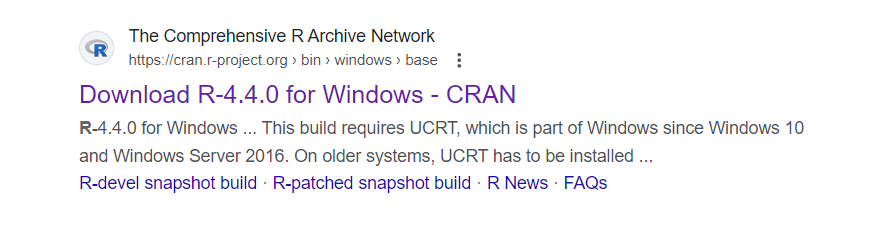
|  |  |  |
| --- | --- | --- |
| **Alumno**: Irving Armando Martinez Torres | | **Fecha:30/05/2024** |
| **Carrera:** Ingenieria en desarrollo de software multiplataforma | | **Grupo:** IDGS91 |
| **Asignatura:** ECBD | **Unidad temática:** Preparación de los datos | |
| **Profesor:** M.G.T.IMaria Eugenia Guerrero Chan | | |

**I.- Ejercicios a resolver:**

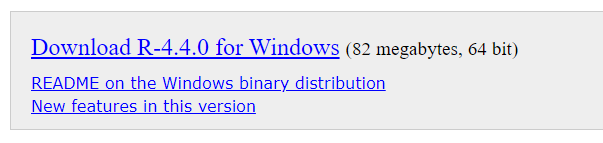
Instalar R y R Studio

**II.-Procedimientos y resultados:**

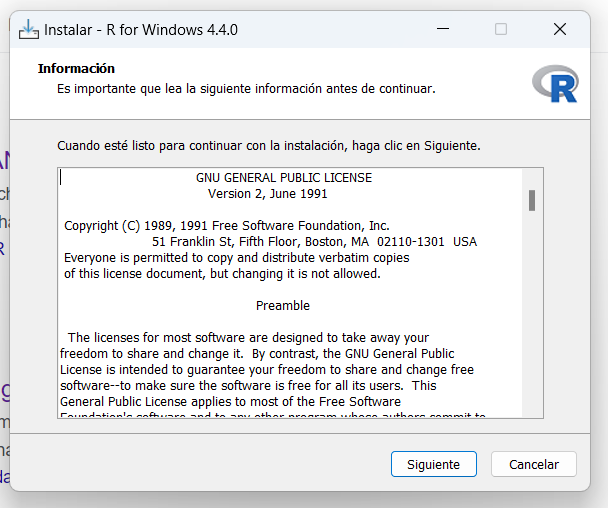
Accedes al primer enlace de la pagina oficial



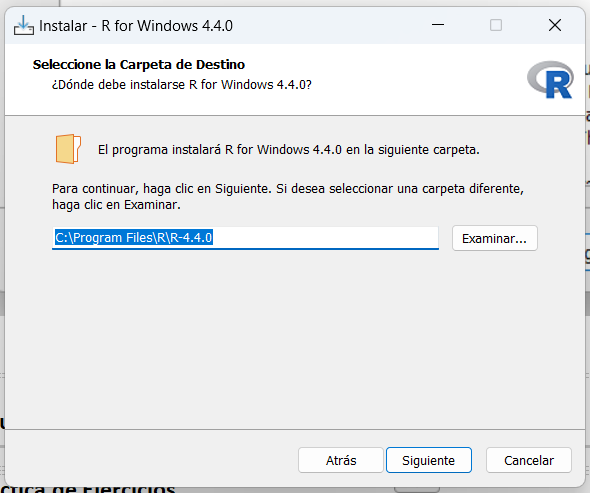
Presionamos aqui para descargar



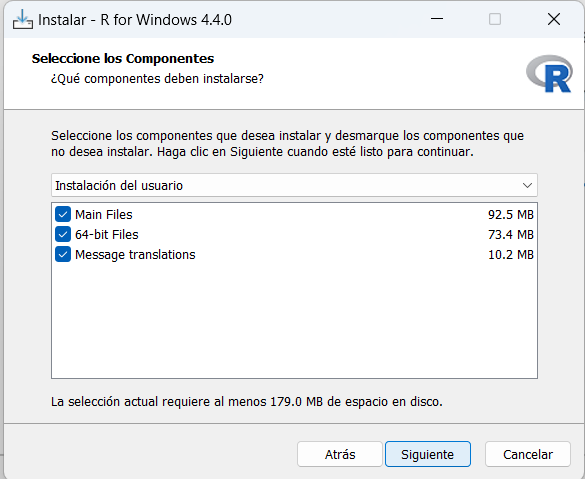
A continuación abrimos el archivo y presionamos en siguiente



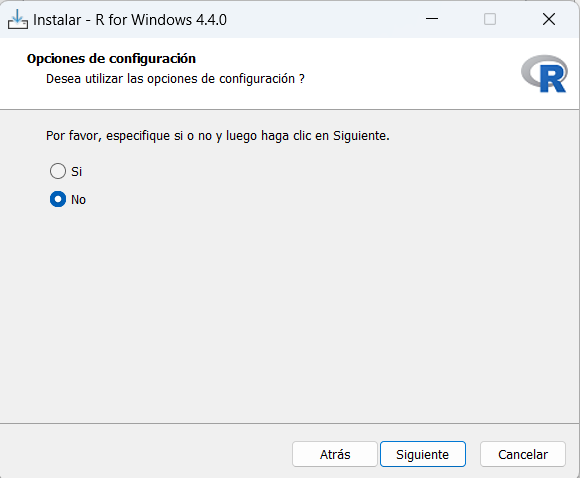
De nuevo en siguiente



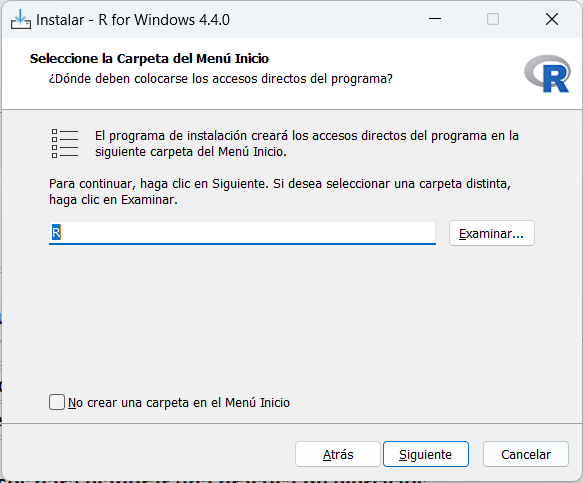
Otra vez en siguiente seleccionando los 3



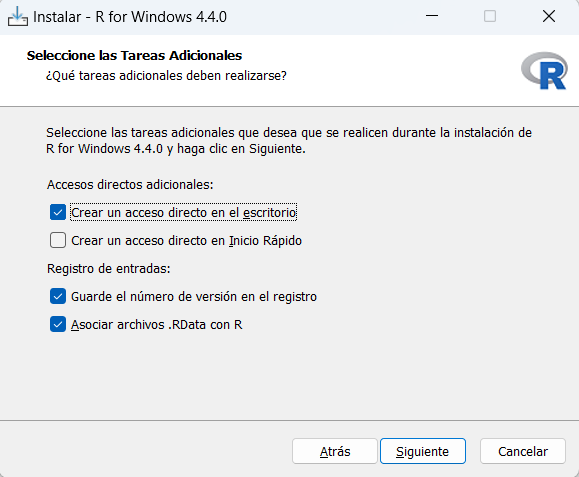
Selecciona la opción no



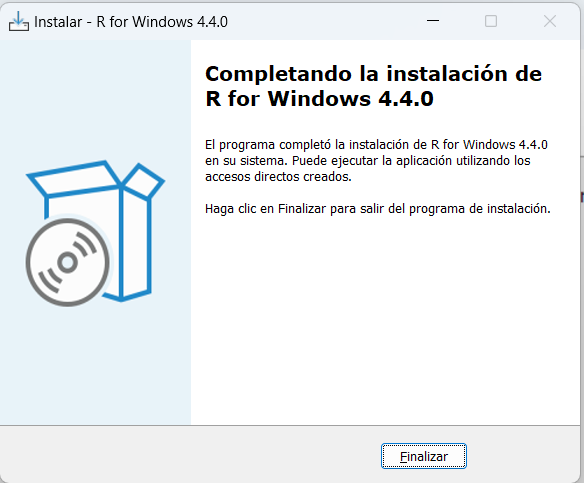
Presionas en siguiente



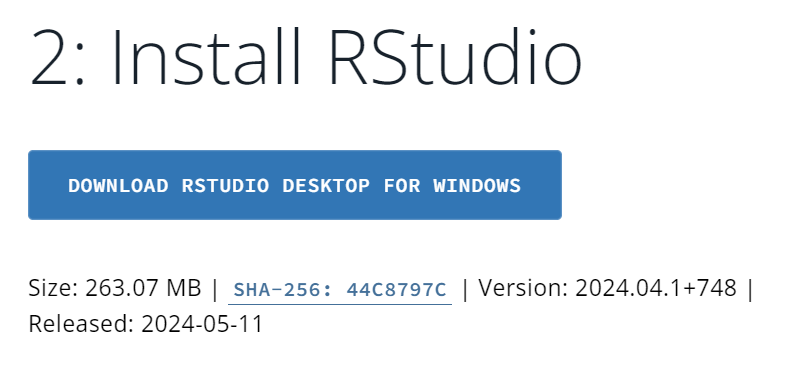
Siguiente



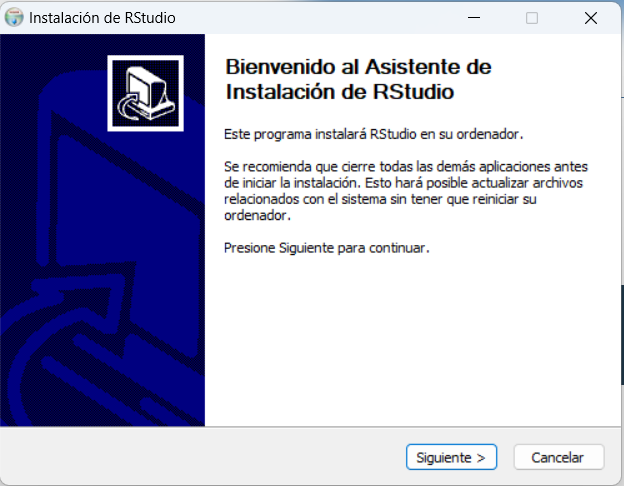
La instalación se ha completado



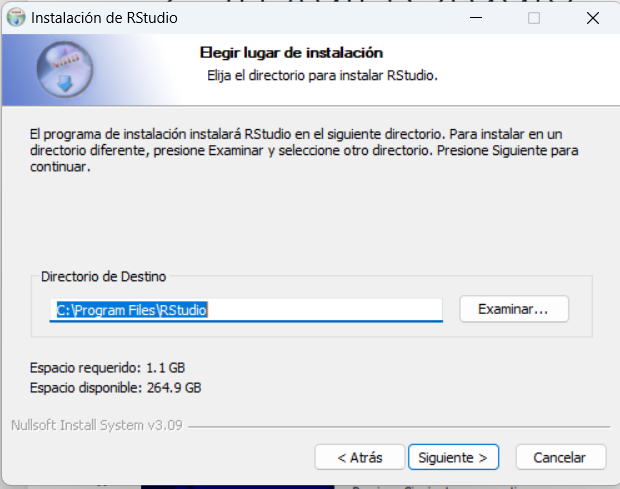
R studio



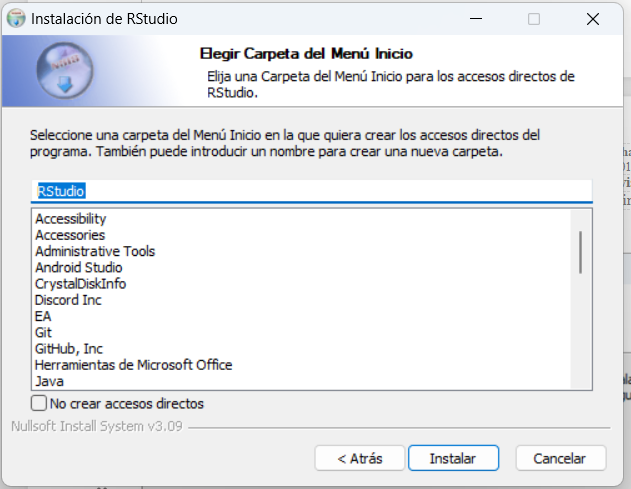
Abrimos el instalador y presionamos siguiente



De nuevo en siguiente



Presionamos instalar



|  |
| --- |
| **Guía de elaboración de práctica de ejercicios** |

|  |
| --- |
| **Práctica de Ejercicios** |
| Es el documento empleado para aplicar conceptos teóricos por medio de ejercicios. | |

**Pasos para elaborar una práctica de ejercicios.**

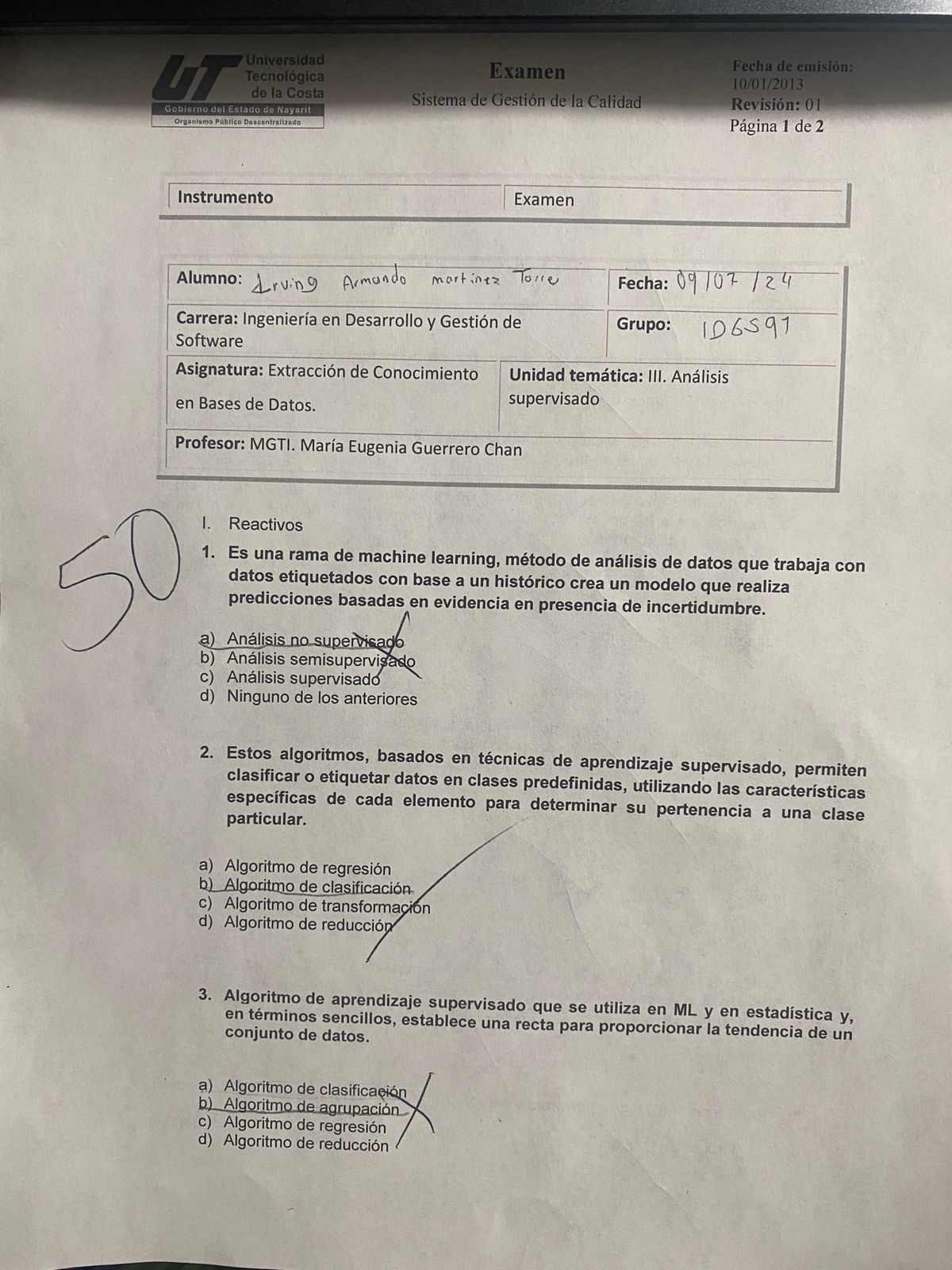
1. Leer todos los materiales y referencias bibliográficas asociadas al tema de la unidad.
2. Identificar el objetivo de la actividad requerida.
3. Leer detalladamente los ejercicios que se solicitan en la práctica.
4. Dar respuesta a los planteamientos, problemas o cuestiones que se soliciten.
5. Redactar el procedimiento que se empleó para llegar a los resultados finales.
6. Resaltar los resultados finales de cada ejercicio, de tal forma que se diferencien de los procedimientos.
7. Verificar todas las actividades realizadas, con el objetivo de evitar omisiones o faltas en la redacción.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Contenido a Evaluar**   |  |  |  | | --- | --- | --- | | **Aspectos a evaluar** | **Descripción** | **Ponderación** | | Ejercicios | En esta sección se especifican los ejercicios que han sido solicitados |  | | Procedimiento y resultados | Evidenciar los razonamientos detallados y ordenados así como las estrategias que se han empleado en el proceso de solución de los ejercicios solicitados. En esta sección se deben presentar los resultados obtenidos de cada ejercicio. | 100 | |

|  |  |
| --- | --- |
| **Consideraciones Generales**  (Serán requisitos indispensables para la entrega) | |
| Datos de identificación | Alumno, Fecha, Carrera, Grupo, Unidad temática, Asignatura, Profesor |
| Bibliografía | Reporte de bibliografía que haya sido utilizada para la elaboración de los ejercicios, de acuerdo a las normas establecidas por el APA (American Psychological Association) para citar referencias bibliográficas. |
| Coherencia y ortografía | Establecimiento de ideas claras y ordenadas en cada párrafo, sin errores ortográficos |

# Unidad 3. Análisis Supervisado

## Examen



## Practica ejercicios1\_U3\_ECBD

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

|  |  |
| --- | --- |
| **Alumno**: Irving Armando Martinez Torres | **Fecha: 20 jun 2024** |
| **Carrera:** Ingeniería en Desarrollo y Gestión de Software | **Grupo: IDGS91** |
| **Asignatura:** Extracción de Conocimiento en Bases de Datos. | **Unidad temática:** II. Preparación de los Datos |
| **Profesor:** MGTI. María Eugenia Guerrero Chan | |

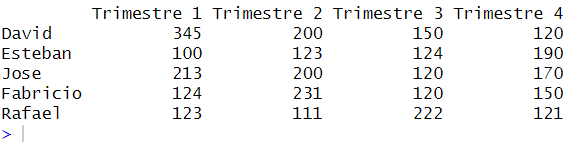
**I.- Ejercicios a resolver:**

**Instrucciones:**

Realiza lo siguiente en el programa de RStudio.

1.- Crea una matriz y titula las columnas y filas, así como se indica en el siguiente ejemplo e ingrese cantidades para las ventas:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Trimestre 1 | Trimestre 2 | Trimestre 3 | Trimestre4 |
| David |  |  |  |  |
| Esteban |  |  |  |  |
| Jose |  |  |  |  |
| Fabricio |  |  |  |  |
| Rafael |  |  |  |  |



Cálculos

1.- Obtener la venta mayor por trimestre.

2.-Obtener la venta menor por trimestre.

3.- Obtener la venta promedio por vendedor.

4.- Obtener la venta promedio por trimestre.

5.- Cantidad de ventas registradas en la tabla.

6.- Mostrar la venta mayor.

7.- Mostrar la venta menor.

Nota 1: El alumno ingresará las ventas

Nota 2: Dejar evidencia en cada uno de los pasos y el resultado final.

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

1.- Obtener la venta mayor por trimestre.

> venta\_mayor\_por\_trimestre <- apply(tabla, 2, max)

IMG_257

2.-Obtener la venta menor por trimestre.

venta\_menor\_por\_trimestre <- apply(tabla, 2, min)

IMG_258

3.- Obtener la venta promedio por vendedor.

venta\_promedio\_por\_vendedor <- apply(tabla, 1, mean)

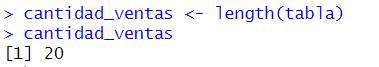
IMG_259

4.- Obtener la venta promedio por trimestre.

venta\_promedio\_por\_trimestre <- apply(tabla, 2, mean)

IMG_260

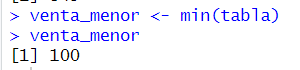
5.- Cantidad de ventas registradas en la tabla.



6.- Mostrar la venta mayor.

IMG_262

7.- Mostrar la venta menor.



## Practica ejercicios2\_U3\_ECBD

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

|  |  |
| --- | --- |
| **Alumno**: Irving Armando Martinez Torres | **Fecha: 21 jun 2024** |
| **Carrera:** Ingeniería en Desarrollo y Gestión de Software | **Grupo:** |
| **Asignatura:** Extracción de Conocimiento en Bases de Datos. | **Unidad temática:** III. Análisis supervisado |
| **Profesor:** MGTI. María Eugenia Guerrero Chan | |

**I.- Ejercicios a resolver:**

**Instrucciones:**

Realiza lo siguiente en el programa de RStudio.

1.- Crea una matriz y titula las columnas y filas, así como se indica en el siguiente ejemplo e ingrese calificaciones:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | U1 | U2 | U3 | U4 |
| Fabian |  |  |  |  |
| Ramon |  |  |  |  |
| Luis |  |  |  |  |
| Liliana |  |  |  |  |
| Diana |  |  |  |  |
| Erika |  |  |  |  |
| Karina |  |  |  |  |
| Jorge |  |  |  |  |

Cálculos

1.- Obtener la calificación mayor.

2.-Obtener la calificación menor.

3.- Obtener la calificación promedio por unidad.

4.- Obtener la calificación promedio por alumno.

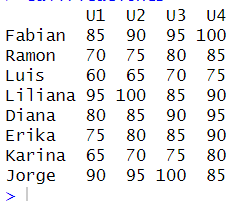
5.- Mostrar cuantas calificaciones están dadas de alta.

6.- Obtener la mediana de las calificaciones.

Nota 1: El alumno ingresará las calificaciones.

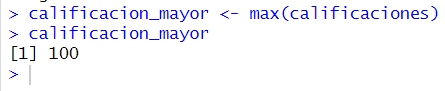
Nota 2: Dejar evidencia en cada uno de los pasos y el resultado final.

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



Cálculos

1.- Obtener la calificación mayor.



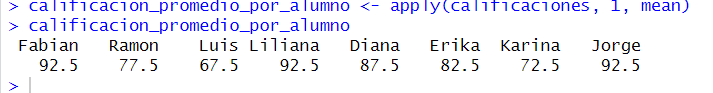
2.-Obtener la calificación menor.

IMG_258

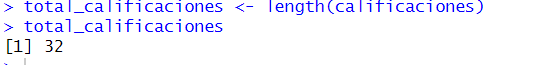
3.- Obtener la calificación promedio por unidad.

IMG_259

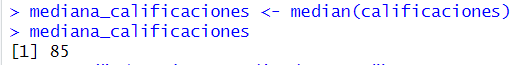
4.- Obtener la calificación promedio por alumno.



5.- Mostrar cuantas calificaciones están dadas de alta.



6.- Obtener la mediana de las calificaciones.



## Practica ejercicios3\_U3\_ECBD

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

|  |  |
| --- | --- |
| **Alumno**: Irving Armando Martinez Torres | **Fecha: 26/6/2024** |
| **Carrera: Ingenieria en gestion y desarrollo de software multiplataforma** | **Grupo: IDGS91** |
| **Asignatura:** ECBD | **Unidad temática: Introduccion al analisis de datos** |
| **Profesor:** María Eugenia Guerrero | |

I.- Ejercicios a resolver:

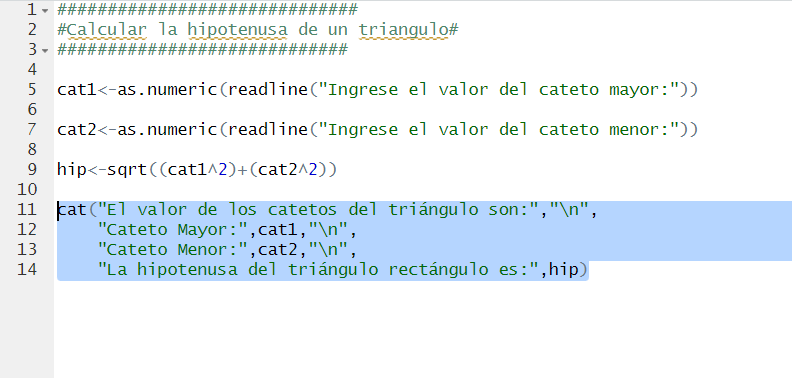
Realiza lo siguiente en el programa de RStudio.

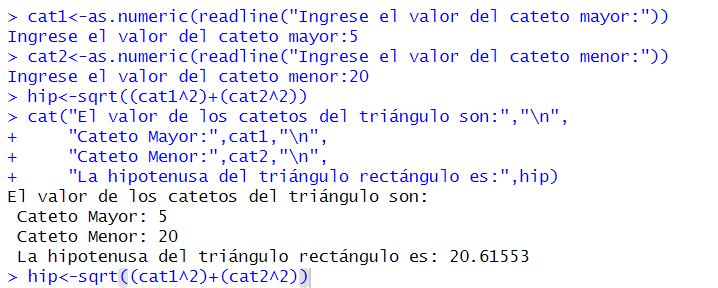
1.- Calcula el área de un triángulo.

2.- Calcula el área de un rectángulo.

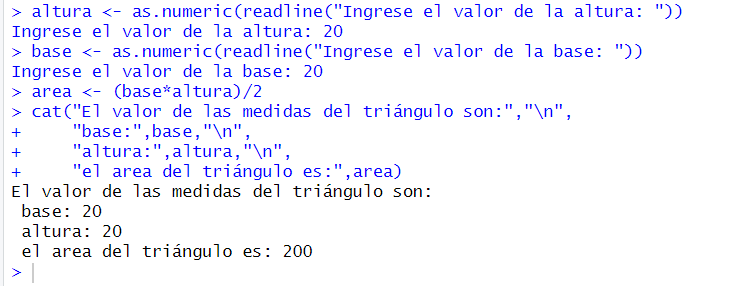
3.- Calcula el perímetro de un rectángulo.

II.-Procedimientos y resultados:



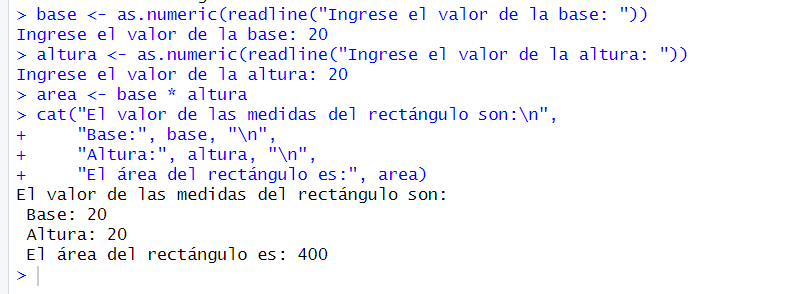


1.- Calcula el área de un triángulo



.

2.- Calcula el área de un rectángulo.



3.- Calcula el perímetro de un rectángulo.

