**Tarea 3 Algoritmos de Aprendizaje Supervisado**

**Yeison Fredy Chala**

**Curso: Análisis de datos**

**Código: 202016908**

**Grupo 50**

**Tutor**

**Handry Orozco**

**Universidad Nacional Abierta Y A Distancia-UNAD**

**Escuela De Ciencias Básicas, Tecnología E Ingeniería-ECBTI**

**Ibagué**

**2024**

**Introducción**

En el siguiente trabajo tiene como objetivo principal aplicar algoritmos de aprendizaje supervisado mediante el uso de Python y Jupyter notebooks, utilizando datasets obtenidos de la plataforma Kaggle.

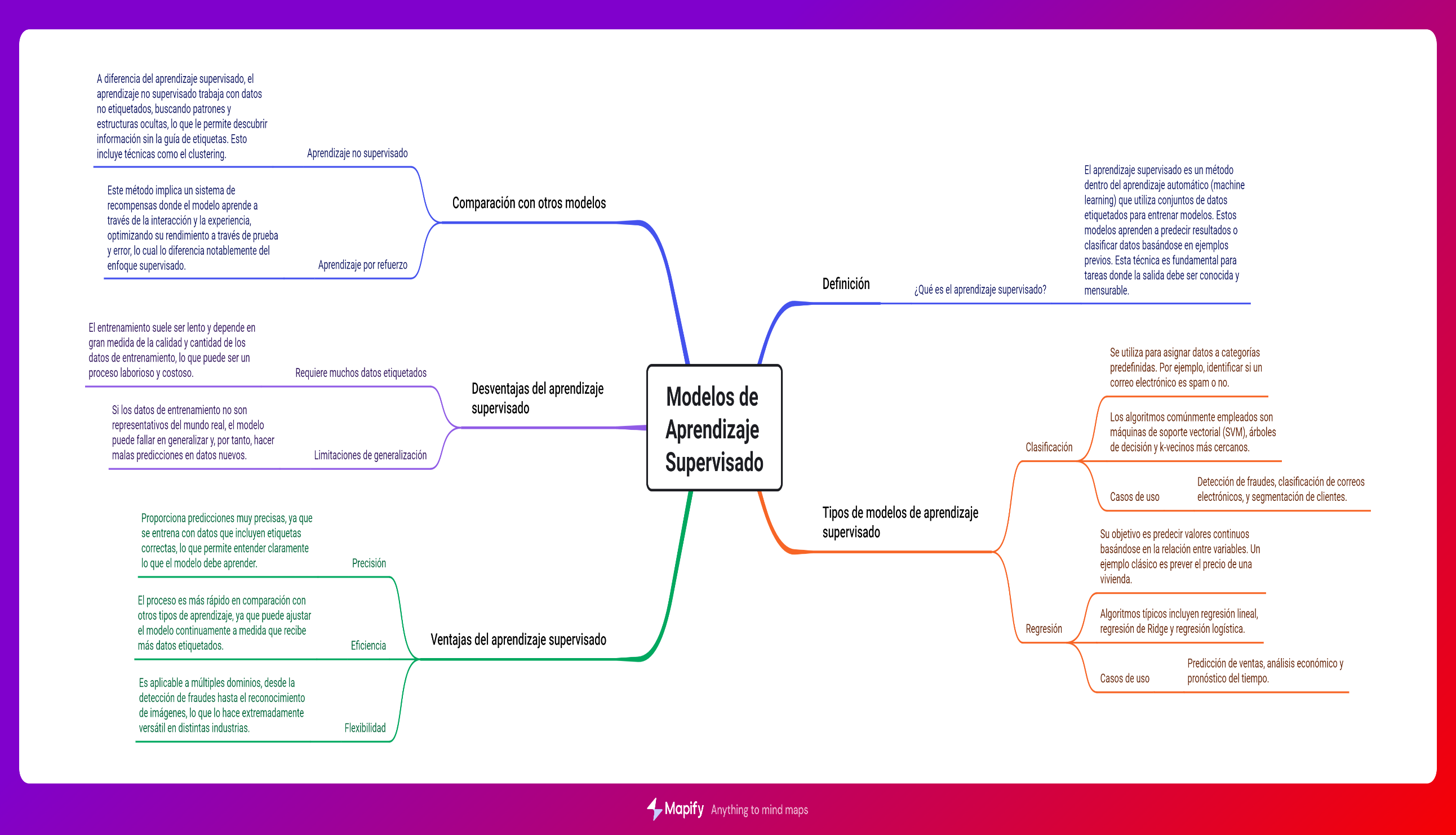
En la presente actividad, se abordarán los conceptos y técnicas esenciales de los algoritmos de aprendizaje supervisado, con el fin de consolidar el conocimiento teórico y práctico en esta área.

se elaborará un cuadro sinóptico sobre los diferentes modelos de aprendizaje supervisado, y se definirá una serie de conceptos clave relacionados con la evaluación de modelos predictivos. Estos pasos proporcionarán las herramientas necesarias para comprender y aplicar adecuadamente los modelos de regresión y clasificación

**Objetivo General**:

Aplicar algoritmos de Machine Learning supervisado según el problema, empleando métodos de modelado predictivo como regresión y clasificación

**Objetivos Específicos:**

1. Desarrollar un cuadro sinóptico que presente los principales modelos de aprendizaje supervisado, destacando su definición, casos de uso, ventajas y desventajas.
2. Implementar y ajustar modelos de regresión y clasificación utilizando Python en Jupyter notebooks, aplicando los métodos adecuados para cada tipo de problema.
3. Evaluar el rendimiento de los modelos mediante el uso de métricas clave como precisión, recall, exactitud, F1 Score y R², utilizando datasets de Kaggle.
4. Explorar y comparar los diferentes enfoques de preprocesamiento de datos, como One Hot Encoding y la división de conjuntos de datos en entrenamiento, validación y prueba, para optimizar los modelos.
5. Aplicar técnicas de ajuste de hiperparámetros, como GridSearchCV, para mejorar el desempeño de los modelos predictivos en problemas reales.
6. Desarrollar un análisis bibliográfico de las técnicas de aprendizaje supervisado más utilizadas, destacando sus aplicaciones y tendencias actuales en la ciencia de datos..
7. cuadro sinóptico sobre los diferentes modelos de Aprendizaje Supervisado

https://www.canva.com/design/DAGUK\_oL-Zg/hEOOgrIgfubAOUgOpAkuUA/edit?utm\_content=DAGUK\_oL-Zg&utm\_campaign=designshare&utm\_medium=link2&utm\_source=sharebutton

2 Elaborar un listado con las siguientes definiciones: Datos de Train, Datos de Validation y Test, GridSearchCV, One Hot Encoding, Matriz de confusión, Precision, Accuracy, Specifiticy, Recall, F1 Score, curva ROC, R cuadrado.

**Datos de Train (Entrenamiento):** Son los datos que se utilizan para entrenar un modelo de Machine Learning. El modelo "aprende" de estos datos para poder hacer predicciones sobre datos nuevos.

En un modelo para predecir el precio de casas, los datos de entrenamiento serían las características (como tamaño, ubicación, número de habitaciones) y los precios de venta de casas anteriores. El modelo usa estos datos para "aprender" cómo predecir precios futuros.

"Los datos de entrenamiento consisten en ejemplos utilizados para ajustar los parámetros (por ejemplo, pesos) del modelo. El modelo aprende patrones de estos datos para hacer predicciones sobre datos invisibles" (p. 104).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press

**Datos de Validation (Validación):** Son los datos que se usan para ajustar los hiperparámetros del modelo (parámetros que no se aprenden del entrenamiento). Ayudan a evitar que el modelo se "sobreajuste" a los datos de entrenamiento.

Ejemplo: Si estás afinando un modelo de predicción del clima, los datos de validación pueden ser usados para ajustar los parámetros que controlan la precisión de la predicción, como qué tan sensibles son los sensores a ciertos cambios de temperatura. No se usan para entrenar el modelo, sino para ajustar estos "detalles".

**Datos de Test (Prueba):** Son los datos utilizados para evaluar el rendimiento final del modelo. No se han utilizado ni en el entrenamiento ni en la validación, por lo que muestran cómo el modelo se comportará con datos nuevos.

Ejemplo: Después de entrenar un modelo para detectar fraudes en transacciones bancarias, los datos de prueba serían transacciones nuevas que el modelo no ha visto antes. Así se evalúa si el modelo realmente puede detectar fraudes en la vida real.

**GridSearchCV:** Es una técnica para encontrar los mejores hiperparámetros para un modelo. Prueba diferentes combinaciones de parámetros para ver cuál da los mejores resultados en el conjunto de validación.

**Ejemplo**: Imagina que estás diseñando un modelo para predecir enfermedades basadas en datos médicos. GridSearchCV te permite probar diferentes configuraciones de parámetros, como el número de árboles en un bosque aleatorio o la profundidad máxima de esos árboles, para encontrar la combinación que da la mejor precisión en tus predicciones

**One Hot Encoding:** Es un método para convertir variables categóricas (como colores o tipos) en una representación numérica que puede usar el modelo. Crea una columna por cada categoría, asignando un 1 a la categoría correcta y 0 a las demás.

Ejemplo: En un modelo que predice qué tipo de coche prefiere una persona, podrías tener una variable categórica con "SUV", "sedán", "deportivo". One Hot Encoding las convierte en columnas separadas donde cada columna tendrá un valor 1 si es la categoría correcta, y 0 si no lo es.

**Matriz de confusión**: Es una tabla que muestra las predicciones correctas e incorrectas del modelo. ayuda a ver cuántas veces el modelo acertó y cuántas veces falló en cada clase.

**Ejemplo**: Si un modelo de clasificación está prediciendo si un correo electrónico es spam o no, una matriz de confusión te mostrará cuántos correos fueron clasificados correctamente como spam, cuántos no spam fueron erróneamente clasificados como spam (falsos positivos), y viceversa (falsos negativos).

**Precisión (Precision):** Mide cuántos de los ejemplos que el modelo predijo como positivos realmente lo eran. Es útil cuando se quiere evitar falsos positivos.

**Ejemplo**: En un modelo que predice si alguien tiene una enfermedad rara, la precisión mide qué porcentaje de las personas a quienes el modelo predijo como enfermas realmente tienen la enfermedad. Si es baja, podrías estar diagnosticando erróneamente a personas sanas como enfermas.

**Exactitud (Accuracy):** Es el porcentaje de predicciones correctas sobre el total de predicciones. Y se encarga de medir el rendimiento general del modelo.

**Ejemplo**: En un modelo que clasifica fotos de perros y gatos, la exactitud mide el porcentaje de imágenes correctamente clasificadas. Si el 90% de las fotos fueron correctamente identificadas como perro o gato, la exactitud es del 90%.

**Especificidad (Specificity)**: se encarga de medir qué tan bien el modelo identifica los ejemplos negativos. Es útil cuando es importante detectar todos los casos negativos.

**Ejemplo**: En una prueba médica para detectar cáncer, la especificidad mide cuántas personas sanas fueron correctamente identificadas como sanas (evitando falsos positivos). Es importante cuando el costo de diagnosticar falsamente a una persona como enferma es alto

**Sensibilidad (Recall):** se encarga de medir cuántos de los ejemplos positivos reales el modelo detectó correctamente. Es útil cuando queremos minimizar los falsos negativos.

**F1 Score:** Es una medida que combina precisión y sensibilidad en un solo valor. Es útil cuando se necesita un balance entre ambas.

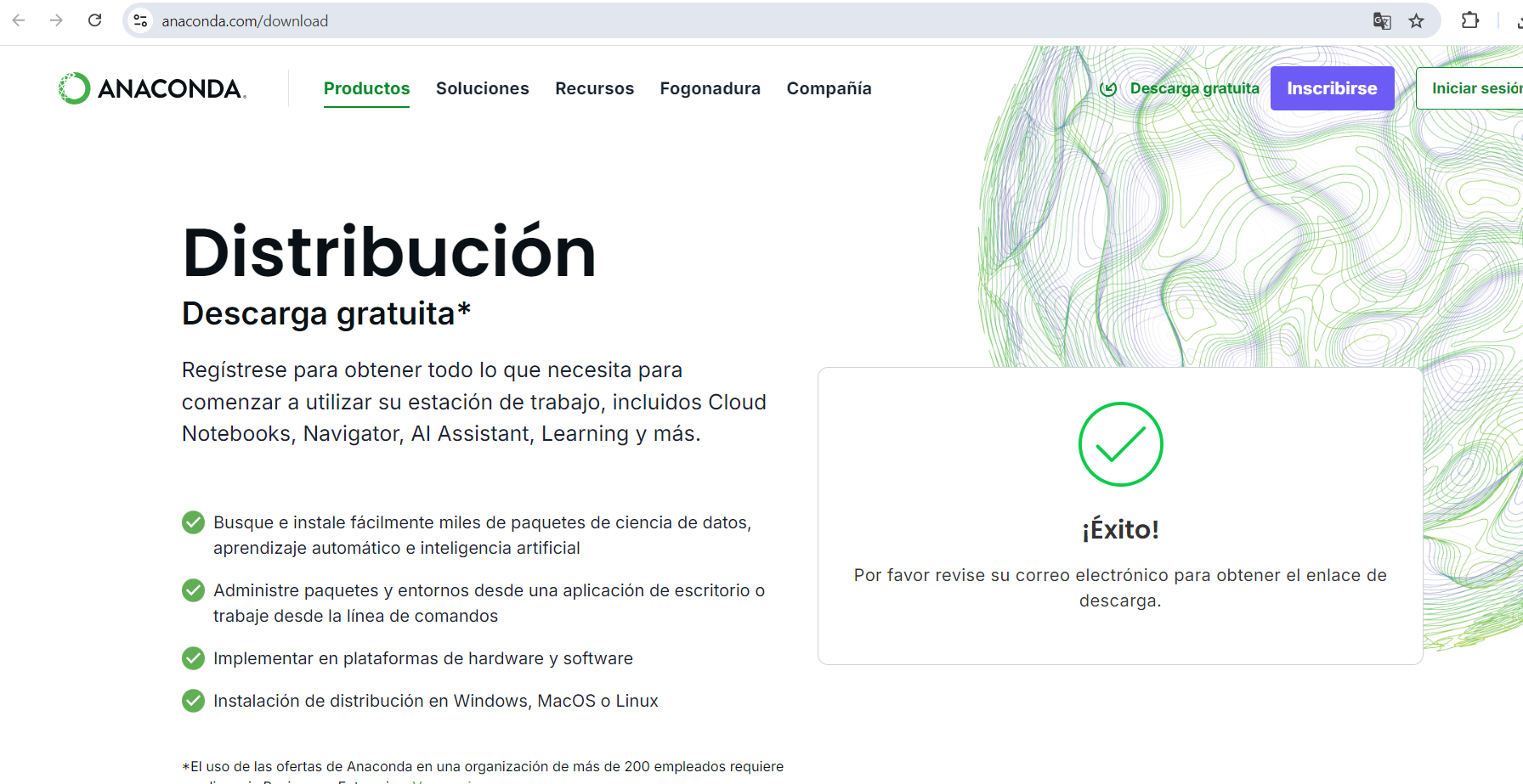
**Curva ROC:** Es un gráfico que muestra cómo varían la sensibilidad y la especificidad a medida que cambias el umbral de decisión del modelo. Ayuda a elegir el mejor umbral para clasificar los datos.

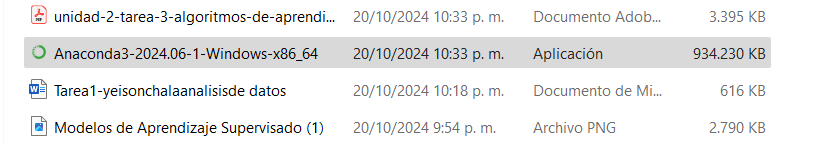
**Ejemplo**: Imagina un modelo que predice si alguien va a pagar un préstamo. La curva ROC te ayuda a ajustar el umbral que determina si alguien debe ser aprobado o no. Un umbral más bajo podría aprobar más gente, pero también aumenta el riesgo de aprobar personas que no pagarán

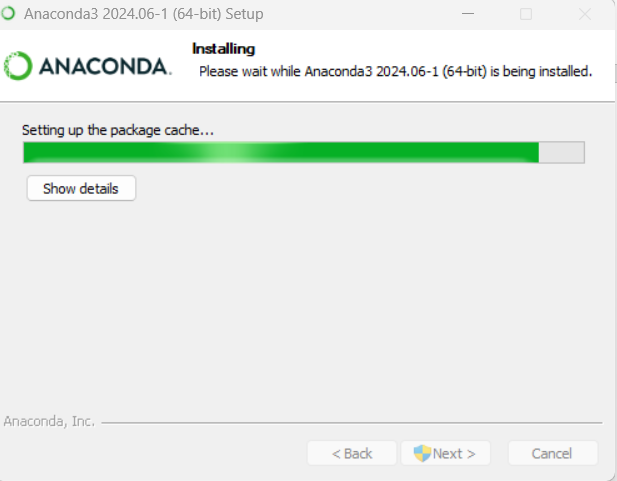
**R cuadrado (R²)**: Es una métrica que indica qué tan bien el modelo de regresión se ajusta a los datos. Un valor de 1 significa un ajuste perfecto, mientras que 0 significa que el modelo no explica nada de la variabilidad de los datos.

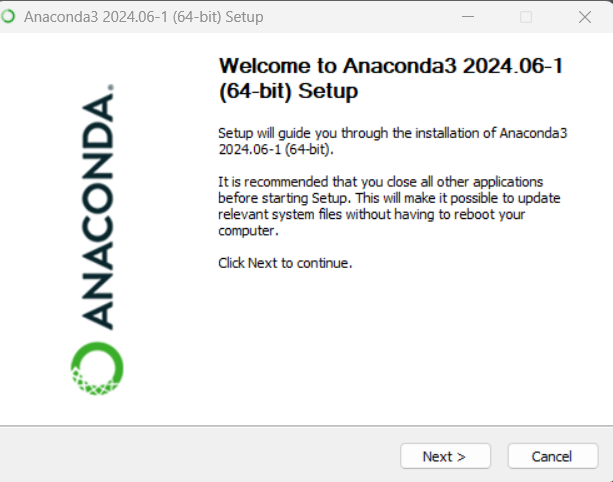
**Ejemplo**: Si estás creando un modelo para predecir las ventas mensuales de una tienda en función del clima, la R² te indica qué tan bien tu modelo está capturando las tendencias reales en las ventas. Un R² de 0.85 significa que el modelo explica el 85% de la variación en las ventas, lo que indica un buen ajuste.

3 El lenguaje a utilizar es Python, el cual se trabajará mediante Jupyter notebooks, para esto es necesario instalar Anaconda, que es una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático. 2 Descargue e instale Anaconda desde: https://www.anaconda.com/products/distribution



****



**Conclusiones**

el desarrollo de estas actividades nos ha proporcionado una base sólida en el tema de análisis de datos, desde la comprensión de conceptos teóricos, hasta la aplicación práctica de técnicas estadísticas. Esta orientación integral es crucial para desarrollar habilidades analíticas firmes y aplicables en una variedad de argumentos de investigación y práctica profesional.

La revisión bibliográfica y el análisis de los conceptos fundamentales del análisis de datos nos permitirá una comprensión más clara y organizada de los elementos clave en esta disciplina. Conceptos como población, muestra, y diversos tipos de muestreo proporcionan el marco necesario para la recolección y análisis de datos, mientras que medidas estadísticas como media, mediana, moda, y desviación estándar son cruciales para resumir y interpretar la información

Con la elaboración del cuadro sinóptico se ha facilitado la visualización de las principales características y relaciones entre Inteligencia Artificial, Machine Learning y Deep Learning. también podemos destacar la importancia del EDA en el proceso de análisis de datos, como una herramienta esencial para la exploración inicial y la comprensión de patrones y incoherencias en los datos.

**Referencias bibliográficas**

Agresti, A., & Franklin, C. (2013). Statistics: The art and science of learning from data (3rd ed.). Pearson.

Bluman, A. G. (2017). Elementary statistics: A step by step approach (10th ed.). McGraw-Hill Education.

Fernández, J. (2020). Métodos estadísticos: Prueba Chi-cuadrado. Revista de Estadística Aplicada, 15(2), 45-58. https://www.revistadeestadisticaaplicada.com/metodos-chicuadrado

Cochran, W. G. (1977). Sampling techniques (3rd ed.). Wiley.

Keller, G. (2018). Statistics for management and economics (11th ed.). Cengage Learning.

Levine, D. M., Stephan, D. F., & Szabat, K. P. (2017). Statistics for managers using Microsoft Excel (8th ed.). Pearson.

Mertler, C. A., & Vannatta, R. A. (2016). Advanced and multivariate statistical methods: Practical application and interpretation (5th ed.). Pyrczak Publishing.

McClave, J. T., Sincich, T., & Dietrich, F. (2019). Statistics for business and economics (13th ed.). Pearson.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). Introduction to the practice of statistics (9th ed.). W.H. Freeman.

Montgomery, D. C. (2017). Design and analysis of experiments (9th ed.). Wiley.

Morales, J. (2020, junio 15). Distribución normal y sus aplicaciones en estadística. Revista de Estadística y Matemáticas Aplicadas. https://www.revistaestadisticaymatematicas.com/distribucion-normal

Triola, M. F. (2018). Elementary statistics (13th ed.). Pearson.