Imagen que contiene dibujo

Descripción generada automáticamente

**Unidad Temática 1 – Práctico Domiciliario 1**

**Asignatura**: Inteligencia Artificial 1

**Alumno:**

Juan M. Pérez

Montevideo, 15 de agosto de 2023

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Definición 1: “El aprendizaje automático es una rama de  [la inteligencia artificial (IA)](https://www.ibm.com/topics/artificial-intelligence)  y la informática que se centra en el uso de datos y algoritmos para imitar la forma en que los humanos aprenden, mejorando gradualmente su precisión.”

Recuperado de IBM <https://www.ibm.com/topics/machine-learning>, 15 de agosto 2023.

Definición 2: “El aprendizaje automático es un subcampo de la inteligencia artificial, que se define ampliamente como la capacidad de una máquina para imitar el comportamiento humano inteligente. Los sistemas de inteligencia artificial se utilizan para realizar tareas complejas de una manera similar a la forma en que los humanos resuelven problemas.”

Recuperado de Sloan de MIT <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>, 16 de agosto de 2023.

Definición 3: “El aprendizaje automático permite que una máquina aprenda automáticamente de los datos, mejore el rendimiento de las experiencias y prediga cosas sin estar programada explícitamente.”

Recuperado de Java T point <https://www.javatpoint.com/machine-learning>, 16 de agosto de 2023.

* Como se compara el Aprendizaje Automático de la “Inteligencia Artificial”, básicamente la Inteligencia artificial es dotar a la maquina con la capacidad de pensar de forma similar a como lo realizan los seres humanos, pero a su vez la fuerte relación que tiene con la ML es que esta última es parte de la IA.

A su vez ambas emplean la automatización de tareas y procesos que normalmente requerirían la intervención humana.

La diferencia más elocuente es que la IA es como se mencionaba anteriormente dotar a la maquina con la capacidad de pensamiento, mientras que la ML esta centrado en el desarrollo de algoritmos y técnicas que permiten a las maquinas aprender patrones de datos sin ser estos programados implícitamente.

* Tanto el Machine Learning como el Estudio Estadístico, son formas de obtener información a través de datos, esta es la principal relación que hay entre ambos.

Como se mencionaba anteriormente ambos utilizan datos, los cuales son requeridos como entradas para poder ser procesados y generar resultados, para esto los datos deben ser coherentes y con la mínima cantidad de errores posible (el tener datos acertados favorece al resultado final).

Como diferencias entre ambos tenemos que, el aprendizaje automático se enfoca en la capacidad de las máquinas para aprender patrones y tomar decisiones sin intervención humana directa. El Análisis Estadístico, por otro lado, se centra en inferir propiedades y relaciones dentro de los datos mediante métodos estadísticos.

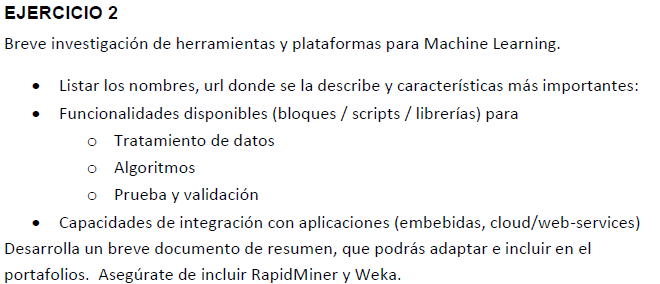
Otra diferencia es que el Machine Learning utiliza algoritmos y modelos matemáticos para entrenar máquinas y generar predicciones basadas en patrones pasados. El Análisis Estadístico se basa en técnicas estadísticas para comprender las propiedades y relaciones de los datos.

* Si bien comúnmente se tiene el concepto de que ML es igual a data mining, ese concepto esta un poco errado, porque si bien su punto objetivo es similar la forma en que cada uno lo realiza es distinto. El Data Mining se enfoca en encontrar patrones interesantes y significativos que puedan ayudar a tomar decisiones informadas, mientras que el Machine Learning utiliza esos patrones para hacer predicciones futuras y tomar acciones automatizadas.

Otra cosa en la que se establece una diferencia es que si bien ambos enfoques dependen de técnicas estadísticas para analizar datos. Mientras que el Data Mining se basa en métodos estadísticos para identificar patrones y relaciones, el Machine Learning utiliza algoritmos estadísticos para entrenar modelos que puedan realizar tareas específicas, como clasificación o regresión.

* El machine learning tiene diferentes aplicaciones según su rama de implementación, actualmente comprende un sinfín de aplicaciones a tareas cotidianas e incluso en tareas con una mayor complejidad. Dentro de estas ramas tenemos la Salud, la Energía, el Procesamiento de Lenguaje Natural (PLN), entre otras tantas.

En cada una de estas se utilizan diferentes enfoques metodológicos de procesamiento que son acordes a las diferentes situaciones, problemas y datos. Al ser esta una herramienta que nos ayuda a facilitar muchas tareas y debe realizarlo con la mayor optimización existen diferentes algoritmos que lo hacen posible.



**Plataformas y herramientas**

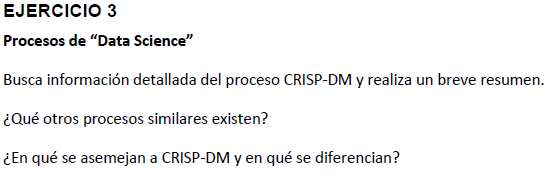
Las plataformas y herramientas de machine learning hoy en día dan soporte a millones de personas en el mundo, la utilización va en aumento por la automatización de procesamiento de datos que poseen.

Actualmente su aplicación se encuentra en diversas ramas, como lo pueden ser desde la inteligencia empresarial, medicina, entre otras hasta las soluciones de aplicaciones más comunes como son la creación de Dashboards.

Existen plataformas y herramientas que son accesibles para todas las personas, ya que las mismas son de código abierto y cuentan con interfaces gráficas que favorecen a la Experiencia del Usuario, quien puede crear contenido simplemente siguiendo una serie de sencillos pasos.

A continuación, se expone un cuadro en el que se reflejan algunas de las características principales de las plataformas y herramientas más utilizadas en la actualidad, sin embargo, estas son solo una pequeña muestra del sinfín que existen.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tipo de Herramienta/**  **Plataforma** | **Descripción** | **Características Importantes** | **Funcionalidades Disponibles** | **Integración en Aplicaciones** |
| Bibliotecas de Python (por ejemplo, scikit-learn) <https://scikit-learn.org/stable/> | Conjunto de herramientas y algoritmos de Machine Learning en Python. | - Fácil de usar y aprender.  - Amplia variedad de algoritmos. - Comunidad activa. | - Algoritmos de clasificación, regresión, agrupación, etc.  - Preprocesamiento de datos.  - Evaluación de modelos. | Embebible en aplicaciones Python o consumible desde servicios web. |
| TensorFlow  [https://www.tensorflow.org/ ?hl=es-419](https://www.tensorflow.org/?hl=es-419) | Plataforma de código abierto de aprendizaje automático desarrollada por Google. | - Redes neuronales y modelos profundos.  - Distribución y entrenamiento en GPUs.  - TensorFlow Lite para dispositivos móviles. | - Creación de redes neuronales.  - Implementación de modelos de aprendizaje profundo.  - Distribución de cómputo. | Integrable en aplicaciones web y móviles, o consumible desde servicios web. |
| PyTorch  <https://pytorch.org/> | Biblioteca de aprendizaje automático de código abierto desarrollada por Facebook. | - Enfoque dinámico y flexible.  - Popular en la comunidad de investigación.  - Compatibilidad con GPUs. | - Construcción de modelos de aprendizaje profundo. - Autograd para diferenciación automática. - Implementación eficiente de redes neuronales. | Integrable en aplicaciones científicas y de investigación, o consumible desde servicios web |
| Microsoft Azure ML  [https://azure.microsoft.com /en-us/free/ machine-learning/](https://azure.microsoft.com/en-us/free/machine-learning/) | Plataforma en la nube para el desarrollo, entrenamiento e implementación de modelos de Machine Learning. | - Entorno basado en la nube.  - Herramientas de colaboración.  - Soporte para diferentes lenguajes. | - Creación y entrenamiento de modelos.  - Integración con herramientas de flujo de trabajo.  - Despliegue en la nube. | Integrable en aplicaciones empresariales y en la nube, o consumible desde servicios web. |
| IBM Watson  <https://www.ibm.com/es-es/cloud/watson-studio> | Plataforma de inteligencia artificial de IBM. | - Amplia gama de servicios de IA.  - Herramientas de automatización.  - Enfoque en el procesamiento del lenguaje natural. | - Creación de chatbots.  - Análisis de sentimiento.  - Reconocimiento de voz y texto. | Integrable en aplicaciones empresariales y servicios web, o consumible desde servicios cloud. |
| RapidMiner  <https://rapidminer.com/> | Plataforma de análisis de datos que incluye capacidades de aprendizaje automático. | - Flujo de trabajo visual.  - Amplia gama de algoritmos.  - Enfoque en análisis predictivo. | - Construcción de flujos de trabajo de análisis.  - Preprocesamiento de datos.  - Modelado predictivo. | Integrable en aplicaciones analíticas y empresariales, o consumible desde servicios web. |
| Weka  <https://www.weka.io/> | Software de minería de datos y aprendizaje automático de código abierto. | - Interfaz gráfica amigable.  - Diversidad de algoritmos.  - Uso educativo y de investigación. | - Preprocesamiento de datos.  - Evaluación de modelos.  - Experimentación con algoritmos. | Ampliamente utilizado en educación y aplicaciones de investigación, o consumible desde servicios web. |

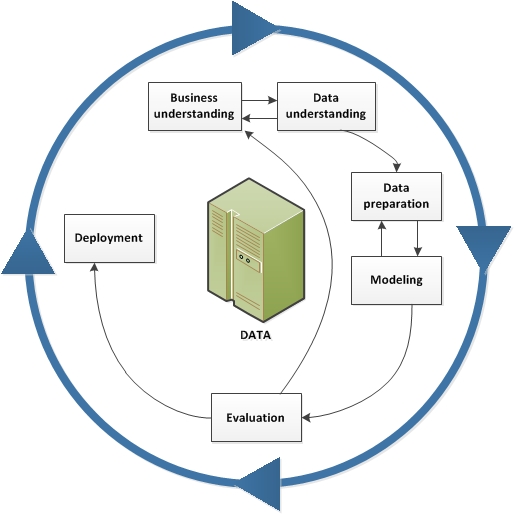


**Concepto:**

Para comprender el concepto de CRISP-DM es necesario conocer que significa la sigla, Cross-Industry Standar Process for Data Mining, lo cual nos induce a pensar que esta basado en un proceso estándar cuyo cometido fundamental es la minería de datos.

Se puede dividir conceptualmente en 2 grandes ramas, una como metodología que como enfoque, abarca las etapas estándar de un proyecto, detallando las actividades requeridas en cada etapa y proporcionando una clara explicación de las interconexiones entre estas actividades.

Y la siguiente trata sobre el modelo de proceso propiamente dicho, ofreciendo un resumen del ciclo de vida de la minería de los datos.



Este modelo de ciclo de vida, se compone de seis etapas interconectadas mediante flechas que destacan las dependencias clave y frecuentes entre ellas. Aunque existe una secuencia sugerida para estas etapas, es importante tener en cuenta que muchos proyectos fluidamente transitan hacia adelante y hacia atrás entre estas fases según las necesidades del proyecto.

Recuperado de IBM <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview> y Canal de YouTube del Instituto Data Science Argentina <https://www.youtube.com/watch?v=UyKkSsEbXkw> 16 de agosto de 2023.

**Metodología similar a CRISP-MD**

La metodología SEMMA, cuya terminología corresponde al acrónimo en ingles de las 5 fases básicas del proceso.



“Este tipo de metodología se define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.” Recuperado de AEIPRO (Asociación Española de Dirección e Ingeniería de Proyectos) <https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf>

En esta metodología, el proceso comienza al tomar una muestra de la población para el análisis, con el fin de elegir una muestra que refleje adecuadamente el problema en estudio. Es crucial que esta muestra sea representativa, ya que su falta de representatividad afectaría la validez del modelo y la aceptabilidad de los resultados. La forma más típica de obtener la muestra es mediante la selección al azar, donde cada miembro de la población tiene una igual probabilidad de ser seleccionado. Esto se conoce como muestreo aleatorio simple.

**Semejanzas y diferencias entre ambas Metodologías**

Tanto en SEMMA como en CRISP-DM, el objetivo final es el mismo poder obtener patrones desconocidos a través del procesamiento de grandes cúmulos de información.

Ambos emplean diferentes fases para llegar a este cometido, aunque como vimos anteriormente SEMMA cuenta con 5 y CRISP-DM tiene 6 fases entre las que divide el procesamiento de los datos.

Continuando con las similitudes en ambos casos se cuentan con procesos que son iterativos e interactivos, lo que quiere decir que los pasos o fases de cada uno están directamente relacionados con los demás.

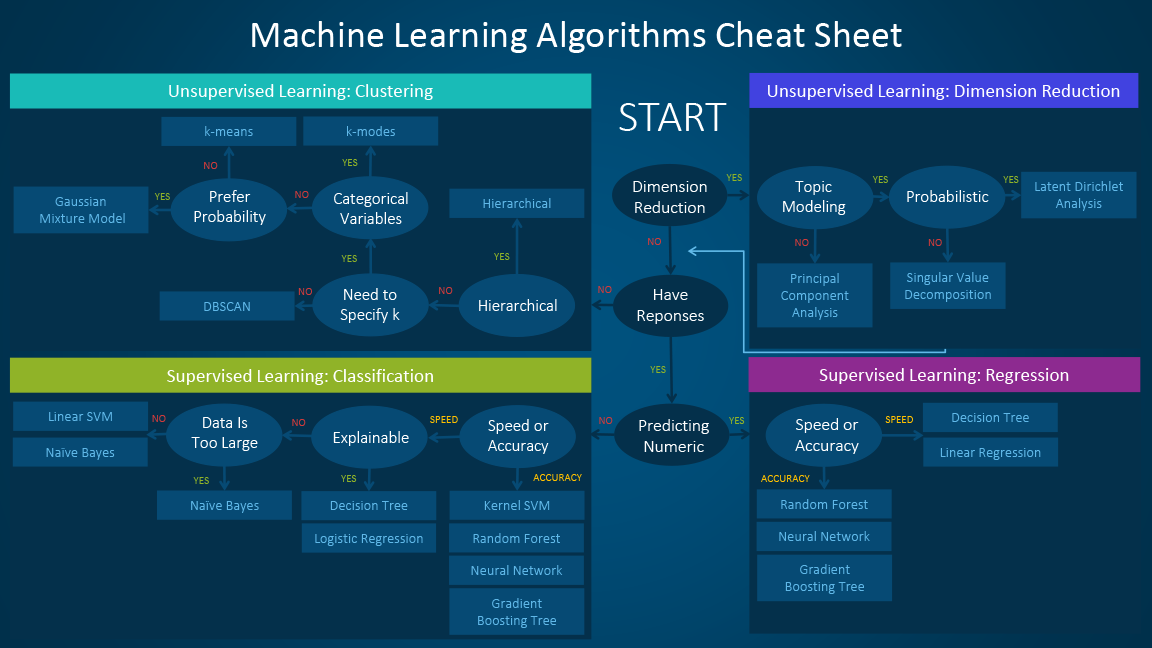
Una de las mayores diferencias se presenta en cuanto a la implementación, siendo SEMMA sin una etapa especifica para la implementación de los resultados en el negocio, mientras que CRISP-DM se dedica a incorporar los resultados en los procesos operativos y en la toma de decisiones.

Diagrama

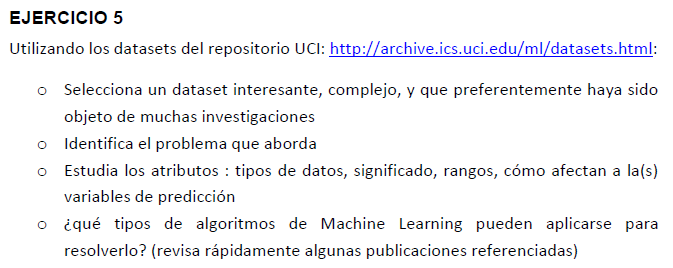
Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza baja



|  |  |  |
| --- | --- | --- |
|  | RAPIDMINER | IBM WATSON |
| Tipos de Algoritmos | -Clustering (K-means, DBSCAN, etc)  -Regresión (Linear, Polinomial, etc)  -Clasificación (Decision Trees, Naive Bayes, etc) | -Clustering (K-means, Hierarchical, etc)  -Regresión (Linear Regression, etc)  -Clasificación (Random Forest, SVM, etc) |
| Atributos Soportados | -Datos numéricos  -Datos categóricos  -Datos Textuales  -Imágenes  -Series temporales | -Datos numéricos  -Datos categóricos  -Datos textuales  -imágenes  -Texto Plano para NLP |



El dataset escogido fue el de Heart Disease, se adjunta imagen de presentación.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Este dataset trata de encontrar un algoritmo de probabilidad para el diagnostico de la enfermedad arterial coronaria.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| N° | Atributo | Tipo dato | Descripción | Rango |
| 1 | age | Entero | edad en años |  |
| 2 | sex | categórico |  | 1 = masculino  0 = femenino |
| 3 | cp | categórico | tipo de dolor de pecho | 1 = angina típica 2= angina atípica 3= sin dolor de angina 4= asintomático |
| 4 | trestbps | Entero | presión de sangre en reposo |  |
| 5 | chol | Entero | colesterol graso en mg/dl |  |
| 6 | fbs | categórico | azúcar en sangre rápida > 120 mg/dl | 1 = true  0 = false |
| 7 | restecg | categórico | resultado de electrocardiograma en reposo | 0 = normal 1= tener anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV) 2= muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes |
| 8 | thalach | Entero | frecuencia cardiaca máxima alcanzada |  |
| 9 | exang | categórico | ejercicio indujo angina | 1= yes 2= no |
| 10 | oldpeak | Entero | Depresión del ST inducida por el ejercicio en relación con el reposo |  |
| 11 | slope | categórico | la pendiente del segmento ST de ejercicio máximo | 1 = pendiente ascendente 2= plano 3= pendiente descendente |
| 12 | ca | Entero | número de vasos principales (0-3) coloreados por fluoroscopia | 0, 1, 2, 3 |
| 13 | thal | categórico | resultado de la prueba de estrés con Thalio | 3= normal 6= defecto fijo 7= defecto reversible |
| 14 | num | Entero | diagnóstico de enfermedad cardíaca | 0= < 50% estrechamiento del diámetro 1= > 50% estrechamiento del diámetro |

Propiamente desde la pagina de UCI, aparece la referencia de que algoritmos utilizar para este caso de estudio, siendo los 2 más óptimos el XGboost y la Regreción Logística, quedando en tercer lugar el Random Forest.