

“Sistema clasificador de datos n–dimensional implementado por 5 algoritmos a través de Inteligencia Artificial”

Trabajo terminal No.

Alumnos: Castillo Flores Magali, *Torres Uruga Saul

Director: Tonáhtiu Arturo Ramírez Romero

*e-mail: tt.protocolocfntus@gmail.com

Resumen – Esta propuesta de trabajo terminal busca el desarrollo de un sistema clasificador de datos que implemente algoritmos por medio de técnicas de Inteligencia Artificial (IA), para la asignación de elementos de n dimensión de datos entrantes no etiquetados en una categoría concreta conocida, con el fin de obtener clasificación supervisada a través de los paradigmas desarrollados por la estadística.

Palabras clave – Clasificador, Estadística, Inteligencia Artificial, Estructura de Datos, Análisis de algoritmos.

1. Introducción

Un sistema de clasificación de datos nos permite identificar mediante un algoritmo una asignación de un elemento no etiquetado a una categoría concreta conocidas, entre proceso es parte de la terminología de aprendizaje automático, se le considera a la clasificación como aprendizaje supervisado que contiene múltiples datos correctamente identificados.

El sistema de clasificación de datos tiene dos etapas:

- **Aprendizaje:** La construcción del modelo. El modelo se construye a partir de un conjunto de elementos ya etiquetados. El modelo obtenido se representa como un conjunto de reglas de clasificación, árboles de decisión, fórmula matemática, etc. [1].
- **Validación:** La estimación de la precisión del modelo. Se prueba el modelo obtenido con un conjunto de ejemplos diferentes al utilizado para la construcción del modelo, para cada ejemplo se compara su clase real con la clase predicha por el clasificador [1].

Un sistema clasificador de datos tiene distintos fines, ya que al ordenar información es útil en varios campos, por mencionar algunos, en la minería de datos se puede asignar valores discretos capaces de indicar si un correo es spam o no, o bien el descubrimiento de conocimiento en bases de datos de investigación clínica ya que permite conocer si un tumor es benigno o no, así como clasificar flores según la forma en que los datos de entrada se agrupan entre otras aplicaciones. [2]

Se considera estudiar los siguientes métodos de clasificación y algoritmos más utilizados en el aprendizaje supervisado:

Clasificador Bayesiano	Simple y rápido.	Saber cuál es la hipótesis más probable entre varios conjuntos de datos.	Algoritmo.
Validación cruzada	Muy rápido.	Evalúa los resultados de un análisis estadístico.	Método.
Árboles de clasificación	Depende del crecimiento.	Clasificar utilizando particiones sucesivas.	Conjunto de algoritmos.

Máquina soporte vector (SVM)	Muy lento.	Predice a que categoría pertenece un elemento no etiquetado.	Conjunto de algoritmos.
K vecinos más cercanos (KNN)	Muy lento.	Estima la función de densidad de los elementos no etiquetados por cada clase.	Algoritmo.
Redes Neuronales	Lento.	Aprender mediante ensayos repetidos para conseguir maximizar la predicción.	Algoritmo.
K - Medias	Simple y muy rápido.	Clasifica un conjunto de objetos en un determinado número K de clústeres, K determinado a priori.	Algoritmo.

El desarrollo de este trabajo terminal busca implementar un clasificador de datos, que pueda clasificar cualquier conjunto de datos en texto una vez que se encuentren de manera ordenada. También pretende realizar una investigación sobre algunos de los diversos algoritmos para clasificar datos, de esta manera estudiar cuáles serían los algoritmos más eficientes que se podrán utilizar para la implantación de nuestro clasificador.

La motivación del presente trabajo terminal es continuar con el estudio de los algoritmos para clasificar datos, ya que previamente se ha realizado el trabajo terminal “Sistema clasificador de datos n-dimensional por medio de Inteligencia Artificial”[3] que estudió ya algunos de estos algoritmos, de esta manera como se ha mencionado antes se pretende continuar con la investigación de otros algoritmos que nos permitan implementar un clasificador con nuevos algoritmos y mejorar la investigación que anteriormente se ha realizado, así como contrastar con la tesis “Clasificación de grandes conjuntos de datos vía Máquinas de Vectores Soporte y aplicaciones en sistemas biológicos”[4] que aun siendo una aplicación particular resulta de interés el estudio y los avances obtenidos a año 2021, ya que se busca el implemento de los algoritmos realizados en ambos trabajos.

Resultados de las investigaciones en relación a la clasificación de datos obtenidos por los trabajos anteriormente mencionados.

Sistema clasificador de datos n-dimensional por medio de Inteligencia Artificial	Clasificación de grandes conjuntos de datos vía Máquinas de Vectores Soporte y aplicaciones en sistemas biológicos
Algoritmos de interés: <ul style="list-style-type: none"> • Clasificador Bayesiano • K vecinos más cercanos (KNN) • Redes Neuronales 	Algoritmo de interés: <ul style="list-style-type: none"> • Máquina soporte vector (SVM)
Se estudio 10 bases de datos con diferentes combinaciones de datos y frecuencias entre clases, donde se comprobó los siguientes aspectos: <ul style="list-style-type: none"> • Los conjuntos de datos con menos de 10% del total de datos para cada clase no pueden ser clasificados correctamente • Los conjuntos de datos con menos de 80 datos no pueden ser clasificados correctamente por la red neuronal debido 	El algoritmo FCM-SVM2, fue diseñado para trabajar en conjuntos de datos de mediano tamaño, obteniendo los siguientes resultados: <ul style="list-style-type: none"> • La obtención de un clasificador con SVM basado Fuzzy C-Means teniendo como objetivo mejorar el tiempo de entrenamiento de las SVM clásicas • El desempeño del algoritmo es muy bueno para conjuntos de datos grandes con

a que no hay suficientes datos para su entrenamiento <ul style="list-style-type: none"> El clasificador de Bayes y el clasificador KNN trabajan con una efectividad de 60% aun cuando el conjunto de datos es pequeño (menos de 80 datos, más de 15 datos) 	dimensiones menores a 20 <ul style="list-style-type: none"> Cuando el tamaño de la dimensión del conjunto de entrada es mayor a 20 el algoritmo presenta algunos problemas para un correcto agrupamiento del conjunto de datos de entrada
---	--

2. Objetivo

Objetivo General:

- Desarrollar un sistema de software que permita la clasificación de datos a través de algoritmos de aprendizaje supervisado por medio de Inteligencia Artificial.

Objetivos Específicos:

- Implementar un sistema clasificador capaz de etiquetar un conjunto de datos una vez que estos se encuentren ordenados.
- Implementar algoritmos de aprendizaje supervisado.
- Elaborar un reporte técnico.

3. Justificación

La tendencia del manejo de datos ha aumentado a pasos agigantados, la cantidad de datos digitales creados o replicados a nivel mundial se ha multiplicado por más de treinta en la última década, pasando de dos zetabytes en 2010 a 64 zetabytes el año pasado. Pero esta cantidad no es nada en comparación con lo que se espera en los próximos años. Según las previsiones, el volumen de datos generados en todo el mundo superará los 180 zetabytes en 2025, lo que supone un crecimiento medio anual de casi el 40% en cinco años [5], gracias al impacto del Internet de las Cosas, el desarrollo de la 5G y las redes sociales.

La abrumante cantidad de datos ha sobrepasado las necesidades, dado que no solo nos importan los datos, si no el uso que podemos dar a estos elementos causa de la necesidad de clasificar mediante métodos de aprendizaje que permitan resolver problemas. Habitualmente se aplica clasificadores de datos para aplicaciones específicas a su finalidad.

Por esta razón en este proyecto se planea demostrar los conocimientos obtenidos durante el proceso educativo que se ha obtenido durante la carrera de Ingeniería en Sistemas Computacionales en el Instituto Politécnico Nacional, implementando un sistema de clasificación de datos utilizando inteligencia artificial (IA) ya que es una disciplina eminentemente tecnológica que persigue la construcción de máquinas y programas capaces de realizar complejas tareas con una habilidad y eficiencia [6]. De esta manera se podrá obtener un clasificador más preciso dependiendo de los datos de entrada.

Por otro lado, este trabajo terminal pretende complementar una investigación que previamente se había realizado en el TT “Sistema clasificador de datos n-dimensional por medio de Inteligencia Artificial” en 2019, por medio del estudio de otros algoritmos y métodos de estadística e inteligencia artificial, a su vez profundizando en la investigación que se ha realizado previamente. Estudiamos este proyecto debido a que solamente se enfoca en un clasificador de datos, creemos que ayudará a mejorar la clasificación obteniendo más aciertos de esta manera será un sistema con mayor fiabilidad.

4. Producto o resultado esperado

Se espera obtener un software que, dado los elementos no etiquetados, sean agrupados y separados a categorías concretas conocidas, como se muestra en la siguiente figura (Figura 1).

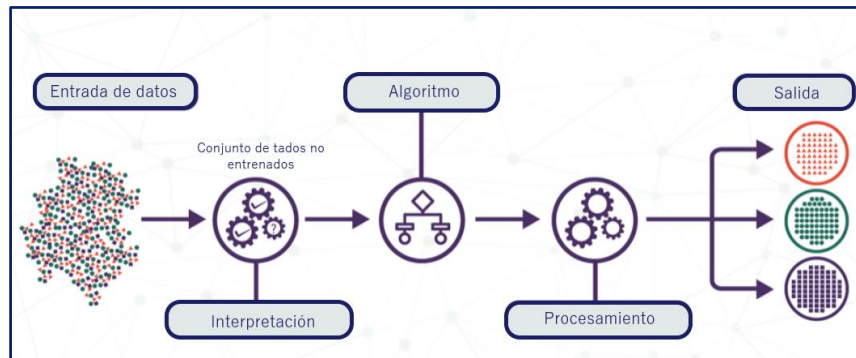


Figura 1.-Software de clasificación de datos

Fuente: <https://juandomingofarnos.files.wordpress.com/2018/11/machine-learning-explained2.png>

Los productos que esperamos una vez que el proyecto se ha desarrollado con éxito, son los siguientes:

- Sistema funcional
- Reporte Técnico

5. Metodología

Desarrollo en espiral: Es un modelo de proceso de software evolutivo que conjuga la naturaleza iterativa de construcción de prototipos con los aspectos controlados y sistemáticos del modelo lineal secuencial [7]. Definido por primera vez por Barry Boehm en 1986, utilizado generalmente en la ingeniería de software.

Las actividades de este modelo se conforman en una espiral, en la que cada bucle o iteración representa un conjunto de actividades. Las actividades no están fijadas a ninguna prioridad, sino que las siguientes se eligen en función del análisis de riesgo, comenzando por el bucle interior. La metodología de trabajo se explica exhaustivamente en “A Spiral Model of Software Development and Enhancement” por Barry Boehm. [8]

En cada ciclo se construye un modelo completo del sistema completo, puede combinarse con modelo cascada y evolutivo

Debido a que en el desarrollo del sistema se contempla la utilización de múltiples algoritmos esta metodología nos permite tener un modelo del sistema con la implementación de un algoritmo por ciclo, además que contamos con objetivos en común de los algoritmos, la cual causa una ventaja en tiempo de desarrollo permitiéndonos analizar los riesgos y enfocarnos más en la etapa de desarrollo y pruebas para la obtención de resultados.

Otra de las ventajas por las cuáles decidimos utilizar esta metodología, es porque una vez que hayamos analizado los riesgos de nuestro proyecto y se comience a realizar la construcción de los prototipos, esta nos ayudará a reducir los riesgos que habíamos planteado ya que esta metodología utiliza la construcción de prototipos como un mecanismo de reducción de riesgos.



Figura 2.- Desarrollo en espiral

Fuente: https://es.wikipedia.org/wiki/Desarrollo_en_espiral#/media/Archivo:ModeloEspiral.svg

6. Cronograma

Vea anexo 1.

7. Referencias

- [1] Galar Idoate, M. (2018, 22 de marzo). *Implementación del algoritmo de los k vecinos más cercanos (k-NN) y estimación del mejor valor local de k para su cálculo*. Academica-e. <https://academica-e.unavarra.es/xmlui/bitstream/handle/2454/29112/Memoria.pdf?sequence=2&isAllowed=y>
- [2] López Briega, R. E. (s. f.). *Introducción al Machine Learning*. Libro online de IAAR. <https://iaarbook.github.io/ML/>
- [3] Cruz Urbina, I. (2019). “Sistema clasificador de datos n–dimensional por medio de Inteligencia Artificial”. ESCOM, Ciudad De México, México.
- [4] Cervantes Canales, J. (2009). “*Clasificación de grandes conjuntos de datos vía Máquinas de Vectores Soporte y aplicaciones en sistemas biológicos*”. CINVESTAV, Ciudad De México, México.
- [5] Mena Roa, M. (2021, 21 de octubre). *El Big Bang del Big Data*. Statista. <https://es.statista.com/grafico/26031/volumen-estimado-de-datos-digitales-creados-o-replicados-en-todo-el-mundo/>
- [6] UNAM. (s. f.). *Inteligencia Artificial*. Apache Tomcat/8.5.68. <http://www.ptolomeo.unam.mx:8080/xmlui/bitstream/handle/132.248.52.100/219/A7.pdf>
- [7] Ecured cu. (2018). *Modelo espiral - EcuRed*. EcuRed. https://www.ecured.cu/Modelo_espiral
- [8] Boehm, B. (1986). *A Spiral Model of Software Development and Enhancement*. ACM SIGSOFT.

8. Alumnos y Director


Castillo Flores Magali. - Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Especialidad Sistemas, Boleta: 2018630172, Tel. 5534630221, email mcastillof1700@alumno.ipn.mx

Firma:



Torres Uraga Saul. - Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Especialidad Sistemas, Boleta: 2015131401, Tel. 5544428594, email storresu1400@alumno.ipn.mx

Firma:



Dr. Tonahtiu Arturo Ramírez Romero. - Doctor en Ingeniería en sistemas, profesor investigador. Áreas de interés: Inteligencia Artificial, bases de datos, desarrollo de sistemas web y sistemas complejos. Publicaciones en congresos nacionales e internacionales, así como en revistas científicas arbitrarias. Departamento de Ciencias e Ingeniería de la Computación, Escuela Superior de Computo, Tel. 57296000 Ext.: 52052 email-e: tonahtiu@yahoo.com

Firma:



CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Art. 3, fracc. II, Art. 18, fracc. II y Art. 21, lineamiento 32, fracc. XVII de la L.F.T.A.I.P.G.
PARTES CONFIDENCIALES: No. de boleta y

Anexo 1.- Cronograma

Castillo Flores Magali

Actividad	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
Planteamiento del problema											
Investigacion											
Entorno de desarrollo											
Analisis											
Requerimientos de sistema											
Analisis de la informacion											
Diseño											
Adaptacion de la arquitectura de software											
Diseño de la base de datos											
Entorno de desarrollo											
Desarrollo											
Implementacion de prototipo											
Pruebas de desarrollo de algoritmo											
Pruebas funcionales											
Documentación											
Reporte Tecnico											
Entrega de Trabajo Terminal I											
Desarrollo											
Implementacion de prototipo											
Pruebas de desarrollo de algoritmo											
Pruebas funcionales											
Documentación											
Reporte Tecnico											
Entrega de Trabajo Terminal II											

Anexo 1.- Cronograma

Torres Uraga Saul

[illegible]