
# Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores,
Unidad Morelia.
Lic. Tecnologías para la Información en Ciencias.
Sistemas Basados en Conocimiento.

---

# Airline Passenger Satisfaction

---

**P R O Y E C T O**

Por:
Erika Montserrat Correa Hernández
Luis Aáron Nieto Cruz

Maestros y ayudante:
Dra. Marisol Flores Garrindo
Dr. Luis Miguel García Velázquez
Lic.Javier Navarro

Morelia, Michoacán. Mayo 2023

# Índice



# Índice de figuras

# 1. Introducción

El Aprendizaje Automático es una rama de la Inteligencia Artificial más estudiada, está surgio apartir de 1940 y hasta la actualidad es una de las herramientas más utilizadas a lo largo del desarrollo de las tecnológias, problemas de clasificación y predicción.
El próposito de este reporte es dar un ejemplo mediante el uso de distintas técnicas y metodos Mátematicos del Aprendizaje Automático para predecir la satisfacción de los clientes de distintas aerolineas dado un conjunto de datos.
Los métodos con los que se optó trabajar fueron:
Naive Bayes, Regresión Lógistica y K-Nearest Neighbors.
De acuerdo a los resultados obtenidos se hará un análisis de cada método, para así tener las predicciones más optimas y acertadas.

# 2. Marco Teórico

**¿Qué es el Aprendizaje Automático?**

El Aprendizaje Automático es un campo de estudio dentro de la Inteligencia Artificial (IA), el cuál se centra en el desarollo de algoritmos y módelos de aprendizaje que permiten a las máquinas (computadoras) aprender y hacer predicciones.
Para poder llevar a cabo esto, se necesita la aplicación de distintos modelos matemáticos y, a su vez, algoritmos que permitan que las computadoras puedan analizar e interpretar datos, lo que posibilita la detección de patrones. Esto a su vez nos permite obtener resultados más precisos en forma de predicciones..

## 2.1. Tipos de Aprendizaje Automático

El Aprendizaje Automático se puede clasificar en distintos tipos, los cuales se categorizan de acuerdo a las necesidades que se tienen.

**Aprendizaje Supervisado**

El aprendizaje supervisado es el problema de inferir un mapeo a partir de pares de entrada-salida etiquetados. Dado un conjunto de ejemplos de entrenamiento, cada uno de los cuales consta de un vector de entrada y un vector objetivo correspondiente, el objetivo es aprender una función que pueda predecir con precisión la salida para entradas nuevas e invisibles. En otras palabras, el aprendizaje supervisado busca aprender un mapa o límite de decisión que separa diferentes clases o predice valores continuos basados en características de entrada.

**Aprendizaje supervisado por clasificación**

El aprendizaje automático por Clasificación se refiere al proceso de entrenamiento de un modelo de aprendizaje automático para clasificar o categorizar datos en diferentes clases o categorías predefinidas. Implica aprender un límite de decisión o una función de mapeo que pueda asignar con precisión nuevas instancias de datos invisibles a la clase correcta en función de sus características de entrada.

**Aprendizaje supervisado por Regresión**

El aprendizaje automático por regresión se refiere al proceso de entrenamiento de un modelo de aprendizaje automático para predecir valores numéricos continuos en función de las características de entrada. El análisis de regresión tiene como objetivo establecer una relación entre las variables de entrada y la variable de salida continua, lo que permite que el modelo haga predicciones o estime el valor de la variable objetivo para datos nuevos e invisibles.

**Aprendizaje no Supervisado**

El aprendizaje no supervisado implica entrenar un modelo de aprendizaje automático en datos sin etiquetar, donde los datos de entrada no tienen ninguna etiqueta de salida correspondiente. Estos algoritmos descubren patrones ocultos o agrupaciones de datos.

**Semi-Supervisado**

El aprendizaje semisupervisado es un problema de aprendizaje que involucra una pequeña cantidad de ejemplos etiquetados y una gran cantidad de ejemplos no etiquetados.
A continuación se harán experimentos con el uso de Aprendizaje Supervisado.

Figura 1: Imagen representativa
extraida de: link

# 3. Descripcion del conjunto de datos.

El conjunto de datos con el que se trabajó en este proyecto contiene una encuesta de satisfacción de pasajeros de aerolíneas.

Extraido de: Data.Set

Publicado por: TJ Klein.

Lo primero que hicimos fue visualizar los datos y notamos que contamos con 24 columnas, las cuales fueron: Unnamed: 0, id, Gender, Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes, satisfaction. Ninguna contaba con valores nulos, pero aun así decidimos eliminar las columnas 'Unnamed: 0' y 'id' para contar con una mejor organización. Después separamos en listas las columnas categóricas de las columnas numéricas.

La columna que vamos a predecir será la de satisfacción, por lo que se muestra una gráfica del total de pasajeros con su respectiva opinión.

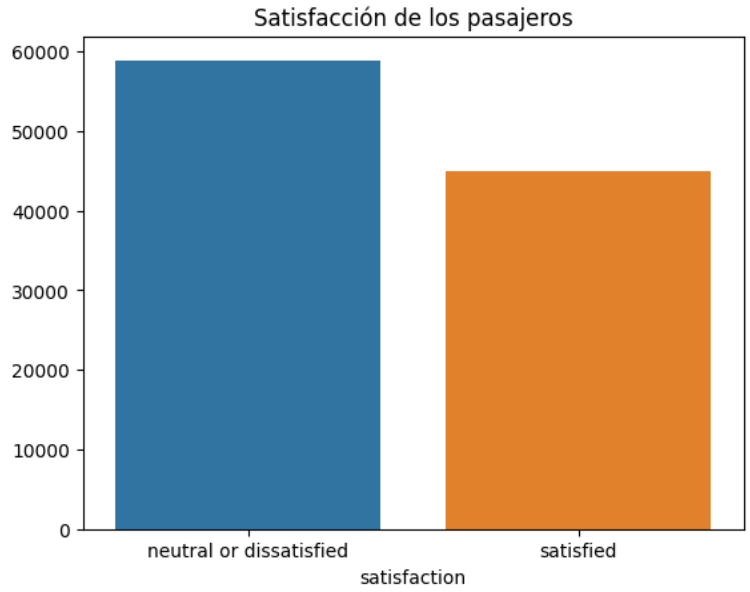


Figura 2: Neutral or dissatisfied: 58879, Satisfied: 45025.

A continuación, se observa la satisfacción de los pasajeros dependiendo de los demás atributos para poder medir y comparar su nivel de satisfacción:

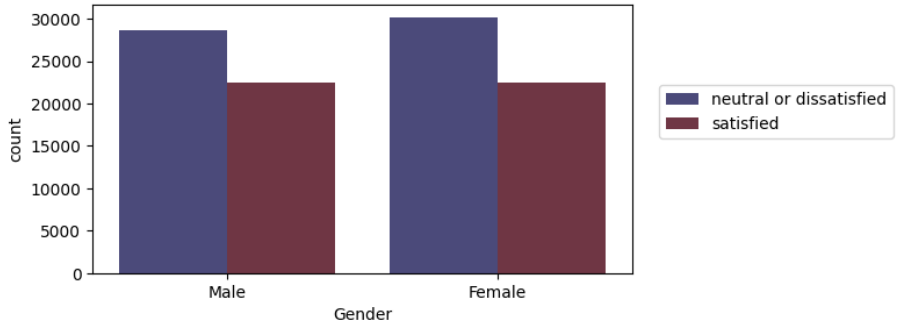


Figura 3: Satisfacción de acuerdo al Género.

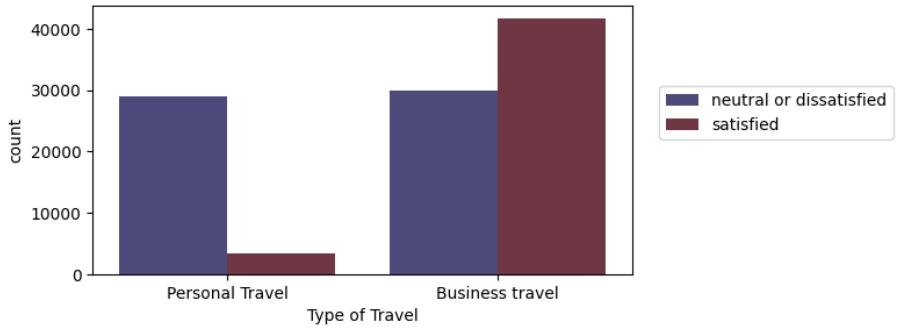


Figura 4: Satisfacción de acuerdo al Tipo de viaje.

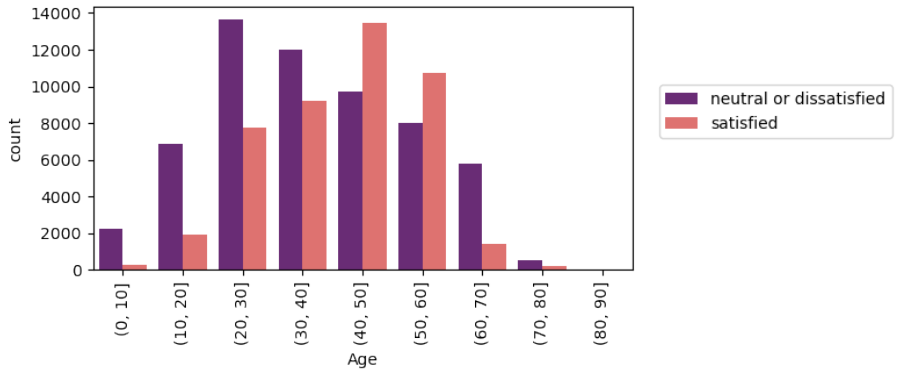


Figura 5: Satisfacción de acuerdo a la edad de los pasajeros.

Tomamos una muestra aleatoria de tamaño 1000 y revisamos la correlación entre los atributos numéricos de nuestro conjunto de datos.

El retraso en la salida y el retraso en la llegada del vuelo se encuentran fuertemente correlacionados, por lo que tenemos que eliminar cualquiera de los dos, en este caso se decidió eliminar el retraso en la salida para así evitar problemas de multicolinealidad.

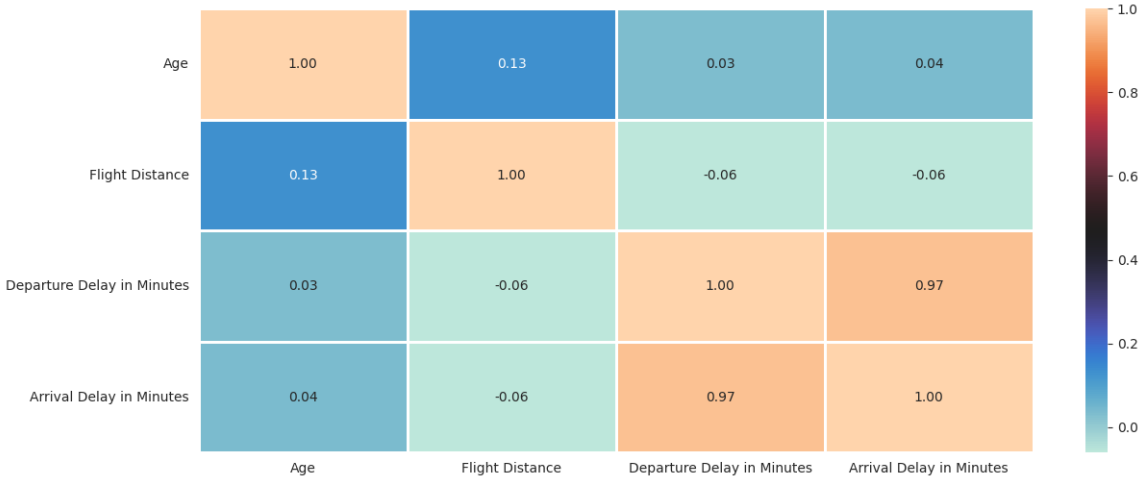


Figura 6: Correlación

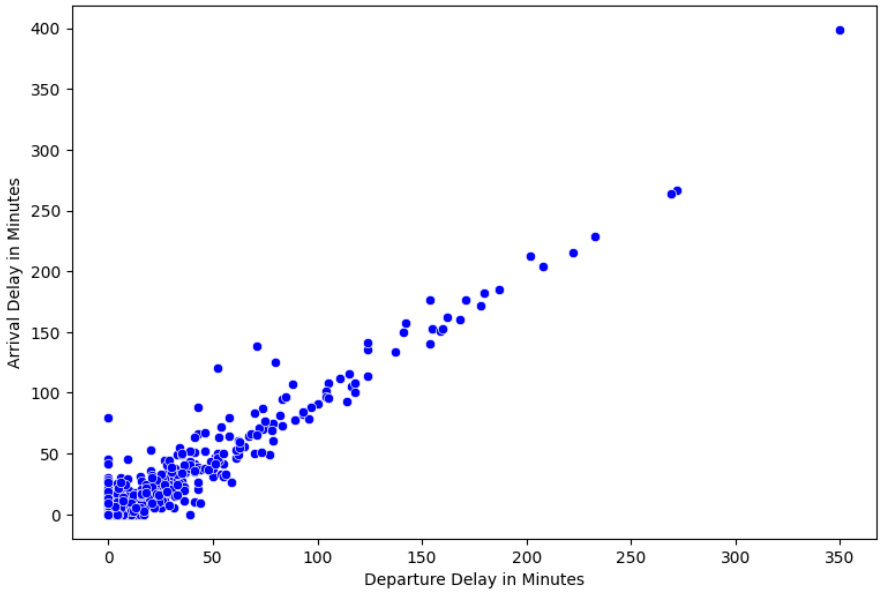


Figura 7: Dispersión

Después reemplazamos los valores nulos de nuestra columna 'Arrival Delay in Minutes' con el valor de la mediana.

# 4. Descripción de los métodos seleccionados

**Decidimos reaizar los experimentos con los siguientes 3 métodos:**

**Naive Bayes**
Decidimos realizar un experimento con el algoritmo Naive Bayes debido a que consideramos que la interpretación de los datos es más sencilla en cierta medida. Esto se debe a que la probabilidad a posteriori calculada por el modelo puede interpretarse directamente como la probabilidad de pertenecer a una clase específica, dado las características observadas..

**Regresión Lógistica**
Decidimos utilizar este método porque es recomendado para problemas de clasificación binaria, además de su eficiencia computacional es una técnica versátil que combina interpretación, eficiencia computacional y capacidad para manejar diferentes tipos de variables. Creímos que este método proporcionaría resultados más óptimos en comparación con los demás.

**K-Nearest Neighbors**
El último método utilizado fue el de KNN. Creíamos que este método proporcionaría buenos resultados debido a su implementación, que no implica un proceso de entrenamiento complejo. Se basa en el almacenamiento de los datos de entrenamiento y el cálculo de la distancia entre los puntos. Sin embargo, también consideramos que podría llevar más tiempo calcular las predicciones debido a la cantidad de datos y su dispersión, lo cual resultó ser cierto. A pesar de esto, los resultados obtenidos fueron óptimos.

# 5. Experimentos

Aplicamos los modelos de aprendizaje automático con sklearn. Utilizando como habíamos mencionado anteriormente:
Naive Bayes
Regresión Logistica
K-Nearest Neighbors.

Medimos el desempeño con la validación cruzada considerando las métricas de:
-Exactitud (Accuracy)
-Precisión
-Sensibilidad (Recall)
-F1

Hacemos una función para que dado un modelo nos devuelva el resultado de la validación cruzada, especificando: el arreglo obtenido, su promedio y el promedio de los promedios (El desempeño total de cada clasificador).

Utilizaremos la función cross val score (k=5) con la que especificaremos: el clasificador, la matriz de datos, el vector de etiquetas y un scoring (para definir la métrica).

Para la Precisión, Sensibilidad (Recall) y F1 utilizaremos el average weighted de sklearn.

## 5.1. Naive Bayes

En la matriz de confusión podemos observar los resutados al aplicar el método de Naive Bayes.

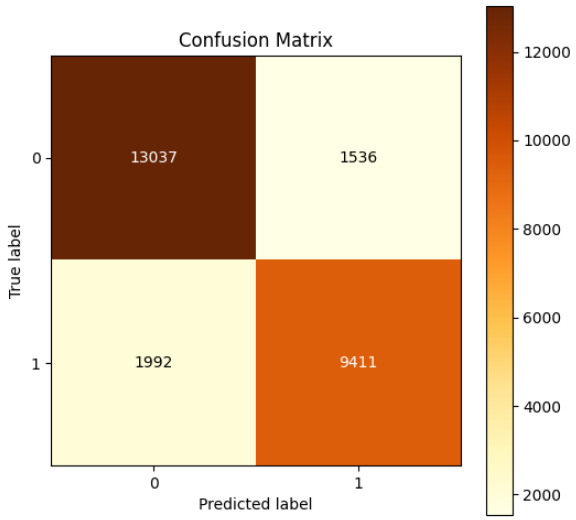


Figura 8: Matriz de Confusión Naive Bayes

## 5.2. Regresión Lógistica

En la matriz de confusión podemos observar los resultados obtenidos al aplicar el método de Regresión Lógistica.

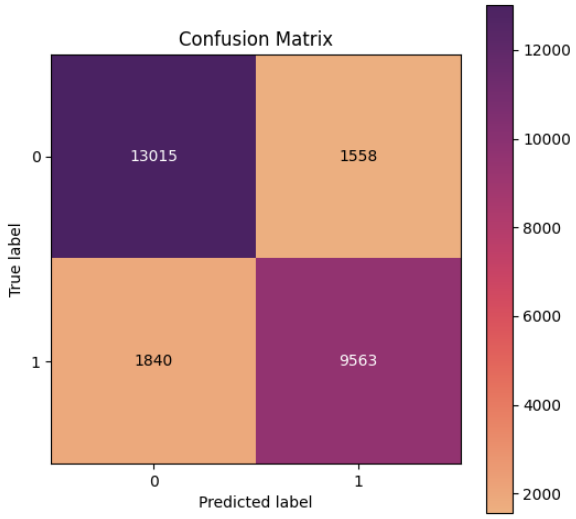


Figura 9: Matriz de Confusión Regresion

## 5.3. K-Nearest Neighbors

En la matriz de confusión podemos observar los resutados al aplicar el método de KNN.

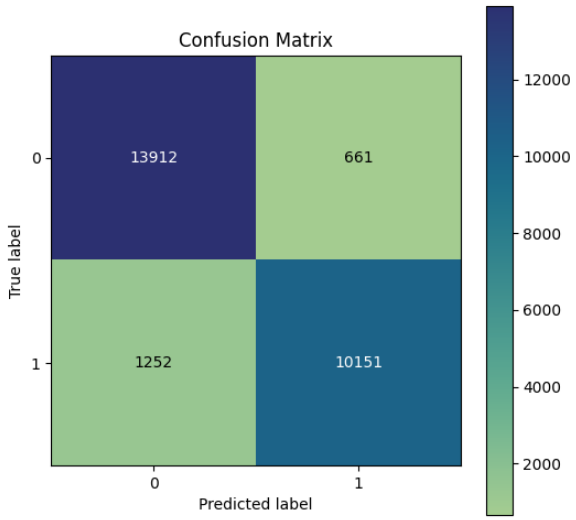


Figura 10: Matriz de Confusión KNN

Como observamos que los niveles de este modelo mejoraban drásticamente en comparación con los demás, decidimos gráficar una Curva de ROC (Característica de Operación del Receptor).

La interpretación de la curva ROC se basa en su forma y posición con respecto a la línea de referencia, que es una línea diagonal que representa el desempeño aleatorio del modelo. Cuanto más se acerque la curva ROC a la esquina superior izquierda del gráfico, mejor será el desempeño del modelo.

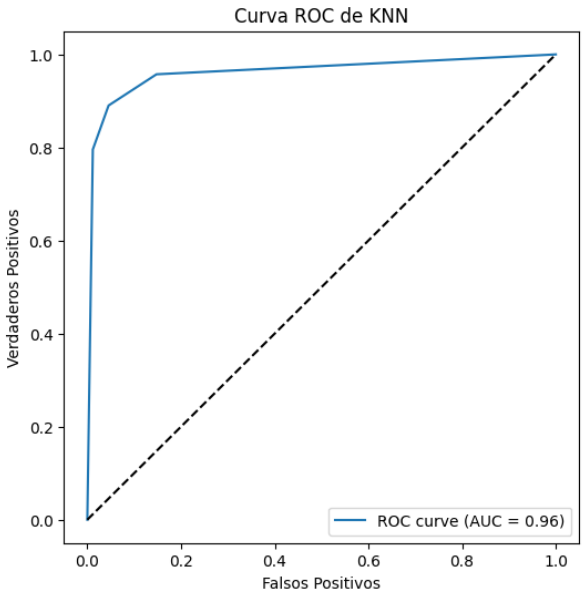


Figura 11: Curva de ROC

# 6. Evaluación de Resultados Obtenidos

Una vez implementados los modelos, estos fueron los resultados que se obtuvieron de cada métrica:.

| Método | Exactitud | Precisión | Sensibilidad | F1 | Promedio |
|---|---|---|---|---|---|
| Naive Bayes | 0.868 | 0.868 | 0.868 | 0.867 | 0.868 |
| Regresión Lógistica | 0.875 | 0.875 | 0.875 | 0.874 | 0.875 |
| KNN | 0.925 | 0.926 | 0.925 | 0.925 | 0.925 |

Cuadro 1: Evaluación de Resultados.

# 7. Análisis y discusión de Resultados

Una vez obtenidos los resultados de cada uno de los experimentos nos dimos la tarea de graficar la comparación de los modelos, donde podemos concluir que el modelo ganador fue KNN ya que tuvo el mayor promedio en todas las métricas, de ahí le sigue Regresión Logística y como el peor tenemos a Naive Bayes.

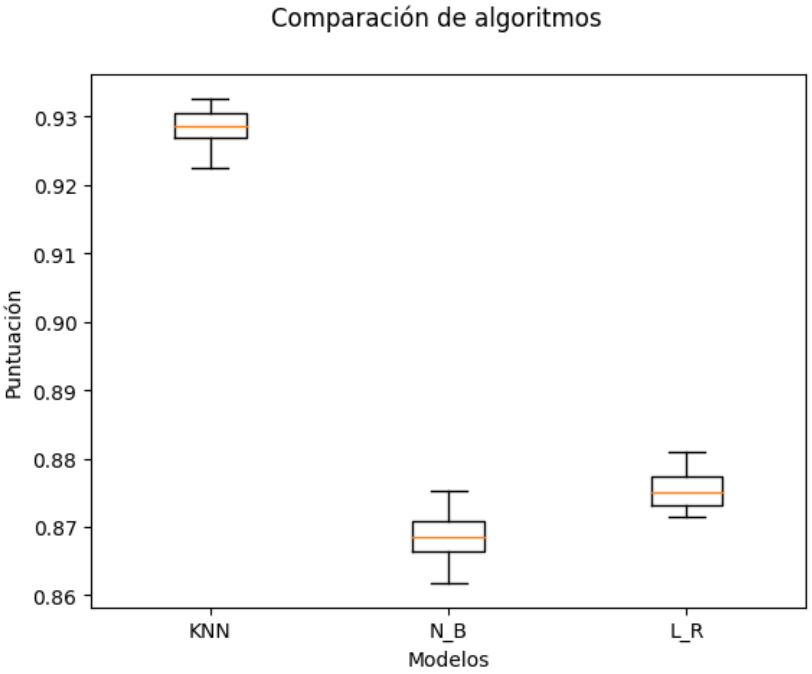


Figura 12: Comparación de Métodos.

# 8. Conclusión

El Machine Learning nos ayuda de diversas formas en diferentes campos dónde se desea realizar una predicción o clasificación. Predice de acuerdo a los datos que se tienen por lo que si en un determinado caso no tenemos datos balanceados, podríamos tener sesgos que nos impidan predecir de manera acertada. También es importante recalcar que realizar distintos experimentos nos ayuda a poder evaluar al Álgoritmo con el Método más eficiente, por más que tu intuyas que alguno será mejor que otro, siempre es mejor realizar pruebas con cada uno y después hacer la comparación de acuerdo a lo que se desea predecir o casificar.

# 9. Referencias

Asiri, S. (2022). An Introduction to Classification in Machine Learning. Built In. https://builtin.com/machine-learning/classification-machine-learning
De Ceupe, B. (s. f.). Aprendizaje supervisado: Qué es, tipos y ejemplo. Ceupe. https://www.ceupe.com/blog/aprendizaje-supervisado.html
Roman, V. (2021, 9 diciembre). Algoritmos Naive Bayes: Fundamentos e Implementación. Medium. https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaciUcan, R. H. (2023, 21 abril). Método de los K vecinos más cercanos - SoldAI - Medium. Medium. https://medium.com/soldai/m