Universidad Autónoma de Sinaloa



Facultad de Ingeniería Mochis

Materia:

Sistema de apoyo a la toma de decisiones

Docente:

Dr. Alan David Ramírez Noriega

Desarrollar un clasificador basado en el método Bayes Ingenuo

Grupo 4-01

Alumnos:

Castro Aguilar Allan Enrique

Jimenez Apodaca Brandon Ulises

Gardea Palafox Irán Alondra

Fecha:10 de diciembre de 2021

**Índice**

Contenido

[Introducción 3](#_Toc92497791)

[¿Qué es el aprendizaje maquina? 3](#_Toc92497792)

[¿Qué es un clasificador? 3](#_Toc92497793)

[Describir que se hará en este proyecto. 3](#_Toc92497794)

[Marco teórico 4](#_Toc92497795)

[Aprendizaje Maquina 4](#_Toc92497796)

[Algoritmos supervisados y no supervisados 4](#_Toc92497797)

[Algoritmo de clasificación Bayes Ingenuo 5](#_Toc92497798)

[Probabilidad 6](#_Toc92497799)

[Teorema de Bayes 6](#_Toc92497800)

[Bayes Ingenuo 7](#_Toc92497801)

[Desarrollo del software 9](#_Toc92497802)

[Análisis 9](#_Toc92497803)

[Diseño 9](#_Toc92497804)

[Arquitectura basada en MVC (Modelo Vista Controlador) 10](#_Toc92497805)

[Casos de uso 11](#_Toc92497806)

[Diagrama de casos de uso 11](#_Toc92497807)

[Caso de uso detallado 11](#_Toc92497808)

[Diagrama de secuencia 13](#_Toc92497809)

[Diagrama de clases 13](#_Toc92497810)

[Algoritmos del sistema (pseudocódigo o diagramas de flujo) 13](#_Toc92497811)

[Interfaz 14](#_Toc92497812)

[Implementación 14](#_Toc92497813)

[Experimento 14](#_Toc92497814)

[Experimento 15](#_Toc92497815)

[Dataset 15](#_Toc92497816)

[Resultados 15](#_Toc92497817)

[Discusión 15](#_Toc92497818)

[Conclusiones 16](#_Toc92497819)

[Referencias 17](#_Toc92497820)

[Anexos 18](#_Toc92497821)

[Código 18](#_Toc92497822)

[Generar documentación del código 18](#_Toc92497823)

[Descripción de los roles del proyecto: 18](#_Toc92497824)

# Introducción

## ¿Qué es el aprendizaje maquina?

Es el área de la ciencia computacional que se centra en el análisis y la interpretación de patrones y estructuras de datos que hacen posible el aprendizaje, el razonamiento y la toma de decisiones sin interacción humana.

Dicho de otro modo, el aprendizaje automático permite que el usuario alimente un algoritmo informático con una cantidad ingente de datos, a partir de los cuales el ordenador analice toda la información y sea capaz de tomar decisiones y hacer recomendaciones basándose únicamente en los datos introducidos. En el caso de identificar correcciones, el algoritmo puede incorporar esa información para mejorar la toma de decisiones futura.

## ¿Qué es un clasificador?

Se utiliza en referencia al algoritmo utilizado para asignar un elemento entrante no etiquetado en una categoría concreta conocida. Dicho algoritmo, permite pues, ordenar o disponer por clases elementos entrantes, a partir de cierta información característica de estos. Una manera de implementar un clasificador es seleccionar un conjunto de ejemplos etiquetados y tratar de definir una regla que permita asignar una etiqueta a cualquier otro dato de entrada.

## Describir que se hará en este proyecto.

Clasificar cosas es parte de nuestra vida desde que somos pequeños, este proceso está en muchas de nuestras actividades cotidianas, debido a esto, se necesitan herramientas que faciliten el proceso de clasificación de grandes cantidades de información, y que sea fácil catalogar personas u objetos con base en ciertos criterios.

En este proyecto se llevará a cabo el desarrollo de una herramienta de software basado en el algoritmo de Bayes Ingenuo (Naive Bayes) que nos permitirá clasificar instancias para apoyar la toma de decisiones en diversas áreas.

# Marco teórico

## Aprendizaje Maquina

El “machine learning” o aprendizaje automático es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Una habilidad indispensable para hacer sistemas capaces de identificar patrones entre los datos para hacer predicciones. Esta tecnología está presente en un sinfín de aplicaciones como las recomendaciones de Netflix o Spotify, las respuestas inteligentes de Gmail o el habla de Siri y Alexa.

“En definitiva, el “machine learning” es un maestro del reconocimiento de patrones, y es capaz de convertir una muestra de datos en un programa informático capaz de extraer inferencias de nuevos conjuntos de datos para los que no ha sido entrenado previamente”, explica José Luis Espinoza, científico de datos de BBVA México. Esta capacidad de aprendizaje se emplea también para la mejora de motores de búsqueda, la robótica, el diagnóstico médico o incluso la detección del fraude en el uso de tarjetas de crédito.

Algunos de los usos prácticos más habituales del aprendizaje máquina:

* Seguridad informática, diagnóstico de ataques, prevención de fraude Online, detección de anomalías, etc.
* Reconocimiento de imágenes o patrones (facial, dactilar, objetos, voz, etc).
* Conducción autónoma, mediante algoritmos deep learning: identificación de imágenes en tiempo real, detección de obstáculos y señales de tráfico, prevención de accidentes, etc.
* Salud: evaluación automática de pruebas diagnósticas, robótica médica, etc.
* Análisis de mercado de valores (predicciones financieras, evolución de mercados etc).
* Motores de recomendación.

## Algoritmos supervisados y no supervisados

**Supervisado**. En el aprendizaje supervisado, los algoritmos trabajan con datos “etiquetados” (labeled data), intentado encontrar una función que, dadas las variables de entrada (input data), les asigne la etiqueta de salida adecuada. El algoritmo se entrena con un “histórico” de datos y así “aprende” a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor de salida.

Se suele usar en:

* Problemas de clasificación (identificación de dígitos, diagnósticos, o detección de fraude de identidad).
* Problemas de regresión (predicciones meteorológicas, de expectativa de vida, de crecimiento etc).

Estos dos tipos principales de aprendizaje supervisado, clasificación y regresión, se distinguen por el tipo de variable objetivo. En los casos de clasificación, es de tipo categórico, mientras que, en los casos de regresión, la variable objetivo es de tipo numérico.

Los algoritmos más habituales que aplican para el aprendizaje supervisado son:

* Árboles de decisión.
* Clasificación de Naïve Bayes.
* Regresión por mínimos cuadrados.
* Regresión Logística.
* Support Vector Machines (SVM).
* Métodos “Ensemble” (Conjuntos de clasificadores).

**No supervisado**. El aprendizaje no supervisado tiene lugar cuando no se dispone de datos “etiquetados” para el entrenamiento. Sólo conocemos los datos de entrada, pero no existen datos de salida que correspondan a un determinado input. Por tanto, sólo podemos describir la estructura de los datos, para intentar encontrar algún tipo de organización que simplifique el análisis. Por ello, tienen un carácter exploratorio.

El aprendizaje no supervisado se suele usar en:

* Problemas de clustering.
* Agrupamientos de co-ocurrencias.
* Perfilado o profiling.

Sin embargo, los problemas que implican tareas de encontrar similitud, predicción de enlaces o reducción de datos, pueden ser supervisados o no. Los tipos de algoritmo más habituales en aprendizaje no supervisado son:

* Algoritmos de clustering.
* Análisis de componentes principales.
* Descomposición en valores singulares (singular value decomposition).
* Análisis de componentes principales (Independent Component Analysis).

## Algoritmo de clasificación Bayes Ingenuo

Tanto en probabilidad como en minería de datos, un clasificador ingenuo Bayesiano (clasificador Naive Bayes) es un método probabilístico que tiene sus bases en el teorema de Bayes y recibe el apelativo de ingenuo dadas algunas simplificaciones adicionales que determinan la hipótesis de independencia de las variables predictores.

El argumento de Bayes no es que el mundo sea intrínsecamente probabilístico o incierto, sino que aprendemos sobre el mundo a través de la aproximación, acercándonos cada vez más a la verdad, a medida que recogemos más evidencias.

En términos sencillos, el clasificador ingenuo de Bayes asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica. Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro.

Un clasificador ingenuo de Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.

En muchas aplicaciones prácticas, la estimación de parámetros para los modelos de Bayes utiliza el método de máxima verosimilitud, es decir, se puede trabajar con el modelo ingenuo de Bayes sin aceptar la probabilidad bayesiana o cualquiera de los métodos bayesianos.

Una ventaja del clasificador ingenuo de Bayes es que solo se requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros necesarios para la clasificación (las medidas y las varianzas de las variables).

Solo es necesario determinar las varianzas de las variables de cada clase y no toda la matriz de covarianza. Para otros modelos de probabilidad, los clasificadores ingenuos de Bayes se pueden entrenar en entornos de aprendizaje supervisado.

## Probabilidad

La probabilidad es un método por el cual se obtiene la frecuencia de un acontecimiento determinado mediante la realización de un experimento aleatorio, del que se conocen todos los resultados posibles, bajo condiciones suficientemente estables. La teoría de la probabilidad se usa extensamente en áreas como la estadística, la física, la matemática, las ciencias y la filosofía para sacar conclusiones sobre la probabilidad discreta de sucesos potenciales y la mecánica subyacente discreta de sistemas complejos, por lo tanto es la rama de las matemáticas que estudia, mide o determina a los experimentos o fenómenos aleatorios. La definición de probabilidad surge debido al deseo del ser humano por conocer con certeza los eventos que sucederán en el futuro. Es por eso que a través de la historia se han desarrollado diferentes enfoques para tener un concepto de la probabilidad y determinar sus valores.

## Teorema de Bayes

El teorema de Bayes se ha considerado como un método válido que permite calcular probabilidades condicionales. El enunciado de este postulado establece que se puede calcular la probabilidad de que ocurra un suceso A sabiendo que existe un suceso B que condiciona al primero y viceversa. Por lo general, suele ser útil como método predictivo, causal y de diagnóstico, tomando en cuenta información que se conoce con anterioridad, y nueva información que se relaciona con los eventos.

Teniendo esto tres factores de estudio, se define que para casos de predicción, se busca responder interrogantes en torno a la probabilidad de un suceso, utilizando información acerca de la ocurrencia de otros sucesos predictores. En sentido causal, se busca calcular la probabilidad de que un evento sea consecuencia de otro. Mientras que para diagnóstico, se busca conocer la probabilidad de que ocurra alguno de los anteriores, teniendo como información previa las consecuencias de estos. Esto se puede resumir en dos premisas:

* En casos de estudio de predicción y causales se desconocen las consecuencias, pero la información que se maneja son las causas.
* En casos de estudio de diagnóstico se conocen las consecuencias, mientras que no se posee información de las causas.

Lo que hace que el teorema de Bayes sea un método efectivo de usar, es que puede ser aplicado para el cálculo de distintos tipos de probabilidades, y en cualquier ciencia. A pesar de considerarse un método efectivo, muchos estadistas cuestionan su validez; esto se debe a que consideran que solo es aplicable únicamente en casos donde se presentan eventos excluyentes y exhaustivos. También hacen énfasis en el hecho que al momento de llevar a cabo el procedimiento, solo se basan en condiciones subjetivas.

## Bayes Ingenuo

Naïve Bayes o el Ingenuo Bayes es uno de los algoritmos más simples y poderosos para la clasificación basado en el Teorema de Bayes con una suposición de independencia entre los predictores. Naive Bayes es fácil de construir y particularmente útil para conjuntos de datos muy grandes.

El clasificador Naive Bayes asume que el efecto de una característica particular en una clase es independiente de otras características. Por ejemplo, un solicitante de préstamo es deseable o no dependiendo de sus ingresos, historial de préstamos y transacciones anteriores, edad y ubicación. Incluso si estas características son interdependientes, estas características se consideran de forma independiente. Esta suposición simplifica la computación, y por eso se considera ingenua. Esta suposición se denomina independencia condicional de clase.

Bayesiano ingenuo es un algoritmo muy utilizado para resolver problemas de clasificación. El modelo se denomina Naive porque trata todas las variables de predicción propuestas como independientes unas de otras. El bayesiano ingenuo es un algoritmo rápido y escalable que calcula las probabilidades condicionales para las combinaciones de atributos y el atributo de objetivo. A partir de los datos de entrenamiento se establece una probabilidad independiente. Esta probabilidad proporciona la verosimilitud de cada clase objetivo, una vez dada la instancia de cada categoría de valor a partir de cada variable de entrada.

La validación cruzada se utiliza para comprobar la precisión del modelo en los mismos datos que se utilizaron para generar el modelo. Resulta especialmente útil cuando el número de casos disponible para generar un modelo es pequeño.

El resultado del modelo se puede examinar en un formato de matriz. Los números de la matriz son probabilidades condicionales relacionadas con las clases (columnas) predichas y las combinaciones (filas) del valor de la variable predictora.

Ventajas

* Es fácil y rápido predecir la clase de conjunto de datos de prueba. También funciona bien en la predicción multiclase.
* Cuando se mantiene la suposición de independencia, un clasificador Naive Bayes funciona mejor en comparación con otros modelos como la Regresión Logística y se necesitan menos datos de entrenamiento.
* Funciona bien en el caso de variables de entrada categóricas comparada con variables numéricas.

Desventajas

* Si la variable categórica tiene una categoría en el conjunto de datos de prueba, que no se observó en el conjunto de datos de entrenamiento, el modelo asignará una probabilidad de 0 y no podrá hacer una predicción. Esto se conoce a menudo como frecuencia cero. Para resolver esto, podemos utilizar la técnica de alisamiento.
* Otra limitación de Naive Bayes es la asunción de predictores independientes. En la vida real, es casi imposible que obtengamos un conjunto de predictores que sean completamente independientes.
* Naive Bayes es el algoritmo más sencillo y potente. A pesar de los significativos avances de Machine Learning en los últimos años, ha demostrado su valía. Se ha implementado con éxito en muchas aplicaciones, desde el análisis de texto hasta los motores de recomendación.

# Desarrollo del software

## Análisis

Desarrollar un clasificador Bayes ingenuo que reciba un dataset en formato de texto o separado por comas y genere las metricas de evaluación, una matriz de confusión y determine el accuracy.

Dicho clasificador debe contener opciones de configuración para indicar si cuenta con encabezados, indicar si la clase se encuntra al inicio o al final del dataset y un intervalo de discretización. De igual manera la capacidad de poder probar con un porcentaje de entrenamiento indicado por el usuario.

Para el tratado de las variables discretas o el proceso de discretización se utiliza el metodo de frecuencias iguales y para el tratado de los valores en 0 se utiliza el suavizado de Laplace.

Diseño

* **1. Barra de encabezado.** Dentro de ella se puede localizar el título del programa (Clasificador Naive - Bayes).
* **2. Botón de salir.** Para salir del programa, debe presionarse el botón representado con una *x* ubicado en la parte posterior derecha de la ventana del programa.
* **3. Cuadro de texto “Cargar Dataset”.** Dentro del cuadro de texto de Cargar Dataset el cual es de sólo lectura (Read Only – De solo lectura), se puede observar la ruta del archivo que fue seleccionado a través del botón que contiene el símbolo de tres puntos (…). Al dar click sobre este, se abrirá una ventana desde la cual se podrá elegir el archivo deseado y en la extensión requerida (.txt, .csv, .xlsx).
* **4. Apartado “Contiene encabezado”.** El apartado de contiene encabezado sirve para especificar al programa si el documento de texto que fue proporcionado por el usuario contiene una línea de encabezado, así se evitan errores de procesamiento. El usuario debe elegir la opción *sí* si su documento contiene un encabezado, de lo contrario deberá elegir la opción *no*.
* **5. Apartado “Pruebas”.** Activa el modo de pruebas del programa, para realizar una ejecución tomando en cuenta el método de cálculo para la validación simple. Dentro del apartado se encuentra la opción de “Mismo dataset”, la cual deberá marcarse si se desea habilitar el cálculo de validación simple, además de proporcionar un porcentaje de entrenamiento para dicho cálculo.
* A través del botón de análisis puede ejecutarse el programa considerando dicho procedimiento.
* **6. Cuadro de texto “Intervalo de discretización”.** Dentro de este cuadro de texto se coloca el valor que se le asignará al intervalo de discretización que utilizará nuestro programa para realizar los cálculos correspondientes.
* **7. Cuadro de análisis.** Corre el programa de forma general sin incluir cálculo de métodos extra.
* **8. Cuadro de Dataset.** Dentro de dicho apartado se mostrará el Dataset que el usuario proporcionó al programa.
* **9. Cuadro de métricas de evaluación.** Muestra los datos resultantes del cálculo ejecutado sobre nuestro Dataset (Precisión, exhaustividad, F1 score).
* **10. Cuadro de matriz de confusión.** Muestra los resultados después de realizar la comparativa correspondiente en el cálculo de matriz de confusión.
* **11. Cuadro de texto “Accuracy”.** Muestra el valor del porcentaje de precisión de nuestro Dataset proporcionado.

## Arquitectura basada en MVC (Modelo Vista Controlador)

CONTROLADOR

BOTONES QUE COORDINAN EL CLASIFICADOR

(CARGAR DATASET, CALCULAR ACCURACY, GENERAR MATRIZ DE CONFUSIÓN, GENERAR MÉTRICAS DE EVALUACIÓN)

MODELO

CÓDIGO DEL CLASIFICADOR

VISTA

VENTANA PRINCIPAL

DEL CLASIFICADOR

ACTUALIZA

RESPONDE

MANIPULA

## Casos de uso

## Diagrama de casos de uso

## Caso de uso detallado

|  |  |  |
| --- | --- | --- |
| **Nombre del caso de uso:** | | Cargar ruta dataset |
| **Curso normal de los eventos** | | |
| **Acción del actor:** | **Respuesta del sistema:** | |
| 1. Ingresa al programa.  3. Da clic al botón de 3 puntitos.  5. Selecciona el archivo deseado. | 2. Muestra la pantalla principal con todos los campos.  4. Despliega una lista de archivos para que seleccione.  6. Guarda la ruta en el txt que tiene a un lado el botón. | |

|  |  |  |
| --- | --- | --- |
| **Nombre del caso de uso:** | | Cargar dataset con encabezados |
| **Curso normal de los eventos** | | |
| **Acción del actor:** | **Respuesta del sistema:** | |
| 1. Selecciona la opción de contener encabezados. | 2. Muestra en la parte superior el archivo cargado con todos los datos y con la primera línea mostrada como encabezado. | |

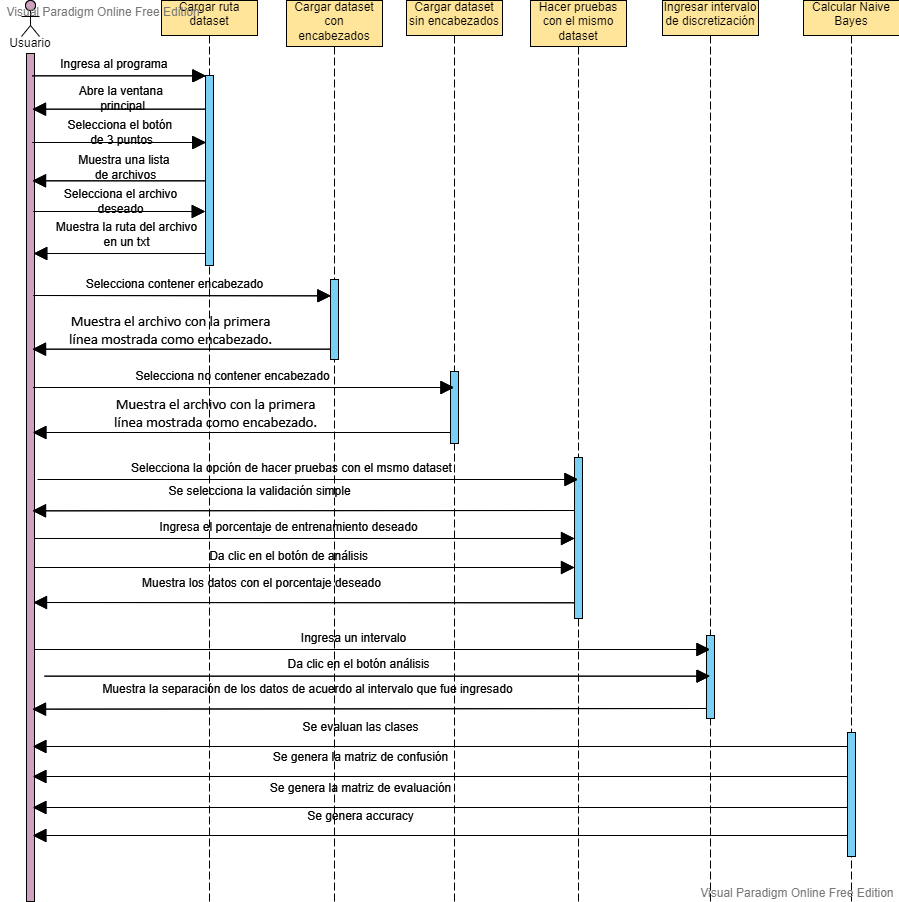
|  |  |  |
| --- | --- | --- |
| **Nombre del caso de uso:** | | Cargar dataset sin encabezados |
| **Curso normal de los eventos** | | |
| **Acción del actor:** | **Respuesta del sistema:** | |
| 1. Selecciona la opción de no contener encabezados. | 2. Muestra en la parte superior el archivo cargado con todos los datos sin mostrar ningún encabezado. | |

|  |  |  |
| --- | --- | --- |
| **Nombre del caso de uso:** | | Hacer pruebas con el mismo dataset |
| **Curso normal de los eventos** | | |
| **Acción del actor:** | **Respuesta del sistema:** | |
| 1. Selecciona la opción de hacer pruebas con el mismo dataset.  3. Ingresa el porcentaje de entrenamiento deseado.  4. Da clic en análisis. | 2. Sé selecciona la validación simple.  5. Muestra los datos pero con el porcentaje ingresado. | |

|  |  |  |
| --- | --- | --- |
| **Nombre del caso de uso:** | | Ingresar intervalo de discretización |
| **Curso normal de los eventos** | | |
| **Acción del actor:** | **Respuesta del sistema:** | |
| 1. Ingresa un intervalo.  2. Da clic al botón análisis. | 3. Muestra la separación de los datos de acuerdo al intervalo que fue ingresado. | |

|  |  |  |
| --- | --- | --- |
| **Nombre del caso de uso:** | | Calcular Naive Bayes |
| **Curso normal de los eventos** | | |
| **Acción del actor:** | **Respuesta del sistema:** | |
|  | 1. Se evalúan las clases.  2. Se genera la matriz de confusión.  3. Se genera la matriz de evaluación  4. Se genera accuracy. | |

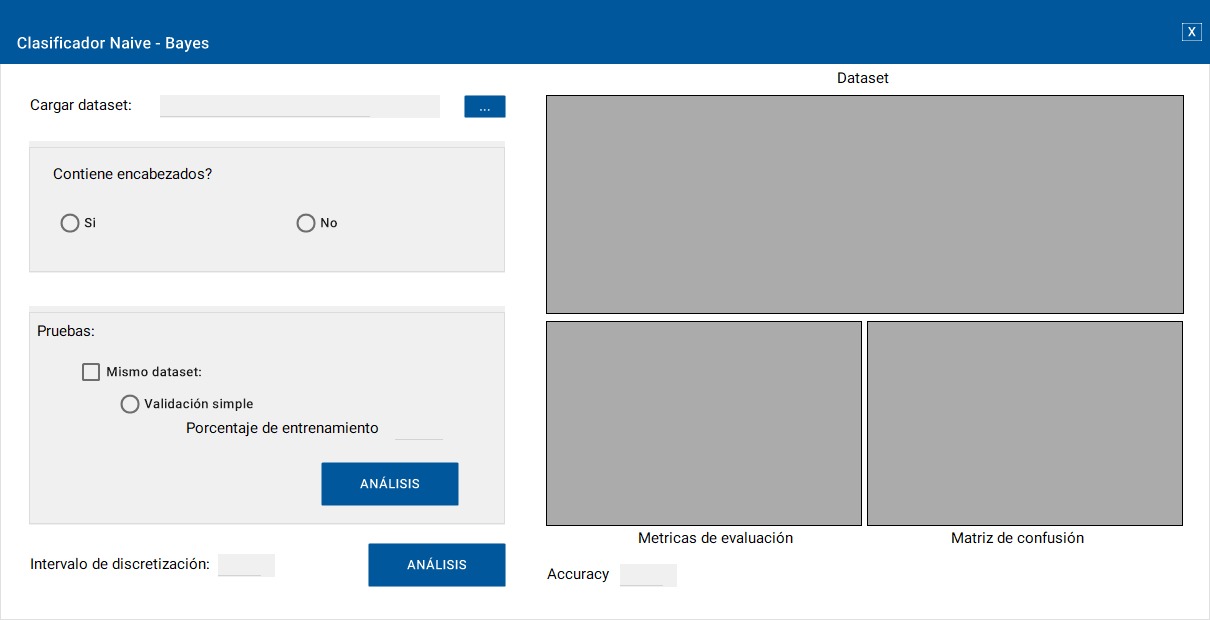
## Diagrama de secuencia



## Diagrama de clases

## Algoritmos del sistema (pseudocódigo o diagramas de flujo)

## Interfaz



## Implementación

Para implementar lo estipulado en el proyecto se hizo uso del lenguaje de programación C# con Windows forms, para el versionado de documentación y desarrollo se utilizó la plataforma de Git con la aplicación Github Desktop.

Las librerías utilizadas fueron:

system.IO: Librería para poder leer y escribir archivos externos o dentro del dispositivo

MaterialSkin: Librería de diseño de formularios para C# de estilo Android.

# Experimento

Experimento**:** Describir la forma en como probaremos el clasificador (elegir un

dataset y probar con el)

Dataset**:** Describir el dataset.

Resultados**:** Mostrar gráficas y tablas de los resultados de las métricas de evaluación

Discusión**:** Describir los resultados, fueron buenos, regulares o malos, justificar el funcionamiento del software.

# Conclusiones

A lo largo del desarrollo de este proyecto comprendimos el uso y adaptación de algoritmos automatizados que puedan funcionar sin contar con datos fijos, más bien datos precargados por dependencias externas y de los cuales la aplicación tiene total desconocimiento.

El desarrollo del clasificador fue todo un reto para todo el equipo, ya que implicó bastantes horas de lectura por parte de todos los integrantes para conocer el funcionamiento o la lógica que debe seguir el programa, así mismo, se dedicaron horas específicamente para probar y validar el programa, corrigiendo cualquier especie de bug o irregularidad.

Fue un trabajo de bastantes horas que pensamos que fueron bien invertidas con un conocimiento que nos sirve como experiencia para futuras materias o proyectos que realicemos.

# Referencias

Gonzalez, L. (20 de Septiembre de 2019). *aprendeia*. Obtenido de https://aprendeia.com/naive-bayes-teoria-machine-learning/

Jacob. (19 de Agosto de 2021). *Jacobsoft*. Obtenido de https://www.jacobsoft.com.mx/es\_mx/clasificador-naive-bayes/

Marta. (Mayo de 2020). *superprof*. Obtenido de https://www.superprof.es/apuntes/escolar/matematicas/probabilidades/combinatoria/teorema-de-bayes.html

Martinez, J. (3 de Junio de 2020). *IArtificial.net*. Obtenido de https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

Microsoft. (16 de Agosto de 2021). *Microsoft*. Obtenido de https://docs.microsoft.com/en-us/dynamics365/finance/finance-insights/confusion-matrix

Roman, V. (25 de Abril de 2019). *Ciencia y datos*. Obtenido de https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f

# Anexos

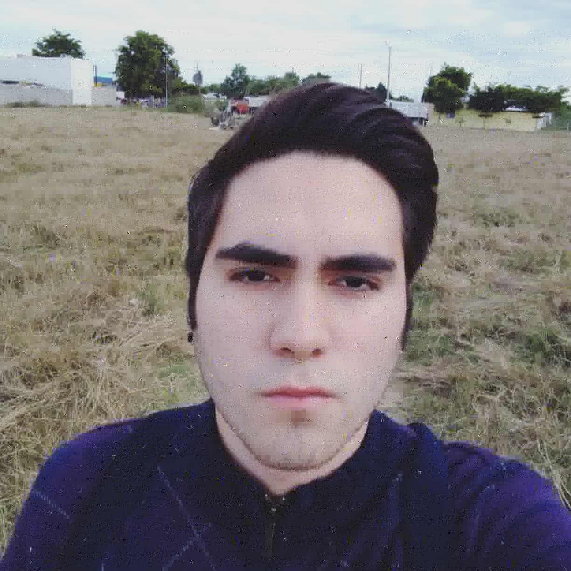
## Código

## Generar documentación del código

## Descripción de los roles del proyecto:



Programador, encargado de revisión de errores.



Programador, encargado de testeos.



Desarrollo de documentación, investigación teórica, creación de diagramas.