PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

# SELECCIÓN DE VARIABLES DE ENTRADA PARA PROCESOS AUTORREGRESIVOS, MEDIANTE EL CÁLCULO DE LA INFORMACIÓN MUTUA USANDO LOS K-VECINOS MÁS CERCANOS

**FRANCISCO JAVIER PACHECO MELO**
**JAVIER LORENZO ROJAS VILCHES**

INFORME DE PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO DE EJECUCIÓN INFORMÁTICA

OCTUBRE, 2014

# Resumen

Para solucionar el problema de la selección de variables de entrada a un sistema, para un proceso autorregresivo, se busca seleccionar las variables más relevantes y menos redundantes, que ayuden a reducir el consumo de tiempo de búsqueda y el exceso de ajustes en los algoritmos de aprendizaje.

Este proyecto de título evalúa el rendimiento del algoritmo de los k-vecinos más cercanos (KNN), utilizando para su cálculo la Información Mutua (IM), para seleccionar las características de un conjunto de variables. Este algoritmo se implementó como un software de estudio, el cual tiene distintas opciones configurables para el análisis del rendimiento bajo distintas condiciones.

El algoritmo, con la configuración óptima determinada en este proyecto, tiene un rendimiento promedio de 83,19 % con una desviación estándar de 8,46 %. Con estos datos, se puede determinar que este algoritmo no es suficientemente confiable para ser utilizado en el proceso de selección de variables.



**Palabras Clave:** Información Mutua - KNN - Redundancia - Relevancia - Correlación.



# Abstract

In order to solve the input feature selection problem into a system, to a autoregressive process, seek the most relevant and less redundant features is wanted, helping to reduce the searching time consumption and the setting excess in the learning algorithms.

This degree work evaluates the performance of the k-nearest neighbours algorithm, using Mutual Information to calculate it, in order to select the features of an input data set. This algorithm is developed as a study software, wich has different settable options to analize the performance under certain conditions.

The algorithm, with the optimal configuration defined on this work, has a average performance of 83.19 % with a standard deviation of 8.46 %. With this information, can be determinate that the algorithm is not sufficiently reliable to be used in the feature selection process.



**Keywords:** Mutual Information - KNN - Redundancy - Relevance - Linear Correlation.

# Índice

# Lista de Figuras

# Lista de Tablas

# 1. Introducción

En el mundo de hoy, donde casi todo puede ser medido y analizado, surgen una serie de dificultades. Para analizar los más variados fenómenos de la vida cotidiana, se deben tomar en cuenta todos los factores o variables que los afectan y, de los cuales, siempre es necesario identificar las que son más relevantes para el caso en cuestión.

Ejemplos de fenómenos a analizar hay muchos: las variaciones de la bolsa, las condiciones del oleaje, el rendimiento de un aparato electrónico, la cosecha de alguna fruta o vegetal, el rendimiento de un equipo de fútbol, entre un sin número de casos los cuales podríamos estudiar.

Para cada uno de los ejemplos mencionados hay aún más variables que influyen en sus comportamientos. Tomando el ejemplo de la cosecha de sandías, algunas de las variables que podría afectar son: la estación del año, la calidad de la tierra, los niveles de agua utilizados para regar, el calor, etc. Teniendo n variables identificadas, el siguiente paso sería ver cuáles de estas variables son las más importantes a la hora del análisis del fenómeno, para así, poner un mayor énfasis en éstas y mejorar ese análisis.

Dentro del campo de análisis de datos y más específicamente de las medidas de independencia de las variables aleatorias, se suele utilizar el coeficiente de correlación lineal, el cual muchas veces es útil para resolver ciertos análisis, pero esta medida es inadecuada e ineficiente ante la presencia de relaciones no lineales. También cabe destacar que el coeficiente de correlación lineal ha sido diseñado para trabajar con datos univariados, presentando problemas al introducir datos multivariados.

Por lo anterior, se hace necesario contar con un modelo más general que permita detectar relaciones no-lineales y esté diseñado para trabajar tanto con datos univariados como con datos multivariados. Una medida que cumple con estos requisitos es la Información Mutua (MI), la cual permite detectar tanto relaciones lineales como no lineales. Ésta, en contraste con el coeficiente de correlación lineal, es sensible a las dependencias que no se manifiestan en la covarianza. Se utiliza como una forma de medir la relevancia de cada clasificación o aproximación entre cada característica y la tarea objetivo.

La Información Mutua (MI, en inglés Mutual Information) mide la información que dos variables aleatorias comparten, cuanto el conocimiento de una variable reduce nuestra incertidumbre sobre la otra. Solo si las dos variables aleatorias son estrictamente independientes el valor de su información mutua es 0. Por el contrario, si son idénticas, entonces saber una de ellas determina el valor de la otra.

El uso de la Información Mutua se hace presente en diversos campos y

áreas de investigación, en donde la MI se combina con distintos algoritmos para intentar captar la relación lineal y no-lineal de las variables.

La selección de características (también conocida como selección de variables o selección de atributos) es el problema de seleccionar un subconjunto del conjunto original de características, para probar la calidad de los datos, de acuerdo a la capacidad de discriminación. Ésta ha atraído mucha atención durante las últimas décadas en reconocimiento de patrones y machine learning para el modelado de clasificación. En el reconocimiento de patrones, por ejemplo, los conjuntos de datos que encontremos pueden ser espacios de características de altas dimensiones, en las cuales no todas de éstas características son relevantes y algunas son redundantes. La reducción de dimensiones se necesita acá no solo para probar la eficiencia del clasificador, sino que reduciendo el costo de la adquisición de características en el sistema. Las características innecesarias aumentan el tamaño del espacio de búsqueda y pueden causar un problema de maldición de dimensionalidad, que hace que el consumo de tiempo y sensibilidad de los algoritmos de aprendizaje sufran de exceso de ajuste. Así, la selección de características ha sido propuesta para solucionar ese problema.

Además de los casos antes mencionados, la selección de características más relevantes y menos redundantes, en adelante (MRmR), puede ser utilizada para realizar estudios en los cuales se busquen las variables más importantes que afecten en los cambios sufridos en un caso de estudio particular. Siendo ésta de gran utilidad para ir en ayuda de la toma de decisiones, logrando saber hacia dónde dirigir los esfuerzos para solucionar las aristas importantes del problema en cuestión. En esta ocasión se estudiará un modelo estimando la Información Mutua utilizando el algoritmo de los k-vecinos más cercanos, dónde el objetivo fundamental es identificar un método que permita identificar las variables de mayor relevancia y menor redundancia.

La primera etapa del proyecto presentado consta de la presentación de la investigación realizada de la literatura existente, respectiva a los temas y modelos a utilizar, la planificación del trabajo y la forma de enfrentar la problemática. Luego se presenta un modelo estimando la MI utilizando los KNN, que será utilizado dentro del algoritmo MRmR para realizar una implementación computacional y evaluación numérica usando datos reales y artificiales, probando el rendimiento en modelos autoregresivos de distintos ordenes. Finalmente se evaluán los resultados con complementos gráficos de comparación.

La figura 1 representa el modelo de la función que se desarrollará, donde $X_1$, ..., $X_n$ representa el conjunto de variables independientes de entrada y $X_1$, ..., $X_m$ representan las variables independientes más relevantes y menos

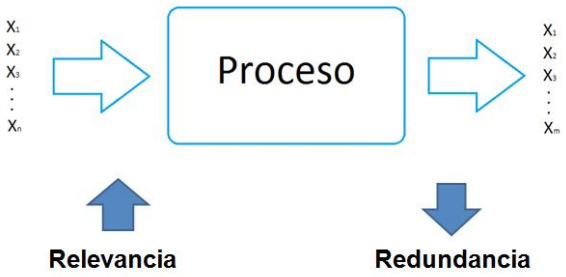


Figura 1: Sistema Propuesto

redundantes que resultan de la selección por medio del modelo propuesto.

# 2. Objetivos

## 2.1. Objetivo General

Evaluar el rendimiento del algoritmo k-vecinos más cercanos, basado en la Información Mutua, para seleccionar las variables endógenas de un modelo autorregresivo lineal.

## 2.2. Objetivos Específicos

- Explicar la estructura y funcionalidad del algoritmo k vecinos.
- Implementar el algoritmo de los k vecinos para selección de variables de un modelo autorregresivo.
- Evaluar el rendimiento del algoritmo k vecinos para procesos autorregresivos lineales, con y sin ruido gaussiano.

# 3. Conceptos Clave

## 3.1. Información Mutua

La información mutua es la medida de la cantidad de información que una variable aleatoria contiene sobre la otra. La información mutua de dos variables aleatorias se define de la siguiente manera:

$$MI(X,N) = \sum_x \sum_y p(x,y) * \log_2 \frac{p_x y(x,y)}{p_x(x) * p_y(y)} \tag{1}$$

Su interpretación es la cantidad de información que ofrecen $X$ e $Y$ en conjunto. El objetivo es obtener la mayor cantidad de información posible de dos variables aleatorias, y esto se conseguirá al maximizar la información mutua.

A partir de los conceptos de entropía conjunta y condicionada, se puede concluir que la información mutua es una medida de la cantidad de la reducción de incertidumbre de una variable $X$ que da el conocimiento de otra $Y$. Esta medida de la información mutua en función de la entropía se puede expresar de la siguiente manera:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

## 3.2. Entropía de Shannon

Se define la entropía de Shannon, $H(X)$, de una variable aleatoria discreta $X$, como la esperanza matemática de la variable aleatoria asociada $I(X)$.

$$H(X) = E(I(X)) = -\sum_{i=1}^{n} p(x_i) * \log_2 p_x \tag{3}$$

La Entropía puede ser considerada como una medida de la incertidumbre y de la información necesaria para, en cualquier proceso, poder acotar, reducir o eliminar la incertidumbre. Resulta que el concepto de información y el de entropía están altamente relacionados entre sí.

## 3.3. K-vecinos más cercanos

El método de los k-vecinos más cercanos (KNN) es usado como método de clasificación de elementos basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos.

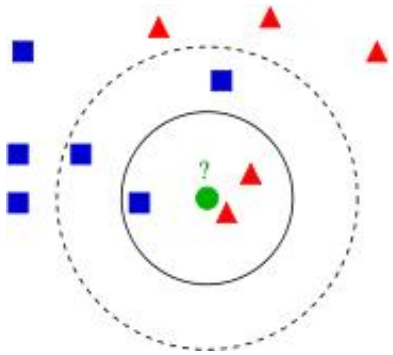


Figura 2: Ejemplo de clasificación usando KNN

- La fase de entrenamiento consiste en almacenar vectores.
- En la fase de clasificación, la evaluación del ejemplo es representada por un vector en el espacio característico.
- Se calcula la distancia entre los vectores almacenados y el nuevo vector y se seleccionan los k ejemplos más cercanos.

Este método supone que los vecinos más cercanos nos dan la mejor clasificación y esto se hace utilizando todos los elementos; el problema de dicha suposición es que es posible que se tengan muchas variables irrelevantes que dominen sobre la clasificación: dos variables relevantes perderían peso entre otras veinte irrelevantes.

El ejemplo que se desea clasificar es el círculo verde. Para $k = 3$ este es clasificado con la clase triángulo, ya que hay solo un cuadrado y 2 triángulos, dentro del círculo que los contiene. Si $k = 5$ este es clasificado con la clase cuadrado, ya que hay 2 triángulos y 3 cuadrados, dentro del circulo externo.

### 3.4. Coeficiente de Correlación de Pearson

El coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente. Esto significa que puede haber variables fuertemente relacionadas, pero no de forma lineal, en cuyo caso el coeficiente de Correlación lineal pierde efectividad. En el caso de que se esté estudiando dos variables aleatorias X e Y sobre una población estadística; el coeficiente de correlación de Pearson se simboliza con la letra

$p_{x,y}$, siendo la expresión que nos permite calcularlo:

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{E((X - \mu_x)(Y - \mu_y))}{\sigma_x \sigma_y} \tag{4}$$

# 4. Estado del Arte

## 4.1. Aplicación de la Información Mutua para la Selección de características Relevantes

### 4.1.1. Selección de Características basado en Información Mutua: Criterios de Máxima-Dependencia, Máxima-Relevancia y Mínima-Redundancia

En esta investigación se estudia cómo seleccionar buenas características, basándose en la máxima dependencia estadística, criterio basado en información mutua. Por la dificultad en implementar directamente la condición de máxima dependencia, primero se deriva una forma equivalente, llamada criterio de mínima-Redundancia-Máxima-Relevancia (mRMR), para selección de características incrementales de primer orden. Luego se presenta un algoritmo de selección de características de dos etapas combinando mRMR y otros selectores de características más sofisticados. Esto permite seleccionar un conjunto compacto de características superiores a un costo muy bajo. Se realiza una comparación experimental extensiva del algoritmo y otros métodos usando 3 clasificadores distintos (naive Bayes, máquina de vectores de soporte y análisis discriminante lineal) y 4 conjuntos de datos diferentes (dígitos escritos a mano, arritmia, NCI líneas celulares de cáncer y tejidos de linfoma). Los resultados confirman que mRMR lleva a una prometedora mejora en la selección de características y la precisión de la clasificación.

### 4.1.2. Selección de variables de entrada para sistemas de recursos hídricos usando una versión modificada del algoritmo mínima-redundancia y máxima-relevancia

La selección de variables de entrada es un paso necesario en el modelado de sistemas de recursos hídricos. Descuidar este paso puede llevar a una innecesaria complejidad del modelo y reducir la precisión del mismo. En esta publicación, se aplica el algoritmo MRMR para identificar los conjuntos de entrada más relevantes en el modelado de un sistema de recursos hídricos. Además se introducen 2 versiones modificadas del algoritmo MRMR (a-MRMR y b-MRMR), donde a y b son factores de corrección que se encuentran para incrementar y disminuir como una función de ley de potencias, respectivamente, con el progreso de los algoritmos de selección de entrada y el incremento del número de variables de entrada seleccionadas. Se aplican los algoritmos propuestos a 22 embalses en California para predecir emisiones diarias basadas en un conjunto de 121 variables de entrada potenciales. Los

resultados indican que los 2 algoritmos propuestos son buenas medidas de las entradas del modelo como se refleja en el desempeño del modelo mejorado. Los valores de a-MRMR y b-MRMR presentan fuerte correlación negativa con el rendimiento del modelo como se muestra en los valores inferiores del error cuadrático medio (RMSE).

### 4.1.3. Mejora del Algoritmo de Selección de Características más Relevantes y Menos redundantes basado en la normalización de la Información Mutua

La selección de características es la técnica de seleccionar un subconjunto de características relevantes, que contiene información para ayudar a distinguir una clase de otras, de un largo número de características extraídas de los datos de entrada. La aplicación de reconocimiento de patrones en el área médica es obviamente de mucho potencial por ejemplo para clasificar imágenes médicas cerebrales, voz patológica, marcha patológica u otros datos biomédicos como la frecuencia cardíaca, presión sanguínea, etc. Sin embargo, como en otras aplicaciones de reconocimiento de patrones, uno de los desafíos más grandes es seleccionar eficientemente un conjunto de característica apropiadas de un largo conjunto de características. Usando un gran número de características no solo se reduce la exactitud de la clasificación debido al ruido y la maldición de la dimensionalidad sino que también aumenta la carga de cálculo del sistema. Por lo tanto, como seleccionar un buen conjunto compacto de características se convierte en una pregunta práctica. Por ahora, varios enfoques para selección de características han sido propuestos. Esos métodos están básicamente fundados en 2 importantes componentes, que son las estrategias de búsqueda y medidas de potencial de características. En este paper, se analizan diferentes medidas de Información Mutua basadas en la bondad del modelo. Algoritmos de búsqueda están más allá del alcance de este paper, por lo tanto, se utiliza una bien conocida estrategia, que se llama Greedy Forward Selection (GFS). El resto del paper está organizado de esta manera. Se presenta el fondo y las limitaciones del trabajo relacionado con la Información Mutua basada en la selección de características. Al final de ésta sección, se propone un método para mejorar las limitaciones indicadas. Luego se presentan los resultados y discusiones. Y su respectiva conclusión.

## 4.2. Aplcaciones de la Información Mutua en otras áreas

### 4.2.1. Extensividad de la Información Mutua en la teoría Relativista

La localización de un estado en una región del espacio en teorías de campos relativistas conlleva a la creación de pares partícula-antipartícula. Estos, impiden la definición de ciertas magnitudes para el sistema localizado, como la entropía, que resultan divergentes. Es posible sustraer contribuciones debidas a las fluctuaciones del vacío que causan estas divergencias considerando magnitudes relacionadas; un ejemplo es la información mutua entre dos regiones.

En este trabajo se muestran los primeros cálculos de entropía para conjuntos alabeados en teorías relativistas; en concreto, se analiza la extensividad de la información mutua entre conjuntos no planeares, para el estado de vacío de un campo de Dirac en 1+1 dimensiones. Se encuentra extensividad de la información mutua en un límite geométrico particular que es relevante para la evaporación de Hawking de agujeros negros.

### 4.2.2. Salud - Epilepsia: fusión de imágenes

En este estudio se analizan distintas alternativas para el procesamiento de imágenes, en los cuales como principal herramienta está la Información Mutua para calcular las dependencias no lineales. Existen numerosos criterios que permiten este recentrado del histograma conjunto y en consecuencia el corregistro de las imágenes. Estos criterios hacen una hipótesis respecto las relaciones entre las intensidades y los voxels de dos imágenes o pretenden optimizar el .ºrdenamiento"del histograma sin a priori. El criterio de los cuadrados mínimos subentiende un orden en torno a la primera diagonal. Pero el criterio de Woods (algoritmo AIR) por ejemplo subentiende una dependencia no lineal entre los valores. El criterio en boga actualmente es el llamado de la información mutua: $IM = H(I) + H(J) - H(I,J)$. $H(I)$ y $H(J)$ son las entropías de cada imagen y $H(I,J)$ es la entropía conjunta de las dos imágenes. Al maximizar este criterio se "pide.el histograma conjunto sin hipótesis a priori sobre las relaciones entre $i$ y $j$. Se reajustan las dos imágenes por una parte favoreciendo la buena cobertura de las imágenes ricas en información gracias a $H(I,J)$ y por otra parte evitando el escollo del recubrimiento parcial gracias al término $H(I) + H(J)$. El poder del corregistro por maximización de la información mutua es que, sin hacer hipótesis a priori sobre las relaciones entre los valores del histograma conjunto, funciona satisfactoriamente incluso en el caso de datos inter modalidades de

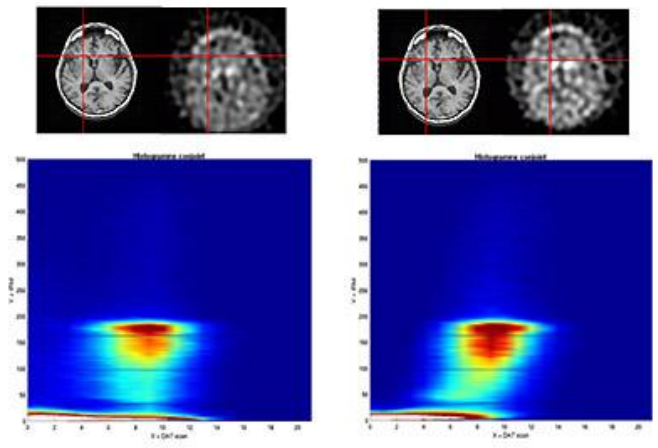

Figura 3: Corregistro SPECT de neurotransmisión y RM por maximización de la información mutua

naturaleza muy diferentes como algunas presentadas anteriormente dónde se puede constatar que se obtiene un corregistro satisfactorio entre, por una parte una RM ponderada en T1 de alta resolución y por otra parte una imagen SPECT de neurotransmisión dopaminérgica en un sujeto que presenta una fijación estriatal muy anormal y una imagen extremadamente ruidosa.

#### 4.2.3. Biología - Aracne: Un algoritmo para la reconstrucción de redes reguladoras de genes en un contexto de celulas de mamíferos

Aracne se presenta como una buena herramienta para la identificación de las interacciones directas de la transcripción en las redes celulares de mamíferos, un problema que ha desafiado a los algoritmos de ingeniería inversa. Este enfoque debería mejorar nuestra capacidad de utilizar los datos de microarrays para dilucidar los mecanismos funcionales que subyacen a los procesos celulares y para identificar objetivos moleculares de compuestos farmacológicos en las redes celulares de mamíferos. El principal porcentaje de error absoluto en la estimación de la MI para densidades normales bivariadas, es comparado con el porcentaje de error en clasificación de valores de la MI relativa para muestras aleatorias pares, para las que la distribución con el menor valor verdadero de MI está entre 70 % y 99 % de la distribución con el valor más alto. La estimación del error de la MI (la línea azul discontinua)

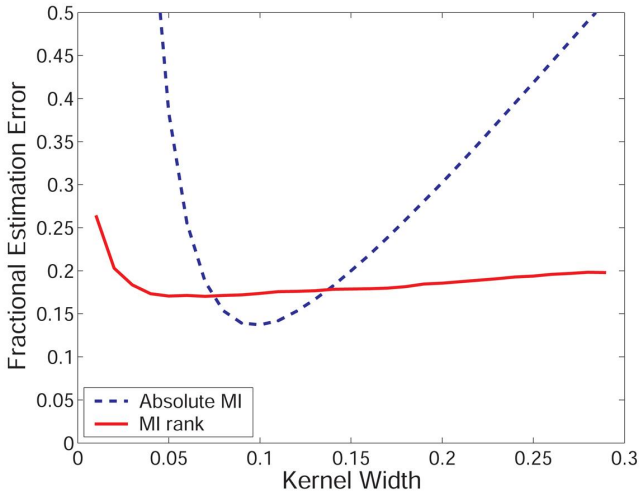


Figura 4: Estimaciones de error de MI y rangos de MI para diferentes anchos del núcleo de Gauss

es altamente sensible a la elección del ancho del núcleo Gausseano usado por el estimador y crece rápido para las elecciones de parámetros no-óptimos. Sin embargo, debido al sesgo similar para distribuciones cercanas a valores de MI, el error en parejas de clasificación de MIs (línea roja continua) es mucho menos sensible a la elección de este parámetro. Estos promedios fueron producidos usando muestras de 1000 densidades normales bivariadas con un coeficiente de correlación aleatorio distribuido aleatoriamente.

#### 4.2.4. Salud - Estudio de la Información Mutua de la Variabilidad del Patrón Respiratorio

Las técnicas tradicionales del análisis de datos en el dominio temporal y en el dominio frecuencial, no son suficientes para caracterizar la compleja dinámica del sistema respiratorio. En el estudio, se analiza la variabilidad del patrón respiratorio usando medidas de la Información Mutua. Estas medidas proporcionan información de las autodependencias estadísticas no lineales de la variabilidad del patrón respiratorio. Se estudiaron señales de un grupo de 20 pacientes sometidos a maniobras de desconexión de la ventilación asistida, para dos niveles de presión de la ventilación soporte. De esta manera se obtuvo señales respiratorias del volumen con diversa variabilidad. Se analizaron las series: duración del ciclo respiratorio, tiempo inspiratorio,

fracción del tiempo inspiratorio, volumen circulante y fracción del flujo inspiratorio medio. Se propuso y estudió siete índices basados en la Información Mutua para caracterizar la variabilidad del patrón respiratorio. Son varias las causas que determinan la variabilidad del patrón respiratorio y las técnicas tradicionales de análisis lineal de datos no son suficientes para caracterizar la dinámica del sistema respiratorio. Así, se hizo necesaria la utilización de técnicas de la dinámica no lineal para el análisis de este complejo sistema fisiológico. La metodología propuesta en el artículo se basa en el estudio de la información mutua a través de la función de autoinformación mutua (AMIF, Auto Mutual Information Function) de las series obtenidas que se mencionan anteriormente.

#### 4.2.5. Análisis de datos - Algoritmo KNN basado en Información Mutua para Clasificación de Patrones con Valores Perdidos

Los datos incompletos son un inconveniente común en los problemas de clasificación de la vida real. Valores perdidos en conjuntos de datos pueden tener distintas causas, tantos como muertes de pacientes, mal funcionamiento de equipos, negativa de encuestados a responder ciertas preguntas, etc. El trabajo presenta un efectivo y robusto enfoque para clasificación con valores de entrada desconocidos. En particular, se propone la presentación de una versión mejorada del algoritmo de los k-vecinos más cercanos usando Información Mutua.

#### 4.2.6. Gobernabilidad -MISSOC- Sistema de Información Mutua sobre protección social

Puesto en marcha en 1990 por la Comisión Europea, el programa MISSOC (Mutual Information System on Social Protection) se ha convertido en una fuente de información privilegiada sobre la situación de la protección social en Europa. Actualmente, participan en MISSOC los 25 Estados miembros de la Unión Europea (UE), los tres Estados del Espacio Económico Europeo, a saber, Islandia, Liechtenstein y Noruega desde el año 2000, así como Suiza desde el año 2002. El programa MISSOC se basa en una estrecha colaboración entre la Dirección General Empleo y asuntos sociales de la Comisión Europea y una red de representantes de los Estados participantes. El programa también cuenta con una Secretaría, designada por la DG ?Empleo y asuntos sociales? de la Comisión, responsable de la coordinación de la red y la preparación de sus publicaciones, de la organización de reuniones, de la recogida de datos, de la gestión del sistema informático y

de la preparación y difusión de publicaciones. El sistema MISSOC se basa en los datos facilitados por los ministerios y las autoridades competentes en materia de protección social. Publica cuadros comparativos actualizados regularmente, que cubren todos los ámbitos de la protección social, así como boletines informativos MISSOC que tratan temas específicos, como la protección social de las personas con discapacidad, la asistencia sanitaria o los sistemas de protección de la vejez o los principales cambios en los sistemas de protección social.

#### 4.2.7. Gobernabilidad - Mecanismo de información mutua para medidas nacionales de asilo e inmigración

El mecanismo de información mutua establece intercambios de información entre la Comisión y los países de la Unión Europea (UE) acerca de los textos nacionales en materia de asilo e inmigración. Por medio de una red Internet y utilizando el formulario de informe que figura en el anexo de la Decisión, los países de la UE transmitirán las medidas que se propongan tomar o hayan tomado recientemente. Esta información deberá transmitirse cuanto antes y, a más tardar, en el momento en que se haga pública.

#### 4.2.8. Biología - Análisis de secuencia de proteínas utilizando motivos información mutua

Secuencia de la proteína motivos son, por definición, breves fragmentos conservados de aminoácidos, a menudo asociados con una función específica. Por consiguiente secuencia de la proteína derivados de múltiples perfiles de alineamientos de secuencias proporcionar una alternativa funcional descripción de los motivos que caracterizan a las familias de secuencias relacionadas. Esos perfiles reflejan convenientemente necesidades funcionales señalando la proximidad de secuencias conservadas en los puestos, así como las distancias a las que representan posiciones variables. Descubriendo la conservación de las características importantes dentro de la variable de los perfiles de puestos de espejos grupo específico y, en particular, las características evolutivas de las secuencias. Se describe la herramienta PROfile análisis basado en la información mutua de Información Mutua (PROMI), que permite el análisis comparativo de los usuarios clasificados secuencias de la proteína. PROMI se ejecuta como un servicio web usando Perl y R así como otras a disposición del público paquetes y herramientas en el lado del servidor. Por el lado del cliente la independencia de la plataforma se logra mediante la entrega de aplicación general a Internet. Como una posible aplicación de los análisis de

dedo de zinc C 2 H 2-tipo de proteína de dominio se introduce para ilustrar la funcionalidad de la herramienta.

#### 4.2.9. Ingeniería Biomédica - Maximización de la Información Mutua para conseguir el registro de imágenes

Aunque la entropía conjunta se ha empleado para el registro de imágenes, la información mutua es una medida más estable, por lo que se ha generalizado su uso para el registro intermodalidad. Incluso en el caso de imágenes de la misma modalidad, se ha propuesto la maximización de la información conjunta como mejor herramienta de registro. También se ha desarrollado mejoras como la información normalizada, que resulta más estable cuando las regiones del paciente presentes en cada estudio no coinciden totalmente, o métodos que incluyen otro tipo de características presentes en las imágenes, como por ejemplo los bordes. Actualmente los algoritmos de registro basados en la maximización de la IM han demostrado claramente su buen funcionamiento para el registro de imágenes PET y SPECT dentro de la misma modalidad y con otras modalidades, incluso para estudios realizados en animales de laboratorio, por lo que este tipo de algoritmos debería ser la elección de preferencia.

# 5. Conceptos de la Información Mutua usando los k-vecinos más cercanos

La IM es una medida no-paramétrica y no-lineal de relevancia que deriva de la teoría de la información. La IM entre dos variables aleatorias $X$ e $Y$ $(I(X,Y))$, es una medida de como $X$ depende de $Y$ y viceversa, y se puede definir a partir del concepto de entropía $H(\Delta)$ :

$$IM(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) \tag{5}$$

donde $H(Y|X)$ es la entropía condicionada de $Y$ dado $X$, y representa la incertidumbre que se tiene sobre Y cuando se conoce el valor de $X$. $I(X,Y)$ mide la reducción en la incertidumbre en $X$ cuando se conoce el valor de $Y$ y viceversa. Si las variables $X$ e $Y$ son independientes, $H(X,Y) = H(X)+H(Y)$, y $H(Y|X) = H(Y)$, la IM entre dos variables independientes es cero. Cuanto mayor sea la medida de IM entre dos variables, mayor será la dependencia existente entre ambas.

Aunque la definición realizada parece sencilla, no siempre resulta así. Normalmente, se cuenta con conjuntos de N mediciones de dos variables, $z_i = (x_i, y_i)$, $i = 1, ..., N$, con densidad $\mu = (x, y)$. De esta manera, $x$ e $y$ pueden ser escalares o elementos de un espacio de mayores dimensiones. En lo que viene, se debe asumir que la densidad es una función fluida adecuada, aunque también podría permitir densidades más singulares. Todo lo se necesita es que las integrales que se definen a continuación existan en algún sentido. En particular, no se debe asumir que las densidades son estrictamente positivas. Las densidades marginales de $X$ e $Y$ son $\mu_x(x) = \int \delta y \mu(x,y)$ y $\mu_y(y) = \int \delta x(x,y)$. La Información Mutua es definida como:

$$I(X,Y) = \int\int \frac{\mu(x,y)\log\mu(x,y)\delta x\delta y}{\mu_x(x)\mu_y(y)} \tag{6}$$

El más sencillo y generalizado avance para estimar la Información Mutua más precisamente, consiste en particionar los soportes de X e Y en ?bins? de tamaño finito, y aproximando la ecuación [6] por la suma finita

$$I(X,Y) = I_{binned}(X,Y) = \sigma_{ij} p(i,j)\log\frac{p(i,j)}{p_x(i)p_y(j)} \tag{7}$$

Mientras algunos estimadores tienen mejores resultados que los estimadores usando tamaños de contenedores específicos, siguen teniendo errores sistemáticos que resultan por un lado aproximando $I(X,Y)$ por medio de $I_{binned}(X,Y)$, y por otro lado aproximando probabilidades por medio

de (algoritmos de) frecuencia de radios. Lo siguiente puede ser presumiblemente minimizado usando correcciones para finitos $n_x(i)$. Estas correcciones son de la forma de series asintóticas que divergen para un finito $N$, pero cuyos primeros 2 términos mejoran la estimación en casos típicos. En esta investigación no seguiremos estas líneas, sino que estimaremos la Información Mutua usando los k-vecinos más cercanos.

Existe una extensa literatura de tales estimadores para la entropía de Shannon simple

$$H(X) = -\int \delta x \mu(x) \log \mu(x) \tag{8}$$

pero según lo estudiado estos métodos nunca han sido utilizados para estimar MI. En muchos casos se asume que x es unidimensional.

Asumiendo algunas métricas que se dan en el espacio abarcado por $X$, $Y$ y $Z = (X, Y)$. Podemos así ordenar, para cada punto $z_i = (x_i, y_i)$ , sus vecinos por distancia $\delta_{i,j} =\| z_i - z_j \|$: $\delta_{i,j1} \leq \delta_{i,j2} \leq \delta_{i,j3}$.... Ordenamientos similares pueden realizarse en el subespacio $X$ e $Y$. La idea básica es estimar la entropía desde las distancias promedio a los k-vecinos más cercanos, promediados sobre $x_i$.

Se presentan a continuación dos algoritmos ligeramente distintos, ambos teniendo como base la idea principal. Para términos prácticos de la investigación actual se utilizará el primero (ecuación 11). Ambos usan para el espacio $Z = (X, Y)$ la norma máxima,

$$\| z - z' \| = max\{\| x - x' \|, \| y - y' \|\} \tag{9}$$

Mientras cualquier norma puede ser usada para $\| x - x' \|$ y $\| y - y' \|$ (no tienen por qué ser iguales, ya que estos espacios pueden ser completamente diferentes). Se denotará $\frac{\epsilon(i)}{2}$ la distancia de $z_i$ con su vecino más cercano, y por $\frac{\epsilon_x(i)}{2}$ y $\frac{\epsilon_y(i)}{2}$ las distancias entre los mismos puntos proyectados en el subespacio de $X$ e$Y$. por tanto, $\epsilon(i)$ se denotará como $\epsilon(i) = max\{\epsilon_x(i), \epsilon_y(i)\}$.

En este primer algoritmo, contamos el número $n_x(i)$ de puntos $x_i$ cuya distancia de $x_i$ es estrictamente menor que $\frac{\epsilon(i)}{2}$, y de la misma forma para $y$ en lugar de $x$. Esto es ilustrado en la figura 5a. Cabe destacar que $\epsilon(i)$ es una variable aleatoria, y por lo tanto también $n_x(i)$ y $n_y(i)$ fluctúan. Denotamos por $\langle ... \rangle$ promedios tanto sobre las $i \in [1, .., N]$ y sobre todas las realizaciones de las muestras aleatorias,

$$\langle ... \rangle = N^{-1} \sum_{i=1}^{N} E[...(i)] \tag{10}$$

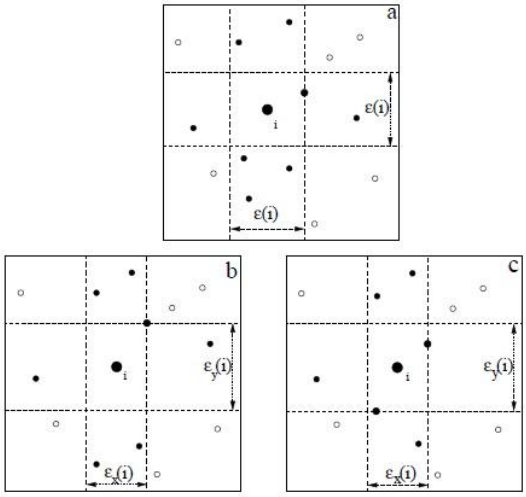


Figura 5: Ejemplos con fórmulas propuestas

El estimador para la Información Mutua es entonces,

$$I^{(1)}(X,Y) = \psi(k) - \langle \psi(n_x+1) + \psi(n_y+1) \rangle + \psi(N) \tag{11}$$

Alternativamente, en el segundo algoritmo, reemplazamos $n_x(i)$ y $n_y(i)$ por el número de puntos con $\parallel x_i - x_j \parallel \leq \frac{\epsilon_x(i)}{2}$ y $\parallel y_i - y_j \parallel \leq \frac{\epsilon_y(i)}{2}$ (figura 5b y c)

El estimador para Información mutua es entonces,

$$I^{(2)}(X,Y) = \psi(k) - \frac{1}{k} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \tag{12}$$

Existen más derivaciones de la ecuación 11 y 12 que son analizadas en la publicación en la cual se basa esta investigación. Allí también se dan fórmulas para redundancias generalizadas en dimensiones más altas,

$$I(X_1, X_2, ..., X_m) = H(X_1) + H(X_2) + ... + H(X_m) - H(X_1, X_2, ..., X_m) \tag{13}$$

En general, ambas fórmulas dan resultados muy similares. Para un mismo k, la ecuación 11 da un poco menos de errores estadísticos (porque $n_x(i)$ y $n_y(i)$ tienden a ser más grandes y tienen un menor fluctuación relativa), pero tiene más errores sistemáticos. El último sólo es fuerte si estamos interesados

en dimensiones muy grandes donde $\epsilon(i)$ tiende a ser mucho más grande que el marginal $\epsilon_{xj}(i)$. En ese caso, el segundo algoritmo parece ser el óptimo. De todas maneras, ambos pueden ser usados igualmente bien.

# 6. Conceptos de máxima dependencia, máxima relevancia y mínima redundancia

Teóricamente, la información mutua entre varias variables aleatorias es equivalente a medir su dependencia; así, determinar el conjunto de variables de entrada con la mayor información mutua es un problema de maximización de la dependencia. Este proceso puede ser implementado usando el algoritmo de la máxima dependencia. Éste algoritmo calcula la dependencia como la información mutua entre un conjunto de variables de entrada ($S_m$) y la variable de salida $Y$ con el método forward, hasta que la adición de cualquier nueva variable se vuelve despreciable:

$$maximize I(S_x, Y), I(X_i, i = 1, ..., m; Y) \tag{14}$$

Esto es cuando ninguna variable de entrada adicional dará una mejora en información mutua ($I$) que sea mayor que un valor de umbral establecido $e$, el algoritmo se detiende. Secuencialmente, el algoritmo encuentra la mejor variable de entrada individual más importante, luego la segunda, tercera, etc., las variables más importantes hasta que la reducción de la incertidumbre de la variable de entrada $Y$ se vuelva menor que $e$. Este algoritmo no es computacionalmente confiable cuando se trata de calcular distibuciones de densidad de probabilidad conjunta de altas dimensiones.

Para reparar esta limitante, una aproximación del algoritmo de máxima dependencia usado comúnmente, es el algoritmo de máxima relevancia ($D$) -también llamado información mutua promedio (AMI)- que se presenta en la ecuación 15. El algoritmo de máxima relevancia trabaja de una manera similar al algoritmo de máxima dependencia. Para cualquier caso, se asume que las variables de entrada $X_i \in S_m, i = 1, ..., m$ son independientes -a menudo una suposición violada cuando la autocorrelación temporal y las variantes espaciales de las variables de entrada hidrológicas son consideradas en el proceso de selección. Este algoritmo no cuenta para cualquier redundancia entre las variables de entrada, por eso, carece de poder discriminar para evitar seleccionar variables de entrada redundantes:

$$maximize D(S_m, Y), D = \frac{1}{m} \sum_{X_i \in S_m} I(X, Y) \tag{15}$$

Para superar las deficiencias de no contemplar la redundancia, Ding y Peng presentan el algoritmo de mínima redundancia ($R$) que trata de encontrar los conjuntos de variables de entrada más mutuamente excluyentes. El algoritmo de mínima redundancia ($R$) por si solo es un pobre estimador de

información mutua porque un conjunto de variables de entrada mutuamente excluyente puede muy bien tener ninguna dependencia con la variable de salida $Y$:

$$maximize R(S_m), R = \frac{1}{m^2} \sum_{X_i X_j \in S_m} I(X_i; X_j) \tag{16}$$

## 6.1. Mínima redundancia, máxima relevancia (MRMR)

Un enfoque más efectivo es combinar los algoritmos de máxima relevancia y mínima redundancia. Peng, propuso la idea de combinar los algoritmo de máxima relevancia y mínima redundancia en un solo problema de maximización. Donde $D$ y $R$ están definidas en las ecuaciones 15 y 16 respectivamente y $\Phi$ es una aproximación del valor real de información mutua $I$:

$$max \Phi(D, R), \Phi = D - R \tag{17}$$

Al método combinado se le denominó algoritmo de mínima redundancia y máxima relevancia (MRMR). En el estudio analizado se demuestra que funciona eficientemente, incluso para conjuntos de entrada relativamente largos y proporciona una prueba analítica que a un modelo de primer orden de MRMR colapsa en un problema de dependencia máxima. Sin embargo, el método es una aproximación, intuitivamente debería dar un mejor algoritmo de selección que otros algoritmos de mínima redundancia y máxima relevancia por si solos. La máxima relevancia lleva el proceso de selección en favor de los conjuntos de variables más relevantes sin atención a la redundancia, mientras que la redundancia mínima favorece a las variables de entrada que son irrelevantes el uno al otro sin ningún tipo de atención a cuán importante son para la variable de salida. Es por esto último la potencialidad que se obtiene al mezclar ambos algoritmos.

## 6.2. Modificación de la mínima redundancia y máxima relevancia

Matemáticamente, la información mutua $I$, entre las variables $X$ e $Y$ es igual a la suma de las entropías marginales de cada variable, $H(X) + H(Y)$, menos su entropía conjunta, $H(X, Y)$. Por lo tanto, el término entropía conjunta es representado por la unión de los dos círculos de Venn. Así mismo, la información mutua entre un conjunto de dos variables aleatorias $(X_1, X_2)$ y una tercera variable aleatoria $(Y)$ puede ser calculada basada en la ecuación 19, que incluye el término de entropía conjunta de 3 dimensiones:

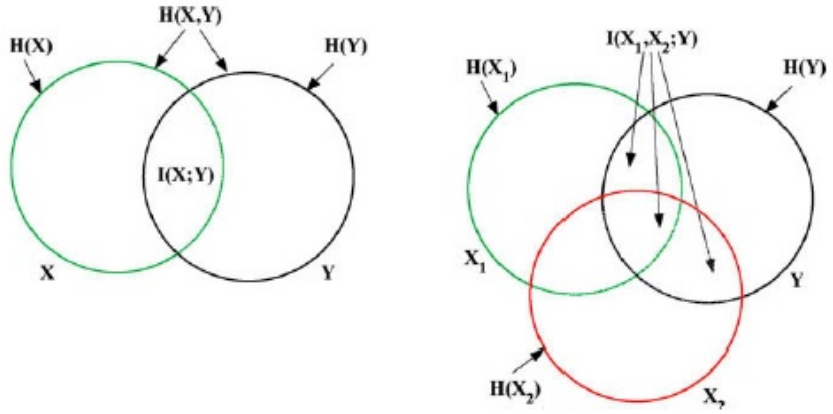


Figura 6: Definición de información mutua en el contexto de los diagramas de Venn

$$I(X_1, Y) = H(X) + H(Y) - H(X, Y) \tag{18}$$

$$I(X_1, X_2; Y) = H(Y) + H(X_1, X_2) - H(X_1, X_2, Y) \tag{19}$$

Se debe notar que el cálculo de $R$ basado en la ecuación 15 (incluyendo autoentropías). Para dejarlo más claro, se supone que $D$ (15) y $R$ (16) son cantidades totales en vez de cantidades promediadas.

Dado que la información mutua es una métrica simétrica ($I(X_i; X_j) = I(X_j; X_i)$), también se excluyen término simétricos ($i >= j$) cuando se calcula $R$. Acá se define $R$ como se indica en la ec 20. El término de redundancia ($R$) en la ec. 15 y 16 es normalizada por el número posible de combinaciones posibles. Por lo tanto, para considerar la autoentropía omitida ($i = j$) y términos simétricos ($i <= j$), el término denominador ($m^2$) se vuelve $2(m^2 - m)$. Esta es la diferencia entre la definición de $MRmR$ utilizada en la investigación Hejazi-Cai y la definición presentada por Peng.

$$minimize R(S_m), R = \frac{1}{2(m^2 - m)} \sum_{X_i X_j \in S_m} I(X_i; X_j) \tag{20}$$

$$R = I(X_1; X_2) + I(X_2; X_1) + I(X_1; X_1) + I(X_2; X_2); R = I(X_1; X_2) \tag{21}$$

| Número de escenario | Información Mutua<br>$I(X_i;Y)$, para $i = 1, ..., 4$ |
|---|---|
| 1 | $I(X_1;Y) = AB$ |
| 2 | $I(X_1, X_2;Y) = AB + AC + ABC$ |
| 3 | $I(X_1, X_2, X_3;Y) =$<br>$AB + AC + AD + ABC$<br>$+ABD + ACD + ABCD$ |
| 4 | $I(X_1, X_2, X_3, X_4;Y) =$<br>$AB + AC + AD + AE$<br>$+ABC + ABD + ABE + ACD+$<br>$ACE + ADE + ABCD + ACDE + ABCDE$ |

Tabla 1: Resumen de las áreas de Venn correspondientes a las medidas de información mutua ($I$) para los escenarios con 1,2,3 y 4 variables de entrada.

| Término | ec. 16 | ec. 20 |
|---|---|---|
| D | $AB + AC + 2ABC$ | $AB + AX + 2ABC$ |
| R | $B + C + AB + AC + 4BC + 4ABC$ | $BC + ABC$ |
| $\Phi$ | $-(B + C + 4BC + 2ABC)$ | $AB + AC + ABC + BC$ |
| I | $AB + AC + ABC$ | $AB + AC + ABC$ |

Tabla 2: Resumen de las areas de Venn correspondientes a las medidas de información mutua ($I$), máxima relevancia ($D$), mínima redundancia ($R$) y mRMR ($\Phi$) para los 3 escenarios de variables aleatorias; $D$, $R$ y $\Phi$ son calculados sin normalizar las cantidades $D$ y $R$.

# 7. Relazación del Software MI-KNN Selector

El Software que nace de la investigación realizada lo hemos llamado MI-KNN Selector, debido a que el núcleo del cálculo del algoritmo de la presente investifación está basado en este algoritmo estudiado.

Luego de haber analizado y elaborado la mixtura de los algoritmos de MI-KNN y MRmR se lo llevó a su implementación en un software, código que está elaborado en MATLAB. En él se ha desarrollado una interfaz de usuario con distintas alternativas, con las cuales el investigador o usuario pueda de una manera sencilla evaluar el rendimiento del algoritmo para la selección de variables relevantes y también poder evaluar datos de algún fenómeno que se quiera estudiar.

## 7.1. Estructura y funcionalidades del Software MI-KNN Selector

Seguiendo el caso de uso general presentado, a continuación se presenta cada uno de los m¸odulos y se detalla cuáles son las funcionalidades disponibles para cada uno de los seis.

### 7.1.1. Cargar Archivo Unitario

Al igual que en el software de Peng, se tiene la funcionalidad de cargar un archivo excel con extensión xls, en este caso, y como alternativa, se tiene la opción de cargar un archivo con una sola variable, vale decir, en este archivo irá almacenada la data de una única variable, a lo largo de un período de tiempo. En la interfaz se solicita el desfase que se le aplicará a la variable, más una variable $k$ que definirá el número de vecinos con el cual se calculará la Información Mutua.

El software obtiene los datos y el archivo ingresado para posteriormente mostrar las gráficas resultantes. En el siguiente capítulo se realizarán pruebas y se mostrarán gráficas de resultado.

### 7.1.2. Cargar Archivo Multivariado

Como segunda alternativa para calcular las variables más relevantes dentro de MI-KNN Selector, se ha desarrollado la opción de cargar un archivo excel con variados set de datos pertenecientes a S número de variables. Para este caso se solicitan sólo 2 datos de entrada: archivo xls y $k$.

Como requerimiento para el usuario se pide que la variable dependiente se ubique en la última columna del archivo.

### 7.1.3. Ingresar Ecuaciones

Para el estudio y análisis investigativo se ha implementado la opción de ingresar ecuaciones autorregresivas.

Un proceso autorregresivo de orden $p$, $AR(p)$, es un proceso estocástico que sigue el modelo

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + a_t \tag{22}$$

Donde $c$,$\phi_1$ y $\phi_p \in R$ son constantes; $a_t$ es un ruido blanco con varianza $\sigma^2$.

En las siguientes 4 interfaces se implementan las funciones Autorregresiva Lineal (AR), Autorregresiva no Lineal (NAR), Autorregresiva Lineal con variables exógenas (ARX) y Autorregresiva no Lineal con variables exógenas (NARX), Permitiendo al usuario ingresar $p$, que determinará el orden de la función autorregresiva, más los factores y un ruido opcional.

Con estas opciones el usuario podrá evaluar y sacar conclusiones respecto a distintos comportamientos según los datos que se ingresen: las variaciones de los resultados según el ruido que sea ingresado, la alteración de los factores, entre otros.

### 7.1.4. Ingresar Ecuación Autorregresiva Lineal (AR)

La primera alternativa de ecuación para ingresar es la Autorregresiva Lineal, de la forma

$$y(t) = \sum_{i=1}^{n} a_i y(n-i) + n(t) \tag{23}$$

En esta interfaz se solicita al usuario un número $n$ que definirá el orden de la función que se ingresará para posteriormente otorgar los factores $a_i$ de la variable autorregresiva para el cálculo. Adicionalmente se puede ingresar un ruido $n(t)$ que afectará el cálculo del algoritmo. Finalmente, se solicita un valor $k$, el cual determinará el número de vecinos que se utilizará para el cálculo de IM.

### 7.1.5. Ingresar Ecuación Autorregresiva Lineal con variables exógenas

Como segunda opción se encuentra la evaluación de una ecuación Autorregresiva Lineal con variable exógena, cuyo modelo es

$$y(t) = \sum_{i=1}^{n} a_i y(n-i) + \sum_{i=1}^{m} b_i \mu(n-i) + n(t) \tag{24}$$

El usuario debe ingresar un valor $n$, el cual define el número de datos que se ingresarán, para posteriormente ingresar los factores $a_i$ y $b_i$ correspondientes. Al igual que todas las opciones de ecuación Autorregresivas se debe ingresar un ruido $n(t)$ y $k$.

### 7.1.6. Ingresar Ecuación Autorregresiva no Lineal (NAR)

Como variante, muchos de los fenómenos que ocurren a diario son del tipo no lineal, por lo que dar soporte y evaluar el rendimiento de éste es escencial. Es por esto que se utiliza IM para el cálculo de MRmR, ya que ella calcula las relaciones lineales y no lineales entre las variables.

El modelo para este caso es:

$$y(t) = \sum_{i=1}^{n} a_{1i} y(n-i) + \sum_{i=1}^{n} \sum_{j=1}^{i} a_{2i} y(n-i) y(n-j) + n(t) \qquad (25)$$

En esta ocasión además de $n$ y los factores asociados, se deben ingresar los factores $a_{2ij}$ de la matriz para calcular la parte no lineal del modelo. De igual manera, como en las otras funcionalidades, se deben ingresar $k$ y el ruido $n(t)$.

### 7.1.7. Ingresar Ecuación Autorregresiva no Lineal con variables exógenas (NARX)

La ecuación Autorregresiva no lineal con variable exógena es similar al caso anterior, pero, esta vez, como su nombre lo indica, además de la variable endógena se incorpora una variable independiente. El modelo para este proceso es

$$y(t) = ec.25 + \sum_{i=1}^{m} b_{1i} \mu(n-i) + \sum_{i=1}^{n} \sum_{j=1}^{i} b_{2ij} \mu(n-i) \mu(n-j) \qquad (26)$$

En esta última opción, al igual que en NAR, se deben ingresar, además de los factores $a_1$ y $a_{2ij}$, los factores $b_1$ y $b_{2ij}$ para el cálculo de la parte no lineal. Una vez más se debe ingresar el valor $k$ y el ruido $n(t)$.

# 8. Casos de Estudio

## 8.1. Evaluación y prueba de carga de archivo

En la primera prueba se evaluará la función por ingreso de archivos Excel, con datos generados a partir de 5 funciones autorregresivas lineales de distinto orden, las 2 últimas con cambio de fase. A todas se les aplica un desfase detallado en una tabla y utilizando con un $k = 3$, por tanto se utilizarán los 3 vecinos más cercanos para el cálculo de la Información Mutua.

Los modelos calculados son:

(1) AR1

$$X_t = 0{,}9 * X_{t-1} + 0{,}866 * \epsilon_t$$

(2) AR4

$$X_t = 0{,}6 * X_{t-1} - 0{,}4 * X_{t-4} + \epsilon_t$$

(3) AR9

$$X_t = 0{,}3 * X_{t-1} - 0{,}6 * X_{t-4} - 0{,}5 * X_{t-9} + \epsilon_t$$

(4) TAR1 - umbral autorregresivo de orden 1

$$X_t = \begin{cases} -0{,}9 * X_{t-3} + 0{,}1 * \epsilon_t & si X_{t-6} <= 0 \\ 0{,}4 * X_{t-3} + 0{,}1 * \epsilon_t & si X_{t-6} > 0 \end{cases}$$

(5) TAR2 - umbral autorregresivo de orden 2

$$X_t = \begin{cases} -0{,}5 * X_{t-6} + 0{,}5 * X_{t-10} + 0{,}1 * \epsilon_t & si X_{t-6} <= 0 \\ 0{,}8 * X_{t-10} + 0{,}1 * \epsilon_t & si X_{t-6} > 0 \end{cases}$$

Para este caso se han tomado desfases aleatorios mayores al orden de cada uno de los modelos. En la tabla 3 se puede observar que en el caso de los modelos (1), (2), (4) y (5) los resultados presentan niveles altos de efectividad, especialmente para los modelos (2) y (4), en los que la efectividad del algoritmo es de un 100 %. No obstante, el cálculo general de efectividad es de 83 %.

## 8.2. Evaluación y prueba de Ingreso de función AR Lineal

Para evaluar el rendimiento de este módulo se utilizarán los modelos AR (1), AR (4) y AR (9). La configuración de las pruebas y los resultados se encuentran detallados en la tabla 4. Cada fila de resultados corresponde al promedio de pruebas realizadas con 6 ondas de ruido aleatorias manteniendo la misma intensidad.

| | VP | VN | FP | FN | Desfase | Efectividad |
|---|---|---|---|---|---|---|
| Modelo (1) | 1 | 3 | 1 | 0 | 5 | 80 % |
| Modelo (2) | 2 | 3 | 0 | 0 | 5 | 100 % |
| Modelo (3) | 2 | 3 | 3 | 1 | 9 | 56 % |
| Modelo (4) | 2 | 5 | 0 | 0 | 7 | 100 % |
| Modelo (5) | 2 | 7 | 1 | 0 | 10 | 90 % |
| Totales | 9 | 21 | 5 | 1 | 36 | 83 % |

Tabla 3: Resultados del cálculo de variables ingresadas por excel

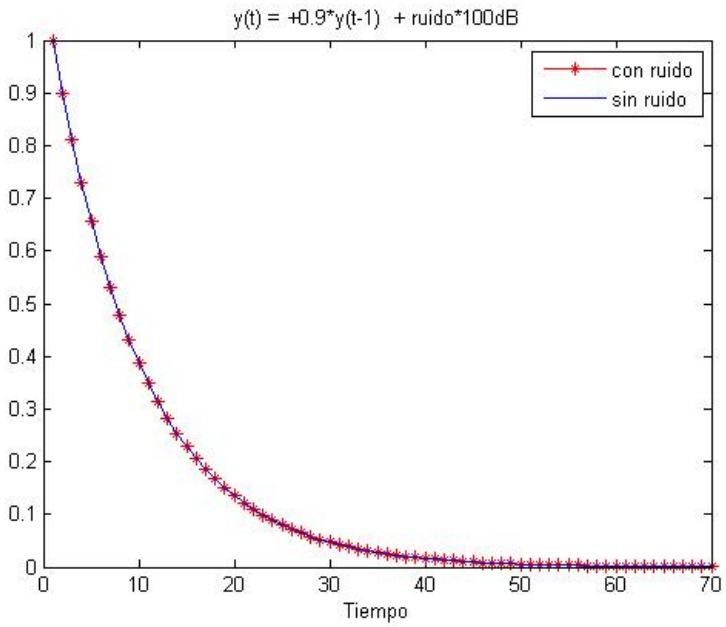


Figura 7: Modelo 1 con -100 dB de ruido

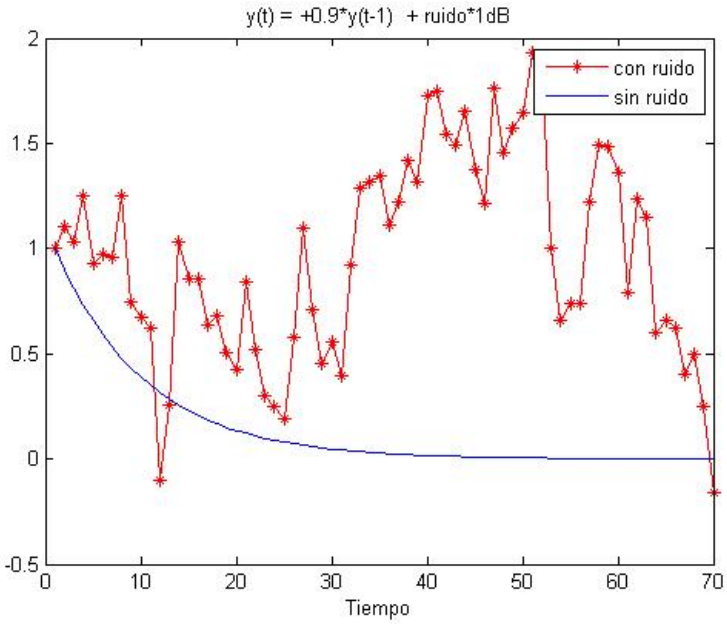


Figura 8: Modelo 1 con -1dB de ruido

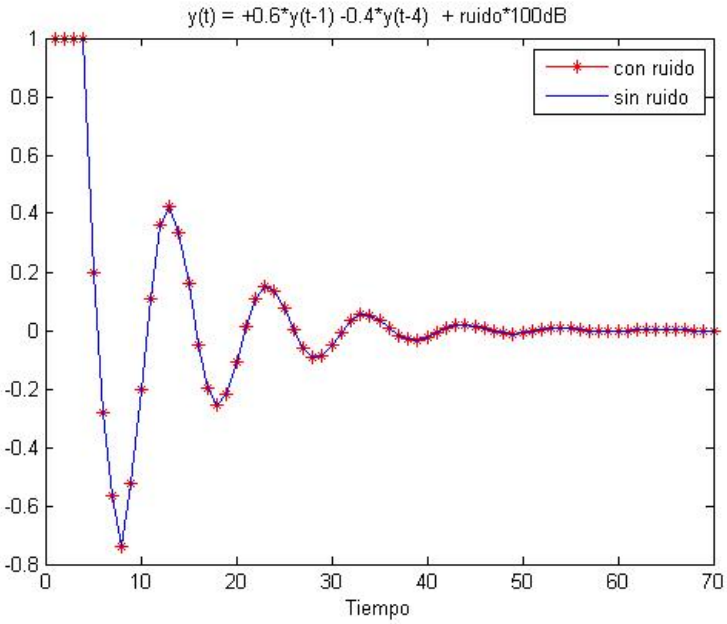


Figura 9: Modelo 2 con -100dB de ruido

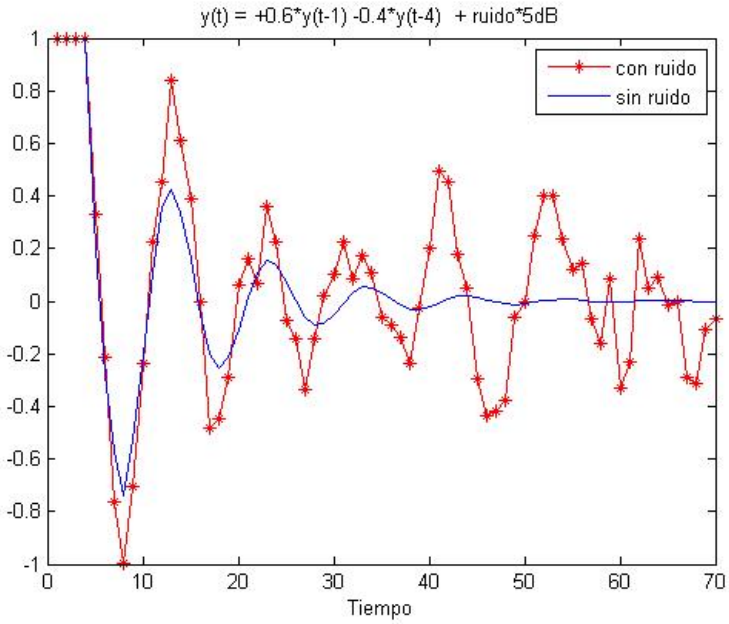


Figura 10: Modelo 2 con -5dB de ruido

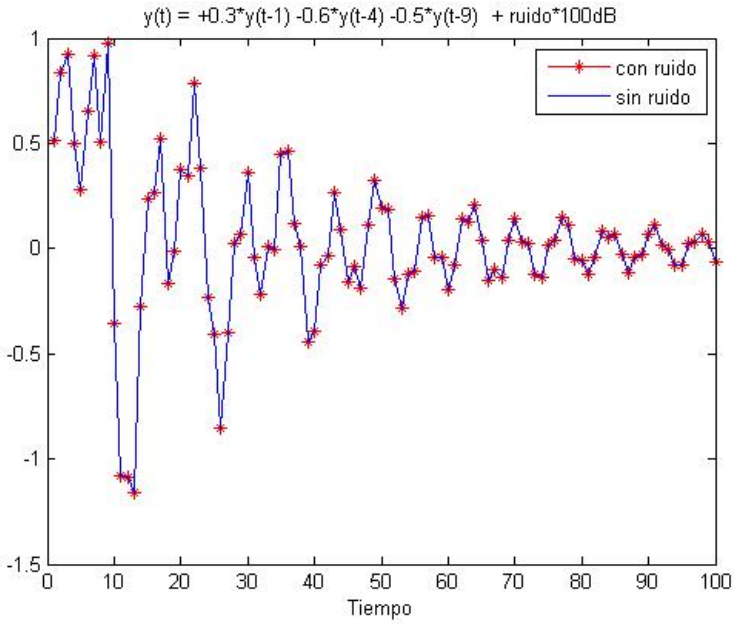


Figura 11: Modelo 3 con -100dB de ruido

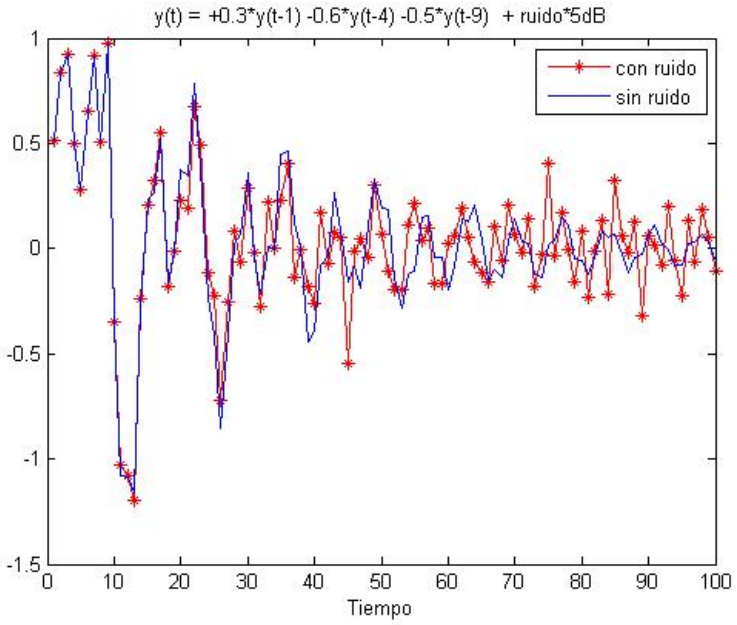


Figura 12: Modelo 3 con -5dB de ruido

| M | D | T | U | dB | K | VP | VN | FP | FN | E |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 70 | 60 | 100 | 2 | 1 | 4 | 0 | 0 | 100 % |
| 1 | 5 | 70 | 60 | 10 | 2 | 1 | 3,8 | 0.2 | 0 | 96 % |
| 1 | 5 | 70 | 60 | 10 | 5 | 1 | 4 | 0 | 0 | 100 % |
| 1 | 5 | 70 | 60 | 5 | 5 | 1 | 4 | 0 | 0 | 100 % |
| 1 | 5 | 70 | 60 | 2 | 5 | 1 | 4 | 0 | 0 | 100 % |
| 1 | 5 | 70 | 60 | 1 | 5 | 1 | 4 | 0 | 0 | 100 % |
| 2 | 4 | 70 | 60 | 100 | 2 | 2 | 2 | 0 | 0 | 100 % |
| 2 | 8 | 70 | 60 | 100 | 2 | 2 | 6 | 0 | 0 | 100 % |
| 2 | 4 | 70 | 60 | 20 | 2 | 2 | 2 | 0 | 0 | 100 % |
| 2 | 8 | 70 | 60 | 20 | 2 | 2 | 6 | 0 | 0 | 100 % |
| 2 | 8 | 70 | 60 | 20 | 5 | 1 | 6 | 0 | 1 | 88 % |
| 2 | 8 | 70 | 70 | 20 | 5 | 2 | 6 | 0 | 1 | 100 % |
| 2 | 8 | 70 | 70 | 50 | 5 | 1,6 | 5,8 | 0,2 | 0,4 | 93 % |
| 2 | 4 | 70 | 60 | 10 | 2 | 0 | 2 | 0 | 2 | 50 % |
| 2 | 4 | 70 | 60 | 10 | 5 | 1 | 2 | 0 | 1 | 75 % |
| 2 | 8 | 70 | 70 | 10 | 5 | 1 | 6 | 0 | 1 | 88 % |
| 2 | 8 | 70 | 70 | 10 | 5 | 1,86 | 5,86 | 0,14 | 0,14 | 96 % |
| 2 | 4 | 70 | 70 | 10 | 5 | 1,4 | 2 | 0 | 0,6 | 85 % |
| 2 | 4 | 70 | 60 | 5 | 2 | 0,1 | 2 | 0 | 1,9 | 53 % |
| 2 | 4 | 70 | 60 | 5 | 5 | 1 | 2 | 0 | 1 | 75 % |
| 2 | 4 | 70 | 70 | 5 | 5 | 1 | 2 | 0 | 1 | 75 % |
| 2 | 8 | 70 | 60 | 5 | 2 | 0,17 | 5,83 | 0,5 | 1,5 | 75 % |
| 2 | 8 | 70 | 60 | 5 | 5 | 1 | 6 | 0 | 1 | 88 % |
| 2 | 8 | 70 | 70 | 5 | 5 | 0,88 | 5,5 | 0,5 | 1,13 | 80 % |
| 3 | 9 | 100 | 60 | 100 | 2 | 1 | 6 | 0 | 2 | 78 % |
| 3 | 9 | 100 | 60 | 100 | 5 | 2 | 4 | 2 | 1 | 67 % |
| 3 | 9 | 100 | 70 | 100 | 2 | 0 | 6 | 0 | 3 | 67 % |
| 3 | 9 | 100 | 70 | 100 | 5 | 1 | 6 | 0 | 2 | 78 % |
| 3 | 9 | 100 | 60 | 20 | 2 | 0,33 | 6 | 0 | 2,67 | 70 % |
| 3 | 9 | 100 | 60 | 20 | 5 | 1,83 | 4,17 | 1,83 | 1,17 | 67 % |
| 3 | 9 | 100 | 70 | 20 | 2 | 0 | 6 | 0 | 3 | 67 % |
| 3 | 9 | 100 | 70 | 20 | 5 | 1 | 6 | 0 | 2 | 78 % |
| 3 | 9 | 100 | 60 | 10 | 2 | 0 | 6 | 0 | 3 | 67 % |
| 3 | 9 | 100 | 60 | 10 | 5 | 1,5 | 4,83 | 1,17 | 1,5 | 70 % |
| 3 | 9 | 100 | 70 | 10 | 2 | 0 | 6 | 0 | 3 | 67 % |
| 3 | 9 | 100 | 70 | 10 | 5 | 0,5 | 6 | 0 | 2,5 | 72 % |
| 3 | 9 | 100 | 60 | 5 | 2 | 0 | 6 | 0 | 3 | 67 % |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 9 | 100 | 60 | 5 | 5 | 1,33 | 5,33 | 0,67 | 1,67 | 74 % |
| 3 | 9 | 100 | 70 | 5 | 2 | 0 | 6 | 0 | 3 | 67 % |
| 3 | 9 | 100 | 70 | 5 | 5 | 0,5 | 6 | 0 | 2,5 | 72 % |
| 3 | 18 | 100 | 60 | 100 | 2 | 0 | 15 | 1 | 2 | 83 % |
| 3 | 18 | 100 | 60 | 100 | 5 | 2 | 13 | 2 | 1 | 83 % |
| 3 | 18 | 100 | 70 | 100 | 2 | 0 | 15 | 0 | 3 | 83 % |
| 3 | 18 | 100 | 70 | 100 | 5 | 1 | 14 | 1 | 2 | 83 % |
| 3 | 18 | 100 | 60 | 20 | 2 | 0 | 14 | 1,5 | 2,5 | 78 % |
| 3 | 18 | 100 | 60 | 20 | 5 | 2 | 13 | 2 | 1 | 83 % |
| 3 | 18 | 100 | 70 | 20 | 2 | 0 | 15 | 0 | 3 | 83 % |
| 3 | 18 | 100 | 70 | 20 | 5 | 1 | 14 | 1 | 2 | 83 % |
| 3 | 18 | 100 | 60 | 10 | 2 | 0 | 15 | 0,33 | 2,67 | 83 % |
| 3 | 18 | 100 | 60 | 10 | 5 | 1,67 | 13 | 2 | 1,33 | 81 % |
| 3 | 18 | 100 | 70 | 10 | 2 | 0 | 15 | 0 | 3 | 83 % |
| 3 | 18 | 100 | 70 | 10 | 5 | 1,17 | 14,17 | 0,83 | 1,83 | 85 % |
| 3 | 18 | 100 | 60 | 5 | 2 | 0 | 15 | 0,17 | 2,83 | 83 % |
| 3 | 18 | 100 | 60 | 5 | 5 | 1,5 | 13,83 | 1,17 | 1,5 | 85 % |
| 3 | 18 | 100 | 70 | 5 | 2 | 0 | 15 | 0 | 3 | 83 % |
| 3 | 18 | 100 | 70 | 5 | 5 | 0,5 | 13,67 | 0,5 | 2,33 | 84 % |

Tabla 4: Resultados del cálculo de variables ingresadas por medio de ecuaciones autorregresivas.

# 9. Conslusiones

En este trabajo de título se ha evaluado el rendimiento de la IM utilizando el algoritmo de los k-vecinos más cercanos, cuyos resultados son sometidos a un proceso de maximización de la relevancia y minimización de la redundancia. Para realizar esta evaluación fue necesario implementar dos procesos: un proceso para generar variables de entrada y un proceso para seleccionar variables. Ambos procesos fueron desarrollados de manera independiente, de forma tal que no existiera relación alguna que pudiera afectar los resultados.

En primera instancia se evaluó la función por ingreso de archivos Excel, con datos generados a partir de 5 funciones autorregresivas lineales de distinto orden, de los cuales 2 presentan cambios de fase. Los resultados presentan niveles altos de efectividad, especialmente para los modelos AR4 (2) y TAR1 (4), en los que la efectividad del algoritmo es de un 100 %, Sin embargo, el promedio general es de un 83 %.

El modelo que baja la efectividad, para estas pruebas, es el modelo AR9 (3). Este resultado se da ya que los tiempos t-9 y t-4, son más determinantes que el tiempo t-1, por lo cual, este se ve relegado a un lugar secundario para el algoritmo, ya que los tiempos derivados de t-1 son más determinantes que este mismo. Es por eso que la efectividad se ve relegada a un 56 %.

En la evaluación mediante ingreso de funciones se utilizaron las funciones AR1 (1), AR4 (2) y AR9 (3). En el primer modelo el rendimiento es bastante parejo y cercano al 100 %; al modificar el desfase, umbral, ruido y el k. Existe una baja particularmente cuando se aplica un ruido de -10 db y un k=2. Este vuelve a ser 100 % efectivo al mantener el ruido pero aumentando el k a 5.

Para el segundo modelo, de orden 4, el rendimiento sin ruido alcanza un rendimiento muy alto, que comienza a bajar al ir aplicando ruidos de mayor magnitud. Sin embargo, al igual que con el primer modelo, al aumentar el valor de k de 2 a 5 y manteniendo el ruido en -10 y -5 db, se nota una clara mejora en el rendimiento. Al mover el umbral de un 60 a un 70 % también se ven variaciones en el rendimiento, que varían positivamente cuando se trabaja con un k=5 y que varía negativamente en menor magnitud con un k=2.

Para el tercer modelo AR9 (3), de orden 9, las tendencias son similares a los modelos anteriores, aunque la diferencia de rendimiento no es tan marcada al aplicar mayores niveles de ruido, debido a que el ruido no distorsiona mayormente la forma de la curva. Si se repite de los modelos anteriores, la mejora al aumentar el valor de k a 5. Al aumentar el desfase al doble, 18, el rendimiento sube considerablemente y se mantiene a lo largo de las pruebas, en las que varían: el ruido, umbral y k. Esto se debe principalmente a que el

algoritmo identifica los verdaderos negativos, pero respecto a los verdaderos positivos, donde debería seleccionar 3 de 18 el porcentaje no es muy alto. A nivel general el algoritmo presentó un 82 % de rendimiento promedio y de un 12 % de desviación estándar.

Como tendencia general, al aumentar el desfase para cada función, el rendimiento aumenta, esto marcado por la presencia y descarte de los verdaderos negativos, para la cual el algoritmo tiene una alta tasa de identificación. La variación en el umbral afecta según el número de k con el cual se ejecute la prueba, ya que se identificó que al trabajar con un k=2 y aumentar el umbral, el rendimiento empeora; por el contrario con un k=5 se identifica una clara mejora.

La aplicación de ruido bajó el rendimiento en todos los modelos a medida que se aumentaba, con excepción del tercer modelo, en el cual, el ruido no hace variar drásticamente la curva y es por esto que los resultados con y sin ruido son similares.

Respecto a la variación en $k$, se encontró como óptimo el valor 5. Se percibe una diferencia marcada en las pruebas al trabajar con un k de 2 y 5, y a la vez que aumentar el valor también empeora los resultados. Esto se debe a que la información con la cual se trabaja al usar un k=5 es mayor, pero al aumentar este valor, esa mayor información puede distorcionar la selección de variables bajo el umbral definido, por lo cual, el umbral y el k están directamente relacionados, mas, esta relación no puede ser definida a priori, sino que debe definirse por heurística.

Teniendo en cuenta lo anterior, desde los datos obtenidos en el estudio se puede extraer un subconjunto $S$ que contenga todos los cálculos realizados con $k = 5$. Teniendo en cuenta sólo los datos de este subconjunto, el rendimiento del algoritmo alcanza un rendimiento de 83.49 % promedio con una desviación estándar de 10.08 %. Incluso, dado que la variable $k$ se encuentra directamente relacionada con el umbral seleccionado, es posible definir dos subconjuntos de $S$, el subconjunto $S_1$ que contenga los resultados obtenidos con un umbral de 60 % y el subconjunto $S_2$ que contenga los resultados obtenidos con un umbral de 70 %.

Al evaluar el rendimiento sobre el subconjunto $S_1$, el algoritmo alcanza un rendimiento de 83,19 % promedio con una desviación estándar de 8,46 %. Y, al evaluar el rendimiento sobre el subconjunto $S_2$, el algoritmo alcanza un rendimiento de 83,74 % promedio con una desviación estándar de 11,5 %.

Si bien, el rendimiento promedio obtenido en el subconjunto $S_2$ es mayor, también lo es su desviación estándar, lo cual disminuye su confianza, por lo cual, luego de realizar este estudio, se puede concluir que el mejor rendimiento se alcanza sobre el subconjunto $S_1$, ya que, a pesar de que su rendimiento

promedio (83,19 %) es menor al del subconjunto $S_2$ su desviación estándar es de sólo un 8,46 %, lo cual aumenta la confiabilidad del algoritmo.

Por tanto, la mejor configuración observada es con la variable $k = 5$ con un umbral de 70 %.

## Referencias


[1] C.A. Cardona M., J.D. Velásquez H., *Selección de Características Relevantes Usando Información Mutua.* Dyna (Medellín, Colombia), vol. 73, n. 149, pp. 149-163, Julio 2006.

[2] P. Carrión, J. Ródenas, J.J. Rieta, *Ingeniería Biomédica, Imágenes Médicas.* Universidad de Castilla-La Mancha, España, pp. 195, 2006.

[3] J. Darcourt, P. Koulibaly, O. Migneco. *Exploración del sistema nervioso central: estado del arte en metodología.* Alasbimn Journal, año 8, n. 30, Artículo N° AJ30-2, Octubre 2005.

[4] F.J. García, *Estudio Comparativo de Métodos de Interpolación para el Cálculo de la Información Mutua en Registro de Imágenes Médicas*, Escuela Técnica Superior de Ingeniería de Telecomunicación, Universidad Politécnica de Cartagena, Septiembre 2008.

[5] P. J. García-Laencina, R. Verdú-Monedero, J. Larrey-Ruiz, J. Morales-Sánchez, JL. Sancho-Gómez, *Algoritmo KNN basado en Información Mutua para clasificación de Patrones con Valores Perdidos.* Simposium Nacional de la Unión Científica Internacional de Radio, Septiembre 2008, Madrid.

[6] Z. GuoDong, Y. LingPeng, S. Jian, J. DongHong, *Mutual Information Independence Model using Kernel Density Estimation for Segmenting and Labeling Sequential Data*, Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing 2005, Cuidad de Méjico, Springer, vol. 3406, pp. 155-166, Febrero 2005

[7] Mohamad I. Hejazi, Ximing Cai, *Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm*, ELSEVIER, vol. 32, n. 4, pp. 582-593, Abril 2009.

[8] E. Hummel, N. Keshvari, W. Weckwerth, J. Selbig, *Species-specific analysis of protein sequence motfits using mutual information*, BMC Bioinformatics, vol. 6, n. 1, p. 164, Junio 2005

[9] A. Kraskov, H. Stögbauer y P. Glassberger, *Estimating Mutual Information*, John-von-Newman Institute for Computing, Forschungszentrum Jülich, pp. 1-3, Febrero 2003.

[10] A. A Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, A. Califano, *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.* BMC Bioinformatics, vol. 7, n. Suppl 1, p. S7, Marzo 2006.

[11] H. Peng, *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-relevance, and Min-Redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, n. 8, pp. 1226-1238, Agosto 2005

[12] O. Tibaduisa, M. Vallverdú, B. Giraldo, P. Caminal. *Estudio de la Información Mutua de la Variabilidad del Patrón Respiratorio*, XXV Jornadas de Automática, Ciudad Real, Septiembre 2004.

[13] Unión Europea, *MISSOC-Sistema de información mutua sobre protección social.* Disponible vía web en http://europa.eu/legislation_summaries/employment_and_social_policy/social_protection/c10606_es.htm. Revisada por última vez el 14 de noviembre de 2014.