UNIVERSIDAD MICHOACANA DE SAN
NICOLAS DE HIDALGO

FACULTAD DE INGENIERIA ELECTRICA

PREDICCIÓN FOTOVOLTAICA USANDO EL MODELO ARIMA Y LOS MÉTODOS DE APRENDIZAJE DE MÁQUINA: K-NN Y MSV.

TESIS

Que para obtener el grado de
LICENCIATURA EN INGENIERÍA ELÉCTRICA.

presenta

Jesús Manuel Gómez Baños.

Dr. Alejandro Zamora Méndez.
Director de Tesis

M.C. José Ortiz Bejar.
Co-Director de Tesis

Julio 2019

Resumen

Este trabajo de tesis, presenta la predicción de distintas series de tiempo relacionadas con la generación fotovoltaica, tomadas de un sistema de paneles interconectados a la red e instalados en el Laboratorio de Potencia de la Division de estudios de Posgrado de la Facultad de Ingenierá Eléctrica, de la Universidad Michoacana De San Nicolás De Hidalgo (UMSNH), ubicada en la ciudad de Morelia, en el estado de Michoacán. Se tomaron datos de voltaje generado y temperatura de la página web del fabricante de los inversores. También se tomaron en cuenta los datos de radiación solar de la estación meteorológica ubicada ahí mismo, más específico en el techo del laboratorio. Con estas series de tiempo, se realiza su predicción usando la MSV y K-NN, además de el modelo ARIMA. Estos métodos fueron ejecutados utilizando funciones de Python y Matlab. Para las primeras pruebas, se usa la serie de tiempo de temperatura, para lo cual se utilizaron diversas configuraciones de estos métodos basados en aprendizaje de máquina y del modelo ARIMA, tales como fueron distintos tipos de funciones Kernel para el método MSV, valores diferentes de K para el método K-NN y diferentes valores para p, d, y q en el modelo ARIMA; esto con el objetivo de saber cuál es la mejor configuración de cada uno de los métodos propuestos para los casos de estudios presentados, los cuales son series de tiempo implicadas en la generación fotovoltaica.

Se hicieron varias pruebas con las mejores configuraciones de cada uno de los métodos propuestos, con el objetivo de determinar cuál de estos tres es el mejor trabajando con las series de tiempo relacionadas con la generación de fotovoltaica. Además, se aplica el teorema de Takens en conjunto con el método de aprendizaje de máquina que mejor resultados arroja, con la finalidad de mejorar la predicción. En todas las pruebas se usaron márgenes de predicción de 1h, 2hr, y hasta de un día.

Palabras claves: Series, Tiempo, Celdas, Solares, Granjas

Abstract

In this thesis, the forecasting of different kind of time series related to the photovoltaic generation is presented. These signals are taken from an interconnected panel system to the grid located at the postgraduate of electrical engineering of the Universidad Michoacana De San Nicolas de Hidalgo (UMSNH), Morelia, Michoacan, Mexico. Several time series for forecasting are obtained from the inverters connected to the solar panels via website. Mainly, two time series are taking into account: voltage generated and temperature. Also, from a meteorological station located in the same place, the solar radiation time series is used for forecasting. In this research, the Autoregressive Integrated Moving Average (ARIMA), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) methods are presented for forecasting the different time series. These methods are implemented in Phython and Matlab. For the ARIMA model, different values for p, d, and q are used. Also, for the SVM method different kind of Kernel are tested, and for the K-NN method, different values of K are used. These variations in their characteristics are implemented with the objective to test which one is the best for forecasting of the time series presented in this thesis.

For the best configurations of the propose methods, several tests are maded with the objective to determined which one is the best to forecasting the time series related with the photovoltaic generation. Also, Takens' theorem is used to improve the forecasting with the K-NN method. The margins of the tests for forecasting are 1 hr, 2 hr, and 1 day.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Símbolos

$\theta_i =$	Coeficientes de los modelos.
$\mu =$	Media aritmética.
$s =$	Desviación estándar de una muestra.
$\sigma =$	Desviación estándar de una población.
$s^2 =$	Varianza de una muestra.
$\sigma^2 =$	Varianza de una población.
$N =$	Número de muestras
$\epsilon_t =$	Ruido blanco o error.
$yt =$	Serie bajo investigación.
$\Delta^d y_t =$	Expresa que sobre la serie original yt, se han aplicado d diferencias.
$P =$	Orden del modelo auto-rregresivo.
$d =$	Número de diferencias aplicadas.
$q =$	Orden del modelo medias moviles.
$rk =$	Coefiente de autocorrelación.
$r =$	Coeficiente de correlación de Karl Pearson.
$cov =$	Covarianza.
$var =$	Varianza.
$SXY =$	Covarianza de dos variables.
$I_{SC} =$	Corriente de corto circuito.
$V_{OC} =$	Voltaje de ciruito abierto.
$P_L =$	Potencia suministrada por la celda.
$V_L =$	Voltaje suministrado por la celda.
$I_L =$	Corriente suministrada por la celda.
$I_{SC}(G) =$	Intensidad de cortocircuito para una irradiación G.
$I_{SC}(CEM) =$	Intensidad de cortocircuito en condiciones CEM.
$G =$	Irrandiancia (W/m^2).
$CEM =$	Condiciones estándar de medida.
$T_c =$	Temperatura de trabajo de la celda $(°C)$.
$T_a =$	Temperatura ambiente $(°C)$.
$TONC =$	Temperatura de operación nominal de la celda $(°C)$.

Capítulo 1

Introducción

En la actualidad, la demanda de energía eléctrica es muy grande por lo tanto se a vuelto importante usar [17] fuentes de generación más amigables con el medio ambiente, como por ejemplo, la generación fotovoltaica, la cual ah tenido una tendencia de crecimiento muy acelerada [18]. Estos sistemas fotovoltaicos cada vez son más comunes de ver, ya sea interconectados a la red o sistemas autónomos. Por lo tanto, es necesario conocer las diferentes partes de estos sistemas. Una parte importante, es conocer de alguna manera, con cuánta energía vamos a contar en un futuro. Esto nos orilla a buscar métodos efectivos de predicción a futuro. Con ayuda de las series de tiempo, se puede ver el comportamiento de las variables de interés. En la actualidad, estas series de tiempo se ocupan en distintas áreas como pueden ser: economía, medicina, física, electricidad, etc., por lo que se pueden utilizar en modelos probabilísticos tales como el AR, ARMA, ARIMA, etc. En este trabajo de tesis, se propone el modelo ARIMA(p,d,q), así como dos métodos basados en aprendizaje de máquina los cuales son, la MSV y K-NN, se hará predicción con series de tiempo relacionadas a la generación fotovoltaica, usando diferentes horizontes de tiempo, tales como minutos, horas y días, con el fin de conocer cual de estos 3 es mejor. Para esto, se va a tomar como objeto de estudio las series de tiempo proporcionadas por la página web del fabricante de los inversores conectados a un sistema fotovoltaico interconectado a la red, el cual está instalado en el laboratorio de la división de estudios de posgrado de la facultad de ingeniería eléctrica de la UMSNH, ubicada en la ciudad de Morelia, Michoacán, México, así como la

medición de la radiación solar provista por la estación meteorológica instalada en el mismo lugar. Además, se compararán los métodos propuestos para determinar su comportamiento ante las distintas series de tiempo presentadas, también se usará el método de aprendizaje de máquina con mejores resultados en conjunto con el teorema de Takens, y así cumplir con los objetivos descritos en esta tesis.

1.1. Motivación

En la actualidad, el consumo eléctrico va al alza debido a la enorme demanda de ésta, ya que es indispensable en la vida cotidiana de cada individuo. Estamos en una era de avances donde la electricidad juega un papel importante para humanidad. Por lo tanto, se busca la manera de proveer de este motor que es la energía eléctrica, pero ya no podemos depender cien porciento de la quema de hidrocarburos, ya que también la contaminación ha aumentado, por lo cual se ha hecho indispensable obtener energía de métodos menos contaminantes, como por ejemplo, la generación eólica, o la generación fotovoltaica; una de las características más importantes de las fuentes alternas de energía, es que su generación no es constante, ya que dependen de factores climatológicos como lo son el flujo del aire, la radiación solar, la temperatura, etc; es ahí donde se vuelve importante conocer el comportamiento a futuro de dichas variables, en nuestro caso, serían las variables que intervienen en la generación fotovoltaica, que es en las que se centra este trabajo de tesis, la radiación solar y la temperatura. Al predecir dicho comportamiento, se puede ayudar en la toma de decisiones en las plantas generadoras y evitar multiples problemas de generación y demanda, entonces para lograr esto, es necesario conocer distintos métodos de predicción y además métodos que nos ayuden a mejorar dicha predicción para así mismo generar confianza en la toma de deciciones usando nuestras predicciones.

1.1.1. Surgimiento de los Paneles solares y evolución en sus aplicaciones

El efecto fotovoltaico fue descubierto por el francés Alexandre Edmond Bequerel, Figura 1.1, en 1838 cuando tenía sólo 19 años. Bequerel estaba experimentando con una pila electrolítica con electrodos de platino cuando comprobó que la corriente subía en uno

de los electrodos cuando este se exponía al sol [1].

Figura 1.1: Alexandre Edmond Bequerel[1].

El siguiente paso se dio en 1873 cuando el ingeniero eléctrico inglés Willoughby Smith descubre el efecto fotovoltaico en sólidos, en este caso sobre el Selenio. Pocos años más tarde, en 1877, el inglés William Grylls Adams, profesor de Filosofía Natural en la King College de Londres, junto con su alumno Richard Evans Day, crearon la primera célula fotovoltaica de selenio. Si bien en todos estos descubrimientos la cantidad de electricidad que se obtenía era muy reducida, con lo cual quedaba descartada cualquier aplicación práctica, se demostraba la posibilidad de transformar la luz solar en electricidad por medio de elementos sólidos sin partes móviles. La posibilidad de una aplicación práctica del fenómeno no llegó hasta 1953 cuando Gerald Pearson de Bell Laboratories, mientras experimentaba con las aplicaciones en la electrónica del silicio, fabricó casi accidentalmente una celda fotovoltaica basada en este material que resultaba mucho más eficiente que cualquiera hecha de selenio. A partir de este descubrimiento, otros dos científicos también de Bell; Daryl Chaplin y Calvin Fuller, perfeccionaron este invento y produjeron celdas solares de silicio capaces de proporcionar suficiente energía eléctrica como para que pudiesen obtener aplicaciones

prácticas de ellas. De esta manera empezaba la carrera de las placas fotovoltaicas como proveedoras de energía [1].

La energía solar fotovoltaica en los últimos años.

En la década de los 90 y en los primeros años del siglo XXI, las celdas fotovoltaicas han experimentado un continuo descenso en su coste junto con una ligera mejora de su eficiencia. Estos factores unidos al apoyo por parte de algunos gobiernos hacia esta tecnología han provocado un espectacular impulso de la electricidad solar en los últimos años como se muestra en la Figura 1.2.

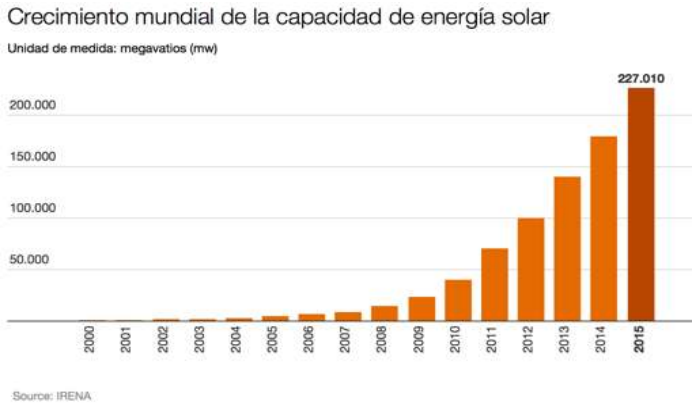


Figura 1.2: Crecimiento mundial de la capacidad de energía solar desde el año 2000 al 2015 [2].

Entre las medidas de apoyo al sector llevadas a cabo por algunos gobiernos, destacan las leyes de primas que obligan a comprar la electricidad fotovoltaica a una tarifa mucho más alta que la de la venta, lo que ayuda a rentabilizar la instalación en un periodo de tiempo pequeño. Esta medida se ha aplicado en España y Alemania, entre otros países, con un enorme éxito propiciando un importante despegue de este tipo de tecnología. Además, las instalaciones de equipo fotovoltaico han contado con muchas subvenciones en diversos países y administraciones que financiaban una parte importante de los costos, facilitando su adquisición. El concepto de granja solar también ha tenido un importante éxito. La granja solar es la asociación de varios inversores en paneles solares que forman una central generadora de energía compartiendo un mismo terreno y los diversos gastos (vigi-

lancia, mantenimiento, conexión a la red, equipamiento etc.). Normalmente se llevan a cabo en países que subvencionan las tarifas de venta de este tipo de energía. Este concepto ha animado a muchos inversores que han visto en ella una fuente de ingreso fija y fiable [1].

1.1.2. Granjas solares en el mundo

La predicción dentro de la generación fotovoltaica a gran escala juega un papel importante, ya que se manejan voltajes al rededor de los GW, por lo tanto al predecir con un buen método podemos evitar un bajo abasto de energía así como también un efecto contrario a este, como puede ser una sobre carga en nuestras lineas de potencia.

Algunas de las plantas más grandes de generación fotovoltaica son las siguientes (de las cuales México de encuentra dentro de la ubicación de una):

Parque Solar del Desierto de Tengger de 1.500MW en China

Figura 1.3: Imagen de la planta solar del Desierto de Tengger en china, con capacidad de 1.547 MW.

La planta solar del Desierto de Tengger es la mayor del mundo conectada hasta la fecha. Tiene una capacidad de 1.547 MW y se instaló en el desierto de Tengger, en Zhongwei, provincia de Ningxia. Se comenzó a construir en 2012 y se concluyó a finales de 2015, aunque no se conectó a la red hasta un año después. Se la conoce en China como la Gran Muralla Solar. El desierto de Tengger es una región natural árida que cubre aproximadamente 36.700 km y se encuentra principalmente en la región autónoma de Mongolia Interior en China. La

planta solar cubre un área de 1.200 Km2, equivalente al 3,2 % de la superficie del desierto. La planta de Tengger Desert es operada por National Grid Zhongwei Power Supply Co.

Kurnool Ultra Mega Solar Park de 1.000 MW en la India

Figura 1.4: Imagen del parque solar Kurnool Ultra Mega Solar Park en India, con capacidad de 1.000 MW.

El pasado 28 de abril los medios de comunicación de la India informaron que ya se habían conectado a la red 900 MW del parque fotovoltaico indio Kurnool Ultra Mega Solar Park, un parque solar que, cuando esté terminado a finales de este mes, contará con 1.000 MW de capacidad, pero que a día de hoy ya es la planta fotovoltaica más grande del mundo, al haber superado los 850 MW de la china Longyangxia Solar Park. El parque ocupa una superficie de 2.400 hectáreas en Panyam Mandal, en el distrito de Kurnool, en Andhra Pradesh. El proyecto está siendo ultimado por Andhra Pradesh Solar Power Corporation Private Limited (APSPCL), una empresa conjunta de Solar Energy Corporation of India, Andhra Pradesh Power Generation Corporation y New and Renewable Energy Development Corporation of Andhra Pradesh Ltd. La construcción del parque ha requerido una inversión de alrededor de 7.000 millones de rupias (unos 1.100 millones de dólares) cuya financiación ha corrido a cargo de los desarrolladores y los gobiernos central y estatal. Los desarrolladores

invirtieron 6.000 millones de rupias (unos 930 millones de dólares), y el resto fue financiado por APSPCL y una subvención del Gobierno de la Unión. El parque utiliza más de 4 millones de paneles solares con una capacidad de 315 vatios cada uno. Los paneles están conectados a cuatro estaciones de 220/33 kV de 250 MW cada una y una subestación eléctrica de 400/220 kV integrada por casi 2.000 kilómetros de circuitos de cables. El parque solar Kurnool genera cerca de 8 GWh al día, producción suficiente para satisfacer el 80 porciento de la demanda eléctrica del distrito de Kurnool. NTPC Limited invitó a los desarrolladores de energía solar a que presentaran sus ofertas para la primera fase del parque el 29 de abril de 2015, y la segunda fase, el 21 de mayo de 2015. Los contratos fueron adjudicados a los desarrolladores de energía solar a mediados de diciembre de 2015. 500 MW fueron otorgados a SunEdison (su parte fue adquirida por Greenko tras la quiebra de la estadounidense) y 350 MW a Softbank Energy, 100 MW a Azure Power y 50 MW a Adani Power[19].

Parque Solar PV Villanueva de 828 MW en México

Figura 1.5: Imagen de la planta solar PV Villanueva en México con capacidad de 828 MW.

El proyecto más grande hasta la fecha en América Latina y el Caribe es el Parque Solar Fotovoltaico Villanueva ubicado en Viesca, México. La planta tiene una capacidad total de 828 MW, una vez finalizadas las obras de ampliación en virtud una opción de

extensión de capacidad de 10 porciento incluida en el contratos de venta de energía para la planta. La planta ha sido desarrollada por Enel Green Power México (EGPM) y fue inaugurada parcialmente el 22 de marzo de 2018. EGPM ha invertido alrededor de 710 millones de dólares en la construcción de Villanueva. Villanueva, cuya construcción comenzó en marzo de 2017, es la mayor planta solar operativa en México, la mayor del continente americano y la tercera del mundo, además de ser el mayor proyecto solar de Enel a nivel mundial. El proyecto comprende más de 2,5 millones de paneles solares, capaces de producir más de 2.000 GWh por año y de evitar la emisión de más de 1 millón de toneladas de CO 2 a la atmósfera. EGPM actualmente solo posee el 20 % de la planta ya que vendió el 80 % de la planta solar al inversor institucional canadiense Caisse de Dépot et Placement du Québec (CDPQ) y el fondo de pensiones mexicano CKD Infraestructura México SA de CV (CKD IM) en octubre de 2017.

Solar Star Solar Farm I y II de 597 MW de Estados Unidos

Figura 1.6: Imagen de la central fotovoltaica Solar Star Solar Farm I y II en estados unidos, con capacidad de 597 MW.

Solar Star es una central fotovoltaica de 597 MW ubicada en las proximidades de Rosamond, California. Consta de dos fases: la primera, de 318 MW, y una segunda de 279

MW. La planta fue finalizada en junio de 2015, y es actualmente la cuarta planta solar más grande del mundo en términos de capacidad instalada, con 1,7 millones de paneles solares fabricados por SunPower y repartidos sobre una superficie de alrededor de 13 kilómetros cuadrados (3.200 acres). La planta es propiedad de MidAmerican Solar, una filial del grupo MidAmerican Renewables. En comparación con otras plantas fotovoltaicas de tamaño similar, Solar Star utiliza un número más pequeño (1,7 millones) de paneles de eficiencia más alta, montados sobre seguidores de eje único. En contraste, la plantas fotovoltaicas Desert Sunlight y el Topaz Solar Farm (de 550 MW cada una) utilizan un número mayor (aproximadamente 9 millones) de módulos fotovoltaicos de teloruro de cadmio en lugar de la tecnología cristalina fotovoltaica de silicio convencional, repartidos en un área más grande (alrededor de 25 kilometros cuadrados). En cualquier caso, ambos tipos de instalaciones son comercialmente viables [19].

1.1.3. Análisis de series de tiempo.

Las series de tiempo han representado un importante conjunto de datos para numerosas disciplinas, sin embargo, no es hasta la década de los 70 en el milenio anterior (1970), donde aparecen importantes descubrimientos al respecto. Sucede así, que George E.P. Box y G.M Jenkins, desarrollan la técnica ARIMA [3].

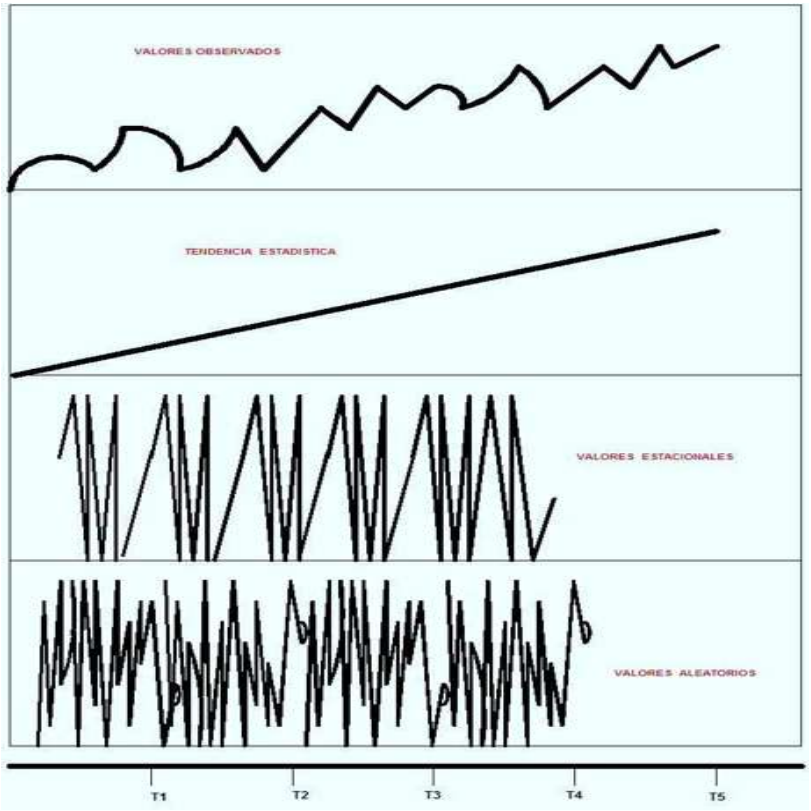


Figura 1.7: Componentes de una serie de tiempo [3].

definen un modelo econométrico que depende de sus estados anteriores, esto es una variable yt va a depender de sus estados anteriores $yt-1, yt-2, yt-3...yt-j$. Box-Jenkins parten de la concepción de la no estacionariedad de la serie, recordemos que este concepto viene de las matemáticas, donde el matemático francés J. Fourier, consigue la aproximación de una serie en términos de funciones de la misma variable, expresadas como senos y cosenos, comprobando su orientación determinística. Pero los economistas necesitaban introducir el concepto estocástico dentro de los modelos, porque existía en el proceso de cálculo con datos

determinísticos un conjunto de resultados de diferencias, que se atribuyen a un proceso aleatorio no controlable. Los trabajos pioneros del matemático A.N. Kolgomorov en 1931 y del estadístico británico George Yule (1871-1951); promueven la investigación y las primeras aplicaciones econométricas basadas en estos modelos autorregresivos de segundo orden. Basado en los trabajos de Yule, el matemático y economista ruso Eugen Slutsky (1880-1948), introduce esta metodología en el análisis de los ciclos económicos, desde un punto de vista dinámico y estocástico, sentando las bases estadísticas que fundamentaran luego las teorías de los ciclos económicos y el análisis econométrico de las series de tiempo. Para los años de 1960 y en adelante, los investigadores habían desarrollado modelos donde era posible descomponer una serie de datos en sus partes como la tendencia, el componente cíclico y una componente aleatoria e irregular, basándose en la aplicación de medias móviles y en las desviaciones de la variable, esto se completa con el procedimiento estadístico denominado X-11 (Ahora se usa un modelo Arima X-12), que consta de un algoritmo muy sencillo resumible en 5 pasos y utilizado por las oficinas estadísticas de gran parte del mundo occidental, cuya premisa fundamental es concebir que una serie temporal se puede descomponer en:

- Tendencia de largo plazo.
- Variación cíclica.
- Variación estacional.
- Variación residual.

El trabajo de Box-Jenkins "Time Series Analysis. Forecasting and Control"(1970) permitió considerar una descripción más fidedigna de la serie temporal, así; a partir de la transformación de la serie real que es estacionaria, tenemos que, mediante la aplicación de un procedimiento de suma, integrando las variables de la serie estacionaria se llega a la serie real; de ahí el nombre de modelos auto regresivos integrados y medias móviles. Desarrollos posteriores permitieron a los econometrías y economistas Dr. Christopher Sims y Dr. Clive Grainger (ambos premios nobel en economía, el Dr. Granger falleció en 2009) y a la Dra. Roselyn Joyeux, trabajar en series con integración fraccional y series no lineales, culminando en la metodología de los vectores autoregresivos y la cointegración [3].

1.1.4. Máquinas de soporte vectorial; surgimiento (regresión).

En épocas históricas, donde los intentos resultaban prematuros en relación con la tecnología disponible, podemos considerar que el camino hacia la construcción de máquinas inteligentes comienza en la segunda guerra mundial, con el diseño de ordenadores analógicos ideados para controlar cañones antiaéreos o para la navegación. A partir de 1937 comienza el desarrollo de las primeras computadoras como la máquina de Turing hasta llegar a 1957 donde A. Newell, H. Simon y J. Shaw presentaron el primer programa capaz de razonar sobre temas arbitrarios. Hacia 1960 John McCarthy, acuña el término de inteligencia artificial, para definir los métodos algorítmicos capaces de simular el pensamiento humano en los ordenadores. Entre los métodos teóricos más utilizados están las redes neuronales artificiales (RNA). Las RNA se han integrado dentro de los métodos ya clásicos del análisis de las relaciones cuantitativas entre la estructura y la actividad biológica u otras propiedades. Constituyen una de las áreas de la inteligencia artificial que ha despertado mayor interés en los últimos años. La razón principal, es que potencialmente son capaces de resolver problemas cuya solución por otros métodos convencionales resulta extremadamente difícil dada su capacidad de aprender. Estos modelos de aprendizaje se clasifican en: híbridos, supervisados, No supervisados y reforzados, dentro de los supervisados, se encuentra la técnica: máquinas de soporte vectorial (MSV). Las MSV son un paradigma aparte de la redes neuronales, pero a pesar de tener similitudes, están mejor fundamentadas en la teoría y tienen mucho mejor capacidad de generalización. En la actualidad, las máquinas de soporte vectorial pueden ser utilizadas para resolver problemas tanto de clasificación como de regresión. Algunas de las aplicaciones de clasificación o reconocimiento de patrones son: reconocimiento de firmas, reconocimiento de imágenes como rostros y categorización de textos. Por otro lado, las aplicaciones de regresión incluyen predicción de series de tiempo y problemas de inversión en general [20].

Máquinas de Soporte Vectorial

Las MSV son una moderna y efectiva técnica de inteligencia artificial (IA), que ha tenido un formidable desarrollo en los últimos años, a continuación, se presentarán los

fundamentos teóricos que definen estos sistemas de aprendizaje. Uno de los conceptos fundamentales en esta técnica es el algoritmo Vector de Soporte (VS) es una generalización no-lineal del algoritmo Semblanza Generalizada, desarrollado en la Rusia en los años sesenta. El desarrollo de los VS trae consigo el surgimiento de las Máquinas de Soporte Vectorial. Estas son sistemas de aprendizaje que usan un espacio de hipótesis de funciones lineales en un espacio de rasgos de mayor dimensión, entrenadas por un algoritmo proveniente de la teoría de optimización [20]. La Minimización del Riesgo Empírico y la Dimensión de Vapnik-Chervonenkis son fundamentales en las Máquinas de Soporte Vectorial. Dicho de manera más sencilla, el algoritmo se enfoca en el problema general de aprender a discriminar entre miembros positivos y negativos de una clase de vectores de n-dimensiones. Las MSV pertenecen a la familia de clasificadores lineales. Mediante una función matemática denominada kernel, los datos originales se redimensionan para buscar una separabilidad lineal de los mismos. Una característica de las MSV es que realiza un mapeo de los vectores de entrada para determinar la linealidad o no de los casos, los cuales serán integrados a los Multiplicadores de Lagrange para minimizar el riesgo empírico y la dimensión de Vapnik-Chervonenkis. De manera general, las MSV permiten encontrar un hiperplano óptimo que separe las clases [20].

Funciones Kernel

Las funciones kernel son funciones matemáticas que se emplean en las MSV. Estas funciones son las que le permiten convertir lo que sería un problema de clasificación no-lineal en el espacio dimensional original, a un sencillo problema de clasificación lineal en un espacio dimensional mayor [20].

Máquina de Soporte Vectorial para Clasificación

Entre las aplicaciones más relevantes de las MSV se encuentra la clasificación, el problema de la clasificación puede reducirse a examinar dos clases sin pérdida de generalidad. La tarea es encontrar un clasificador que funcione bien en datos futuros, es decir que generalice bien la clasificación [20].

Máquinas de Soporte Vectorial para Regresión

Las MSV se desarrollaron inicialmente para solucionar problemas de clasificación, pero se han ampliado para problemas de regresión. Los resultados finales a los que se puede arribar luego del empleo de las MSV pueden ser cualitativos o cuantitativos, para el análisis cuantitativo se emplean MSV para regresión. Dicho método es una extensión del anteriormente explicado donde se incluyen los estimadores de rangos. El empleo de estos determina los valores que tienen ruido dentro de la predicción a través de funciones de pérdida, donde los primeros pasos en este sentido se dieron por Tuckey quien demostró que, en situaciones reales, se desconoce el modelo del ruido y dista de las distribuciones supuestas. A raíz de esto, Huber crea el concepto de estimadores robustos los cuales están determinados por funciones de pérdida. En la actualidad, las más utilizadas son: las funciones de pérdida cuadrática y lineal, y la de Huber, entre otras. En este tipo de técnicas, su estructura se determina sobre la base del conjunto de entrenamiento necesitándose pocos parámetros para el mismo. Dicho entrenamiento se reduce a la solución de un problema de optimización que se reduce a un problema de programación cuadrática. Al mismo tiempo, el uso de las funciones Kernels muestra una gran eficiencia en el resultado de la predicción [20].

Tipos de Máquinas de Soporte Vectorial para regresión

Dentro de las máquinas de soporte vectorial para regresión se encuentran:

- Epsilon_SVR.
- NU_SVR.

Trabajos usando aprendizaje de máquina, ARIMA, SVM y K-NN para predicción

La contaminación a nivel mundial, es un tema que se ha vuelto de suma importancia, ya que los efectos de la contaminación repercuten directamente en el comportamiento climático de la tierra. Actualmente, se han venido desarrollando distintas fuentes de energía renovables, como son la generación a base de radiación solar y la generación a base de

la velocidad del viento; más conocidas como energías fotovoltaicas y energías eólicas, respectivamente. Para estos métodos de generación de energía, se han desarrollado distintos proyectos para aprovecharlas de manera eficiente. En [21] se trata de un método de predicción de radiación solar usando aprendizaje de máquina; esto con el fin de predecir cuanta energía se va generar a ciertas horas mediante la predicción de radiación solar. Con los resultados obtenidos, estos se usan para meterlos a una red inteligente y de esta manera distribuir la energía generada de manera eficiente; sin duda un proyecto fascinante donde se combina predicción con redes inteligentes, otro proyecto similar de predicción es el presentado en [22]; donde se abordan tres métodos de predicción, dos con aprendizaje de máquina y el tercero con el método ARIMA; con el fin de predecir la velocidad del viento y la radiación solar a ciertas horas, para combatir la fluctuación provocada por la variación de generación eléctrica a ciertas horas, ya que la radiación solar y la velocidad del viento no son constantes.

1.2. Objetivos de la Tesis

1.2.1. Objetivos generales

- Analizar y comparar los métodos basados en aprendizaje de máquina (K-NN y MSV), así como también con el modelo ARIMA, haciendo predicción en series de tiempo relacionadas con la generación fotovoltaica.

- Analizar en que tipo de casos (casos donde se cuenten con pocos datos o muchos datos), son mejores los métodos basados en aprendizaje de máquina (K-NN y SVM) como regresores, así como también el modelo ARIMA.

- Cual de estos 3 métodos de predicción obtiene mejores resultados en las pruebas establecidas en este trabajo de tesis.

- Conocer que tanto pueden mejorar los resultados de prediccion de un método, al trabajar en conjunto con el teorema de Takens.

1.2.2. Objetivos particulares

- Aprender a calcular los parámetros de un modelo AR(p).
- Entender los principios básicos del funcionamiento de los métodos basados en aprendizaje de máquina(K-NN y SVM) en su modalidad como regresores, además del modelo ARIMA.
- Aplicar las funciones tanto de Matlab como de Python para hacer predicción con los métodos K-NN y MSV, así como también del modelo ARIMA.

1.3. Descripción de Capítulos

En el Capítulo 1 se realiza una breve reseña de los antecedentes asociados del presente trabajo de tesis, así como el alcance que se pretende tener con esta investigación.

En el Capítulo 2 se presenta el marco teórico, el cual corresponde el conjunto de definiciones y ecuaciones que se utilizarón para fundamentar y realizar esta tesis, se divide en 3 partes, series de tiempo, modelos probabilísticos y aprendizaje de máquina.

En el capítulo 3 se describe más ampliamente el problema a resolver, también se encuentran todas las pruebas que se realizaron en el presente trabajo, se presentan los resultados de dichas pruebas.

En el Capítulo 4 se presentan los trabajos futuros, además de las conclusiones finales del presente trabajo, conclusiones que expresan lo que significo realizar este trabajo y una explicación de los resultados obtenidos en el mismo.

Capítulo 2

Algoritmos MSV, K-NN y ARIMA para predicción de series de tiempo

2.1. Introducción

En este Capítulo se define lo que es una serie de tiempo, su utilidad y las partes que la componen y también se define lo que es el modelo Arima (p,d,q) y los modelos que lo componen, los cuales son los modelos AR(p) y MA(q), se define también lo que es el aprendizaje de máquina, así como algunos métodos que emplean este aprendizaje de máquina, tales métodos son, la máquina de soporte vectorial y la regla K vecinos más cercanos, además se presenta una breve introducción al teorema de Takens.

2.1.1. Introducción a las series de tiempo

Series de tiempo

Una serie temporal es una secuencia de datos, observaciones o valores medidos en determinados momentos y ordenados cronológicamente. Los datos pueden estar espaciados a intervalos iguales, estas observaciones pueden ser: observaciones de variables eléctricas, en medicina pueden ser distintos valores del peso de una persona, en economía, distintos valores que puede tomar una divisa o ganancias monetarias, etc. Todos esos valores espaciados en intervalos iguales de tiempo. En la figura 2.1 se ilustra un ejemplo de una serie de tiempo

y sus componentes [4].

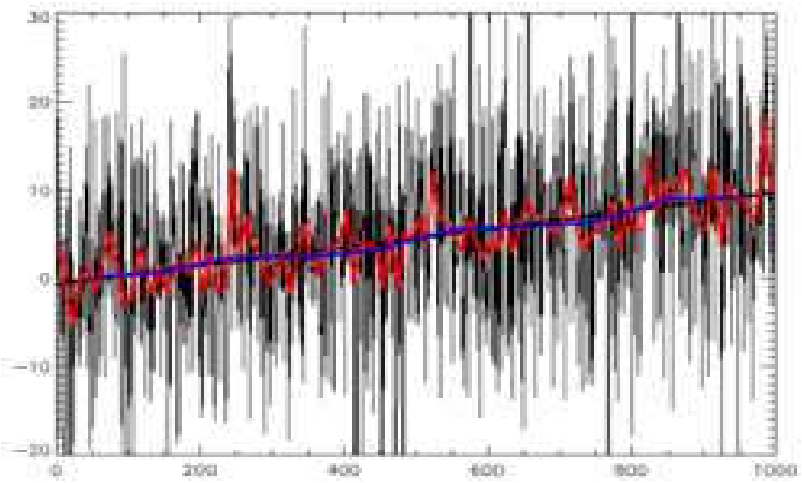


Figura 2.1: Una serie temporal formada por fluctuaciones aleatorias superpuesta a una tendencia creciente, la línea de mejor ajuste y diferentes suavizados de la serie [4].

¿Para qué nos sirven las series de tiempo?

Para el análisis de las series temporales se usan métodos que ayudan a interpretarlas, y que permiten extraer información representativa sobre las relaciones subyacentes entre los datos de la serie o de diversas series. Además, permiten en diferente medida y con distinta confianza, extrapolar o interpolar los datos y así predecir el comportamiento de la serie en momentos no observados, sean en el futuro (extrapolación pronóstica), en el pasado (extrapolación retrógrada) o en momentos intermedios (interpolación). Uno de los usos más habituales de las series de datos temporales es su análisis para predicción y pronóstico (así se hace por ejemplo con los datos climáticos, las acciones de bolsa, o las series de datos demográficos). Resulta difícil imaginar una rama de la ciencia en la que no aparezcan datos que puedan ser considerados como series temporales. Las series temporales se estudian en estadística, procesamiento de señales, econometría y muchas otras áreas [23].

Pronóstico de series de tiempo

El pronóstico de las series de tiempo, significa que extendemos los valores históricos al futuro, donde aún no hay mediciones disponibles. El pronóstico se realiza generalmente

para optimizar áreas como los niveles de inventario, la capacidad de producción o los niveles de personal [23].

Existen dos variables estructurales principales que definen un pronóstico de serie de tiempo:

- **Periodo:**

  El periodo es la designación al intervalo de tiempo necesario para completar un ciclo repetitivo, o la duración de un espacio de tiempo. Los periodos más comunes son meses, semanas y días en la cadena de suministro (para la optimización del inventario). Los centros de atención telefónica utilizan períodos de cuartos de hora (para la optimización del personal) existentes [23].

- **Horizonte:**

  El horizonte representa la cantidad de períodos por adelantado que deben ser pronosticados. En la cadena de suministro, el horizonte es generalmente igual o mayor que el tiempo de entrega [23].

Luego, existen algunas sutilezas relacionadas con la definición del período mismo, principalmente debido a irregularidades del calendario. Por ejemplo, uno puede decidir que la agregación mensual comienza el día N de cada mes (en lugar del 1°), pero si N es mayor que 28, causa problemas porque no todos los meses tienen más de 28 días [23].

2.2. Componentes de una serie de tiempo

Los datos de una serie de tiempo, se pueden descomponer en componentes individuales para facilitar su estudio, los cuales se explican a continuación [5].

Tendencia

La tendencia de una serie de tiempo, es la componente de largo plazo que representa el crecimiento o disminución en la serie sobre un periodo amplio. Como se puede ver,

la tendencia es la propensión al aumento o disminución en los valores de los datos de una serie de tiempo, que permanece a lo largo de un lapso muy extendido de tiempo, es decir, que no cambiará en el futuro lejano mientras no hayan cambios significativos o radicales en el entorno en el que se encuentra inmersa, y que determina el comportamiento de la serie de tiempo en estudio, cambios que podrían ser originados como por ejemplo, por descubrimientos científicos, avances tecnológicos, cambios culturales, geopolíticos, demográficos, religiosos, etc. Un ejemplo se muestra en la figura 2.2 [5], en la cual se observa un crecimiento notable en la señal al transcurrir un periodo de tiempo en la serie.

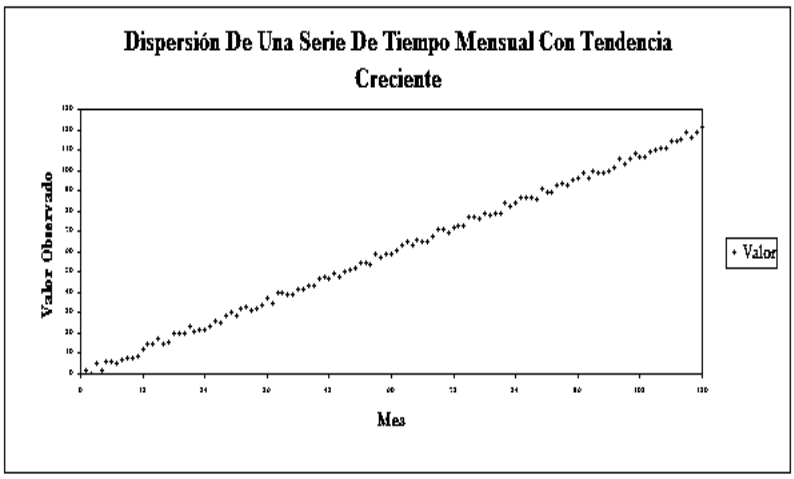


Figura 2.2: Serie de tiempo mensual con tendecia creciente [5].

Estacionalidad

La componente estacional, es un patrón de cambio que se repite a sí mismo año tras año. El patrón de cambio, por lo general es un aumento o una disminución cuantitativa en los valores observados de una serie de tiempo específica. Cabe mencionar que, aunque en la mayor parte de los casos, el patrón estacional es un fenómeno que se presenta en lapsos de tiempo de duración aproximada a un año; también puede manifestarse éste fenómeno en periodos de tiempo, ya sean menores o mayores a un año. Como por ejemplo, el caso de la verificación de vehículos que se eleva en las dos primeras semanas de cada período de verificacioń, ocurriendo esto cada dos meses, siendo éste lapso de tiempo menor a un año. O el caso del aumento en las ventas de panfletos publicitarios, sucedido esto cada seis años y ocasionado por las elecciones presidenciales, siendo éste un lapso mayor a un año. Un ejemplo se muestra en la figura 2.3 [5].

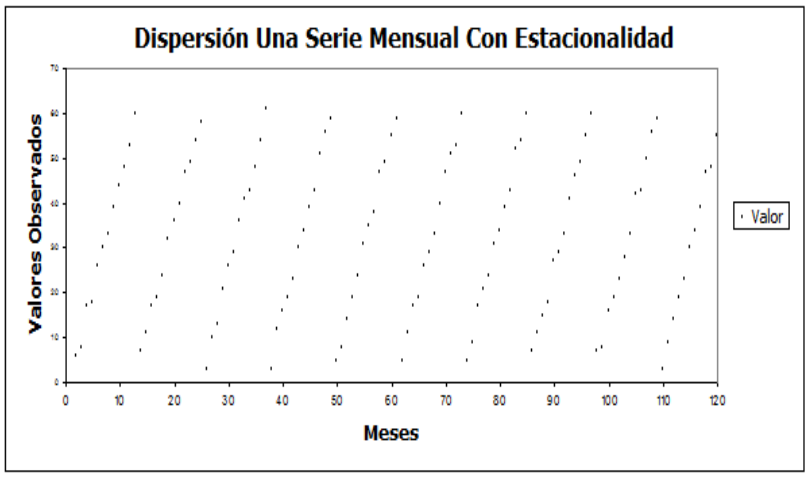


Figura 2.3: Serie de tiempo mensual con estacionalidad [5].

Esta gráfica muestra un notorio patrón en los valores de la serie de tiempo, que parece repetirse en lapsos de tiempo aproximados a un año.

Ciclicidad

El componente cíclico, es la fluctuación en forma de onda alrededor de la tendencia. La ciclicidad es un fenómeno, que en lo general parece estar relacionado con la variación

de la actividad económica ocurrida durante periodos de crisis o prosperidad. La fluctuación también puede presentarse en series de tiempo estacionarias, un ejemplo se muestra en la figura 2.4 [5].

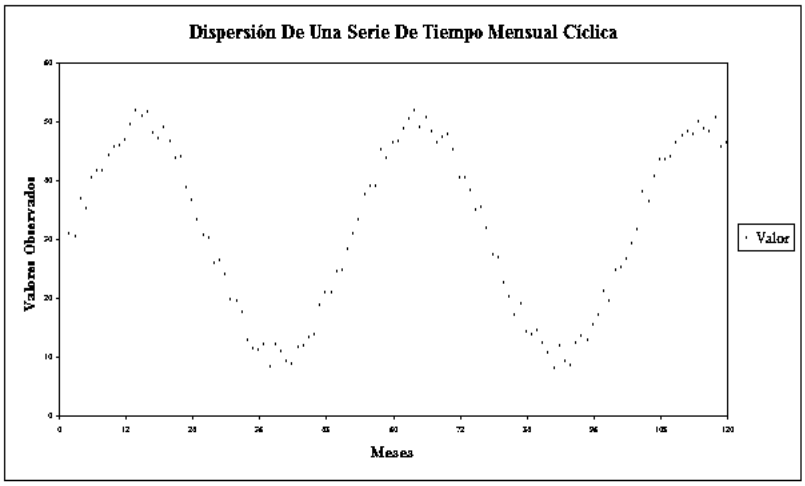


Figura 2.4: Serie de tiempo mensual con ciclicidad [5].

Aleatoriedad

El componente aleatorio mide la variabilidad de las series de tiempo después de retirar los otros componentes.

La aleatoriedad se presenta en todas las series de tiempo y no es otra cosa, que el cambio producido en los valores de una serie de tiempo debido a fenómenos que son en extremo difíciles de explicar y que por lo tanto, su ocurrencia cae en el ámbito del azar; un ejemplo se muestra en la figura 2.5 [5].

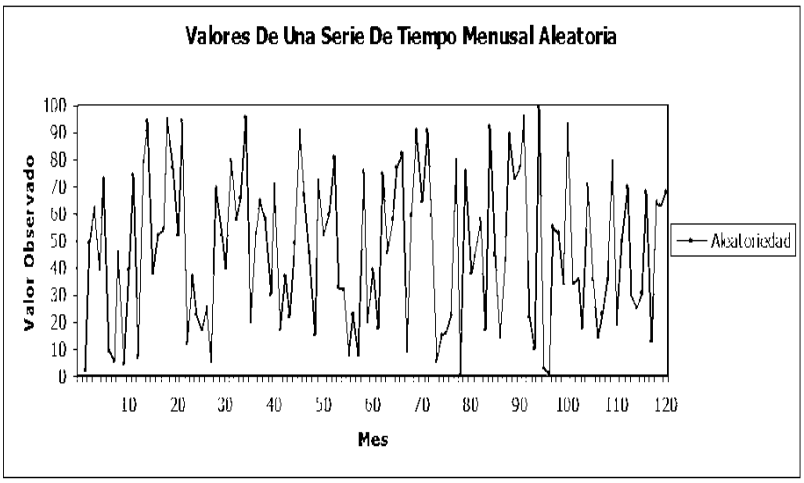


Figura 2.5: Serie de tiempo mensual aleatorio [5].

2.3. Modelos probabilísticos

Los modelos probabilísticos son una representación matemática deducida de un conjunto de supuestos, con el doble propósito de estudiar resultados de un experimento aleatorio y predecir su comportamiento futuro [24].

2.3.1. Modelo AR(p)

En estadística y procesamiento de señales, un modelo autorregresivo (AR), es una representación de un tipo de proceso aleatorio, que como tal, describe ciertos procesos variables en el tiempo, ya sea en la naturaleza, la economía, etc. El modelo autorregresivo especifica que la variable de salida depende linealmente de sus propios valores anteriores [25] se expresa mediante la ecuación (2.1).

$$y_t = \theta_0 + \theta_1 y_{t-1} + ... + \theta_p y_{t-p} + \epsilon_t = \sum_{i=1}^{p} \theta_0 + \theta_i y_{t-i} + \epsilon_t \tag{2.1}$$

Donde:

θ_i: Son los coeficientes de auto-regresión

y_t: Es la serie bajo investigación

p: Es el orden (longitud) del modelo.
ϵ_t: Es ruido, casi siempre se supone que es ruido blanco Gaussiano. i: Es el índice inferior de la sumatoria.

El problema en el análisis AR(P), es derivar los "mejores" valores para θ_i dada una serie y_t . La mayoría de los métodos asumen que la serie y_t es lineal y estacionaria. Por convención, se asume que la serie y_i es media cero, si no, esto es simplemente otro término θ_0 en frente de la suma en la ecuación anterior (2.1) [25].

2.3.2. Modelo AR(p) para la predicción de valores

En ocasiones, se pretende predecir el comportamiento de una variable y en un momento futuro t, a partir del comportamiento que la variable tuvo en un momento pasado, por ejemplo, en el período anterior, y_{t-1} . Formalmente notaríamos que:

$$y_t = f(y_{t-1}) \tag{2.2}$$

Es decir, que el valor de la variable y en el momento t es función del valor tomado en el periodo $t-1$ [26].

Puesto que, en el comportamiento de una variable influyen más aspectos, debemos incluir en la relación anterior un término de error, ϵ_t, que es una variable aleatoria a la que suponemos ciertas características estadísticas apropiadas [26]. Es decir:

$$y_t = f(y_{t-1,\epsilon_t}) \tag{2.3}$$

Ahora debemos elegir una forma funcional concreta para esta expresión [26]. Por ejemplo, una forma lineal como la ecuación (2.4)

$$y_t = \theta_0 + \theta_1 y_{t-1} + \epsilon_t \tag{2.4}$$

donde θ_0 es un término independiente y θ_1 es un parámetro que multiplica al valor de la variable y en el período t-1. Utilizando métodos estadísticos adecuados, se puede estimar los parámetros θ_0 y θ_1 de forma que estos cumplan propiedades estadísticas razonables y sean una buena (la mejor posible) estimación. Con esto se obtiene una expresión que es la

esencia de los modelos autorregresivos (modelos AR). Se realiza una regresión de la variable y_t sobre si misma (auto-regresión) o, mejor dicho, sobre los valores que la variable tomó en el período anterior [26].

Un aspecto importante, es el orden del modelo AR. Por ejemplo, el modelo $y_t = \theta_0 + \theta_1 y_{t-1} + \epsilon_t$ es de orden 1, y se denota como AR(1). Si tomamos en el modelo como explicativas los valores de la variable y en los 2 períodos anteriores, es decir: $y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \epsilon_2$, entonces hemos especificado un AR(2). De igual forma, un AR(3) vendría dado por $y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \theta_3 y_{t-3} + \epsilon_t$. En general, un AR(p) viene dado por $y_t = \theta_0 + \theta_1 y_{t-1} + \theta_2 y_{t-2} + ... + \theta_p y_{t-p} + \epsilon_t$. Es frecuente encontrarnos con modelos AR de bajo orden (1 o 2) [26].

Coeficientes de un modelo AR(p)

Existen varias técnicas para calcular los coeficientes AR. Las dos categorías principales son: mínimos cuadrados y el método Burg. Dentro de cada una de estas, hay algunas variantes, el método de mínimos cuadrados más común se basa en las ecuaciones de Yule-Walker [26].

2.3.3. Modelo Ma(q)

Una alternativa de modelización pasa por tratar de explicar el comportamiento de una variable y, no en función de los valores que tomó en el pasado (modelos AR) sino a través de los errores al estimar el valor de la variable en los períodos anteriores. Ello da lugar a los modelos de medias móviles (modelos MA, por sus siglas del inglés) [26].

Por ejemplo, un modelo MA(1) viene dado por la ecuación (2.5)

$$y_t = \mu + \alpha_t + \theta_1 \alpha_{t-1} + ... + \theta_q \alpha_{t-q} \tag{2.5}$$

Donde:

μ: Es la media aritmetica.

θ_q: Son los coeficientes del modelo.

q: Es el orden del modelo.

α_t: Elementos de la serie.

Al igual que ocurre con los modelos AR, en series con componente estacional, es frecuente que el retardo coincida con la periodicidad de los datos. Entre los modelos AR y MA existe una relación, bajo ciertas condiciones, que es útil conocer [26]. Los modelos ARMA integran a los Modelos AR y MA en una única expresión. Por tanto, la variable queda en función de los valores tomados por la variable en períodos anteriores, y los errores cometidos en la estimación. Una expresión general de un modelo ARMA (p, q) viene dado por la ecuación (2.6)

$$y_t = \mu + \theta_1 y_{t-1} + ... + \theta_p y_{t-p} + \alpha + \theta_1 \alpha_{t-1} + ... + \theta_q \alpha_{t-q} \tag{2.6}$$

que es la unión de un modelo AR(p) y un modelo MA(q) [26]. Obviamente, los modelos AR (p) se corresponden con el modelo ARMA (p, 0), mientras que los modelos MA (q) se corresponden con el modelo ARMA (0, q).

2.3.4. Modelo ARIMA(p,d,q)

El modelo autorregresivo integrado de media móvil, es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Se trata de un modelo dinámico de series temporales, es decir, las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes [26]. Se suele expresar como ARIMA(p, d, q), donde los parámetros p, d y q son números enteros no negativos que indican el orden de las distintas componentes del modelo, las componentes autorregresiva, integrada y de media móvil, respectivamente [26]. Para la obtención de estimaciones con propiedades estadísticas adecuadas de los parámetros de un modelo ARMA, es necesario que la serie muestral que utilizamos para la estimación, sea estacionaria en media y varianza. En un sentido laxo del término, diríamos que precisamos que la serie no tenga tendencia, y que presente un grado de dispersión similar en cualquier momento de tiempo [26]. A efectos prácticos, el cumplimiento de esta propiedad

pasa por tomar logaritmos y diferenciar adecuadamente la serie original objeto de estudio. Con la serie ya tratada para convertirla en estacionaria, ya es posible estimar un modelo ARMA. Pues bien, un Modelo Autorregresivo-Integrado de Medias Móviles de orden p, d, q, o abreviadamente ARIMA (p,d,q), no es más que un modelo ARMA (p,q) aplicado a una serie integrada de orden d (I(d)), es decir, a la que ha sido necesario diferenciar d veces para eliminar la tendencia [26]. Por lo tanto, la expresión general de un modelo ARIMA (p,d,q) viene dada por la ecuación (2.7)

$$\Delta^d y_t = \theta_1 \Delta^d y_{t-1} + ... + \theta_p \Delta^d y_{t-p} + \alpha + \theta_1 \alpha_{t-1} + ... + \theta_q \alpha_{t-q} \tag{2.7}$$

Donde $\Delta^d y_t$, expresa que sobre la serie original y_t, se han aplicado d diferencias. Por lo tanto, sobre una serie integrada de orden 2, necesitaría una doble diferenciación, lo cual se expresa como: $\Delta^2 y_t = \Delta(\Delta y_t) = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$. Obsérvese que en la expresión del ARIMA (p, d, q), desaparece el término independiente justamente por la aplicación de las diferencias sucesivas. Cuando la estimación del modelo ARMA se haya realizado con una serie diferenciada, a efectos de predicción, es necesario recalcularla integrando nuevamente la serie. Un aspecto importante de notación hace referencia al operador retardo [26].

2.4. Aprendizaje de máquina

El aprendizaje de máquina, es el diseño y estudio de las herramientas informáticas que utilizan la experiencia pasada para tomar decisiones futuras; es el estudio de programas que pueden aprender de los datos. El objetivo fundamental del aprendizaje de máquina, es generalizar, o inducir una regla desconocida a partir de ejemplos donde esa regla es aplicada. El ejemplo más típico donde podemos ver el uso del aprendizaje de máquina es en el filtrado de correos basura o spam. Mediante la observación de miles de correos electrónicos que han sido marcados previamente como basura, los filtros de spam aprenden a clasificar los mensajes nuevos. El aprendizaje de máquina combina conceptos y técnicas de diferentes áreas del conocimiento, como las matemáticas, estadísticas y las ciencias de la computación; por tal motivo, hay muchas maneras de aprender la disciplina [27].

Tipos de aprendizaje de máquina

El aprendizaje de máquina tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica. Para abordar cada uno de estos temas es crucial en primer lugar distingir los distintos tipos de problemas del aprendizaje de máquina[27].

Aprendizaje supervisado

En los problemas de aprendizaje supervisado se enseña o entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta. Cuanto mayor es el conjunto de datos, el algoritmo puede aprender más sobre el tema. Una vez concluido el entrenamiento, se le brindan nuevos datos, ya sin las etiquetas de las respuestas correctas, y el algoritmo de aprendizaje utiliza la experiencia pasada que adquirió durante la etapa de entrenamiento para predecir un resultado. Esto es similar al método de aprendizaje que se utiliza en las escuelas, donde se enseñan problemas y las formas de resolverlos, para luego aplicar los mismos métodos en situaciones similares.

Aprendizaje no supervisado

En los problemas de aprendizaje no supervisado, el algoritmo es entrenado usando un conjunto de datos que no tienen ninguna etiqueta; en este caso, nunca se le dice al algoritmo lo que representan los datos. La idea es que el algoritmo pueda encontrar por si solo patrones que ayuden a entender el conjunto de datos. El aprendizaje no supervisado es similar al método que utilizamos para aprender a hablar cuando somos bebés, en un principio escuchamos hablar a nuestros padres y no entendemos nada; pero a medida que vamos escuchando miles de conversaciones, nuestro cerebro comenzará a formar un modelo sobre cómo funciona el lenguaje y comienza a reconocer patrones y a esperar ciertos sonidos [27].

Aprendizaje por refuerzo

En los problemas de aprendizaje por refuerzo, el algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error. Un buen ejemplo de este tipo de aprendizaje lo podemos encontrar en los juegos, donde vamos probando nuevas estrategias y vamos seleccionando y perfeccionando aquellas que nos ayudan a ganar el juego. A medida que vamos adquiriendo más práctica, el efecto acumulativo del refuerzo a nuestras acciones victoriosas terminará creando una estrategia ganadora [27].

Sobreentrenamiento

Como mencionamos cuando definimos el aprendizaje de máquina, la idea fundamental es encontrar patrones que podamos generalizar para luego poder aplicar esta generalización sobre los casos que todavía no hemos observado y realizar predicciones. Pero también puede ocurrir que durante el entrenamiento solo descubramos casualidades en los datos que se parecen a patrones interesantes, pero que no generalicen. Esto es lo que se conoce con el nombre de sobreentrenamiento o sobreajuste. El sobreentrenamiento es la tendencia que tienen la mayoría de los algoritmos de aprendizaje de máquina a ajustarse a unas características muy específicas de los datos de entrenamiento que no tienen relación causal con la función objetivo que estamos buscando para generalizar. El ejemplo más extremo de un modelo sobreentrenado, es un modelo que solo memoriza las respuestas correctas; este modelo al ser utilizado con datos que nunca ha visto el resultado, va a tener un rendimiento azaroso, ya que nunca logró generalizar un patrón para predecir [27].

Como evitar el sobreentrenamiento

Como mencionamos anteriormente, todos los modelos de aprendizaje de máquina tienen tendencia al sobreentrenamiento; es por esto que debemos aprender tratar de tomar medidas preventivas para reducirlo lo más posible. Las dos principales estrategias para lidiar con el sobreentrenamiento son: la retención de datos y la validación cruzada. En el

primer caso, la idea es dividir nuestro conjunto de datos, en uno o varios conjuntos de entrenamiento y otro/s conjuntos de evaluación. Es decir, que no se le pasan todos los datos al algoritmo durante el entrenamiento, sino que vamos a retener una parte de los datos de entrenamiento para realizar una evaluación de la efectividad del modelo. Con esto lo que buscamos es evitar que los mismos datos que usamos para entrenar, sean los mismos que utilizamos para evaluar. De esta forma, vamos a poder analizar con más precisión, como el modelo se va comportando a medida que más lo vamos entrenando y poder detectar el punto crítico en el que el modelo deja de generalizar y comienza a sobreajustarse a los datos de entrenamiento [27].

La validación cruzada es un procedimiento más sofisticado que el anterior. En lugar de solo obtener una simple estimación de la efectividad de la generalización; la idea es realizar un análisis estadístico para obtener otras medidas del rendimiento estimado, como la media y la varianza, y así poder entender cómo se espera que el rendimiento varíe a través de los distintos conjuntos de datos. Esta variación es fundamental para la evaluación de la confianza en la estimación del rendimiento. La validación cruzada también hace un mejor uso de un conjunto de datos limitado; ya que a diferencia de la simple división de los datos, uno para el entrenamiento y otro para evaluación; la validación cruzada calcula sus estimaciones sobre todo el conjunto de datos mediante la realización de múltiples divisiones e intercambios sistemáticos entre datos de entrenamiento y datos de evaluación [27].

Modelo de aprendizaje de máquina

Construir un modelo de aprendizaje de máquina, no se reduce solo a utilizar un algoritmo de aprendizaje o utilizar una librería de aprendizaje de máquina; sino que es todo un proceso que suele involucrar los siguientes pasos [27]:

1. **Recolectar los datos:**

   Se puede recolectar los datos desde muchas fuentes, podemos por ejemplo extraer los datos de un sitio web u obtener los datos utilizando una base de datos. Podemos también utilizar otros dispositivos que recolectan los datos por nosotros; o utilizar

datos que son de dominio público [27].

2. **Preprocesar los datos:**

   Una vez que tenemos los datos, tenemos que asegurarnos que tiene el formato correcto para nutrir el algoritmo de aprendizaje. Es prácticamente inevitable tener que realizar varias tareas de preprocesamiento antes de poder utilizar los datos. Igualmente, este punto suele ser mucho más sencillo que el paso anterior [27].

3. **Explorar los datos:**

   Una vez que ya tenemos los datos y están con el formato correcto, podemos realizar un preanálisis para corregir los casos de valores faltantes o intentar encontrar a simple vista algún patrón en los mismos que nos facilite la construcción del modelo. En esta etapa, suelen ser de mucha utilidad las medidas estadísticas y los gráficos en 2 y 3 dimensiones para tener una idea visual de cómo se comportan los datos. En este punto, se pueden detectar valores atípicos que debamos descartar; o encontrar las características que más influencia tienen para realizar una predicción[27].

4. **Entrenamiento del algoritmo:**

   En este paso, se comienza a utilizar las técnicas de aprendizaje de máquina realmente. En esta etapa nutrimos al o los algoritmos de aprendizaje con los datos que venimos procesando en las etapas anteriores. La idea es que los algoritmos puedan extraer información útil de los datos para luego poder hacer predicciones [27].

5. **Evaluación del algoritmo:**

   En esta etapa, se pone a prueba la información o conocimiento que el algoritmo obtuvo del entrenamiento del paso anterior. Se evalua que tan preciso es el algoritmo en sus predicciones y si no estamos muy conforme con su rendimiento, podemos volver a la

etapa anterior y continuar entrenando el algoritmo, cambiando algunos parámetros hasta lograr un rendimiento aceptable [27].

6. **Utilización del modelo:**

   En esta última etapa, ya ponemos a nuestro modelo a enfrentarse al problema real. Aquí, también podemos medir su rendimiento, lo que tal vez nos obligue a revisar todos los pasos anteriores [27].

2.4.1. Máquina de soporte vectorial (MSV)

Las máquinas de soporte vectorial (MSV) es un tipo de aprendizaje de máquina. Son aquellas que necesitan primero entrenarse con situaciones en las que se les dice la respuesta correcta sobre muchos ejemplos, y una vez entrenada, entra en fase de "uso", y simplemente se convierte en una caja que devuelve la respuesta ante un nuevo caso, ver la Figura 2.6 (en pocas palabras, es un método de aprendizaje supervisado) [6].

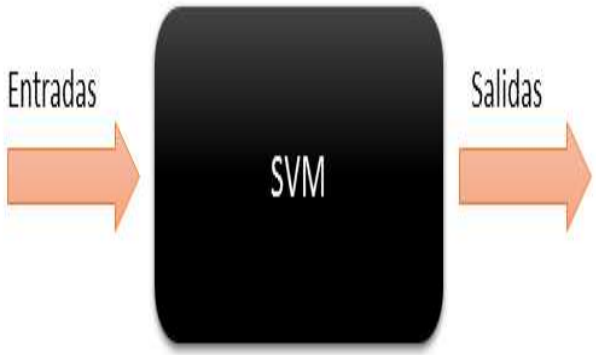


Figura 2.6: Diagrama de entrada, salida de un SVM [6].

¿Para qué sirven, y cómo funcionan?

Los conceptos fundamentales son modelado y predicción en dos vertientes:

- Clasificación

- Regresión.

La forma en que trabaja es muy interesante. Supongamos que tenemos la tarea de realizar predicciones de clasificación binaria (tenemos valores de un examen médico rutinario de una persona, y queremos saber si tiene diabetes o no). Vamos a imaginarnos que los valores recogidos en el examen son sólo 2, en lugar de cientos. Cada paciente que efectivamente tiene diabetes lo podemos poner en un plano cartesiano (donde cada eje es uno de los dos valores que recoge el examen médico). Colocamos a los pacientes que efectivamente tienen diabetes como círculos negros en el plano, en las coordenadas que corresponden a cada uno de ellos (según sus resultados de examen), y a los que no tienen diabetes, como rombos de centro blanco, como se muestra en la Figura 2.7 [7]:

Figura 2.7: Separación de pacientes, rombos blancos: sin diabetes, circulos negros: con diabetes [7].

Las MSV encuentran una "superficie"que intenta separar los ejemplos negativos y positivos con el margen más grande posible a ambos lados del hiperplano. En este caso, bidimencional, la "superficie"sería una línea. En un caso 3D (tres atributos para cada paciente) sería un plano. En un caso de más de 3 dimensiones, sería un hiper-plano o hiper-superficie con el número apropiado de variables [7].

Hay muchas formas de hacer esto, propuestas por métodos estadísticos, por la gente de redes neuronales, por la gente de optimización, etcétera. Lo que distingue a las

MSV es que el hiper-plano resultante separa los datos lo mayor posible [7].

Para los datos que tenemos en el ejemplo, podríamos tener varias posibles superficies (infinitas), pero tomemos como ejemplo la Figura 2.8 [7]:

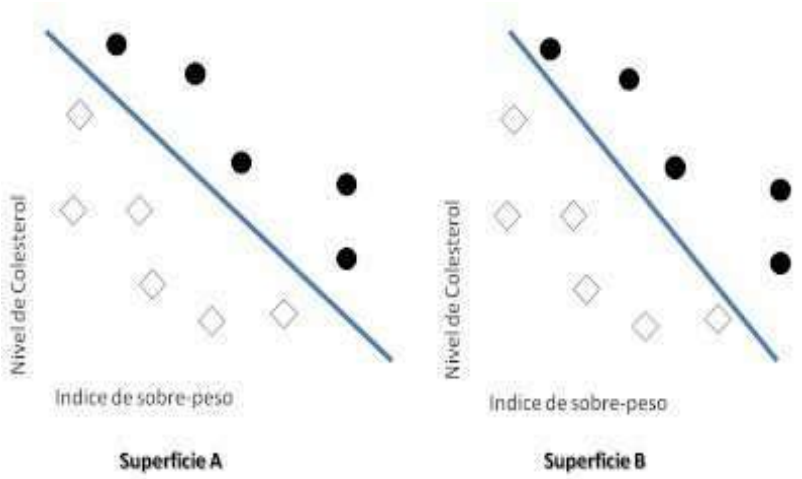


Figura 2.8: Dos tipos de margenes, uno en el centro y otro inclinado [7].

En este caso, pareciera que la Superficie A es mejor que la Superficie B.

Podemos imaginarnos el caso extremo, es decir, que la superficie estuviese adherida a algunos de los puntos de uno de los conjuntos, como en la Figura 2.9 [7]

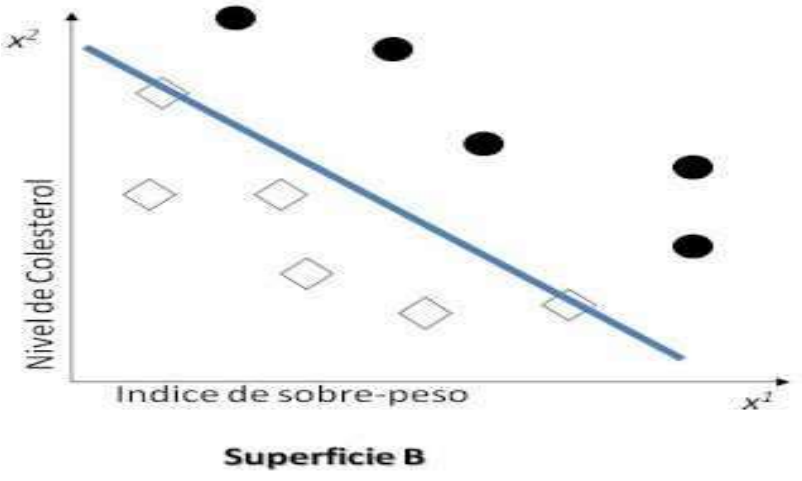


Figura 2.9: Vector de soporte cargado hacia uno de los 2 grupos [7].

Tengamos en cuenta que esos datos son de los pacientes para los cuales, hasta ahora, sabemos si tienen diabetes o no. Si dejamos que la superficie clasificadora esté allí, adherida a los pacientes sanos, intuitivamente podemos imaginar que es bastante probable que aparezca algún paciente con características similares a las de alguno de los pacientes a los cuales está adherido el hiper-plano. Pero cuando se dice "similar", intuitivamente estamos aceptando que no hay dos pacientes exactamente iguales. Debe haber alguna pequeña diferencia; ¿Y si esa pequeña diferencia hiciera que el paciente estuviese justo un ligeramente más allá de la superficie separadora? Si eso ocurriera, la máquina diría que ese paciente pertenece al grupo de los que tienen diabetes, es decir, diría que es un "círculo negro", cuando en realidad el afortunado paciente podría no tener diabetes. Estaríamos dando un falso positivo con cierta frecuencia [7]. Si el hiper-plano estuviese adherido a los pacientes del grupo de entrenamiento que eran diabéticos, estaríamos haciendo una máquina que produciría con cierta frecuencia falsos negativos (porque pacientes muy parecidos a los que ya tienen diabetes, podrían estar ya al otro lado de la superficie separadora). Uno no desearía darle falsas expectativas a un paciente, así que esto tampoco es conveniente [7]. Para lograr alejar la superficie de los puntos de ambos conjuntos, Vapnick define el margen a maximizar como la distancia entre los dos hiper-planos, paralelos al hiper-plano separador, que están, cada uno, adherido a los puntos de uno de los conjuntos. En las Superficies A y B, el margen vendría a ser la distancia entre las líneas punteadas que se muestran en la Figura 2.10 [7].

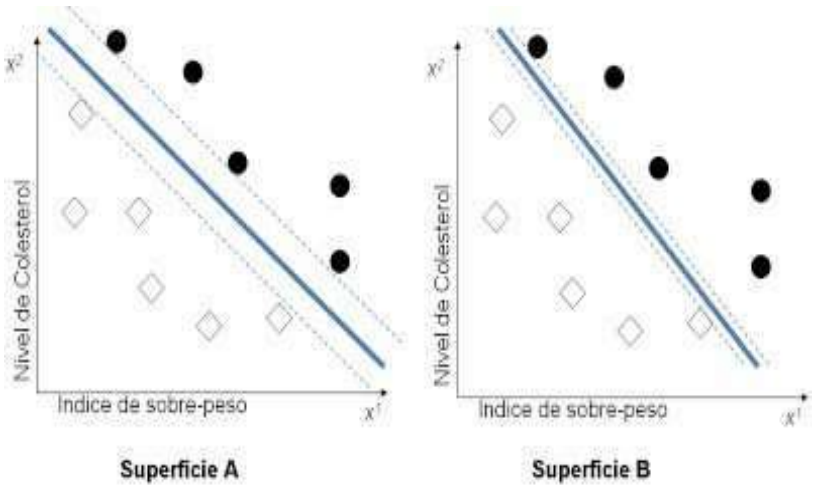


Figura 2.10: Las lineas punteadas son el margen [7].

Como podemos ver, en el caso de la Superficie A, está mucho mejor que en la B. El método, adicionalmente, coloca la superficie, en general, en la mitad de esa distancia [7].

2.4.2. Máquina de soporte vectorial para regresión

Si queremos determinar qué probabilidad hay de que un usuario vuelva a una página web, si queremos predecir el número de clics en el futuro, o qué cantidad de impresiones de un anuncio tendremos, en este caso estamos en un problema de regresión. Siguiendo los mismos principios que para los modelos AR, la regresión se basa en buscar la curva que modele la tendencia de los datos y, según ella, predecir cualquier otro dato en el futuro. Por ejemplo, si disponemos de un caso sencillo como el mostrado en la Figura 2.11, donde la probabilidad de hacer clic en un determinado anuncio depende únicamente de la edad del usuario, podremos definir (siempre minimizando el error, como las SVM garantizan) una línea de tendencia [7].

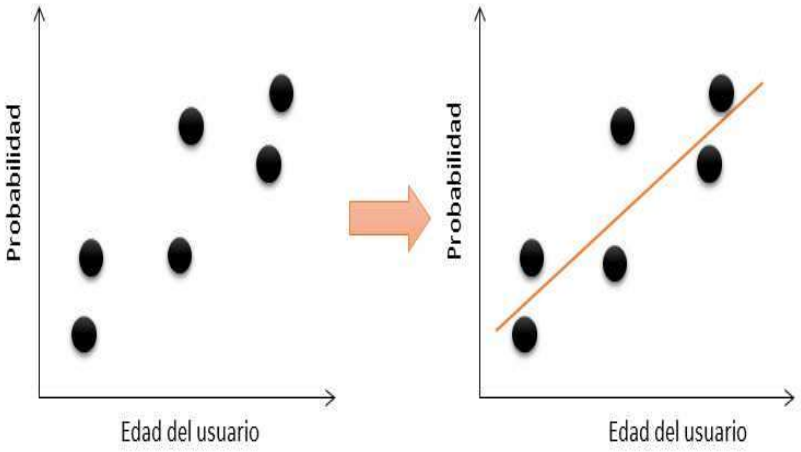


Figura 2.11: Línea de tendencia de la probabilidad de hacer clic en un determinado anuncio dependiendo de la edad del usuario[7].

De forma que se pueda encontrar la respuesta (en este ejemplo, la probabilidad) para un nuevo caso, como se muestra en la Figura 2.12 [7].

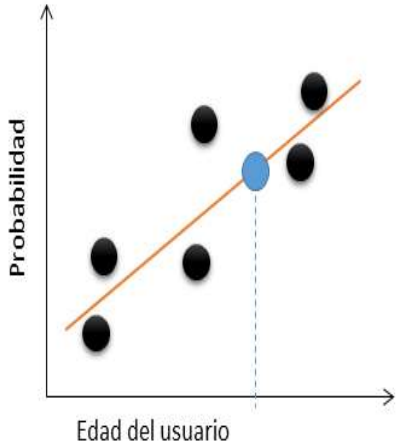


Figura 2.12: Relación en la línea de tendencia que hay entre la probabilidad de hacer clic y la edad del usuario [7].

En problemas no lineales, siempre será posible utilizar una función tipo kernel que, tras resolver el problema en un espacio donde el mismo sea lineal, obtenga la curva que modele los datos, como se muestra en la Figura 2.13 [7].

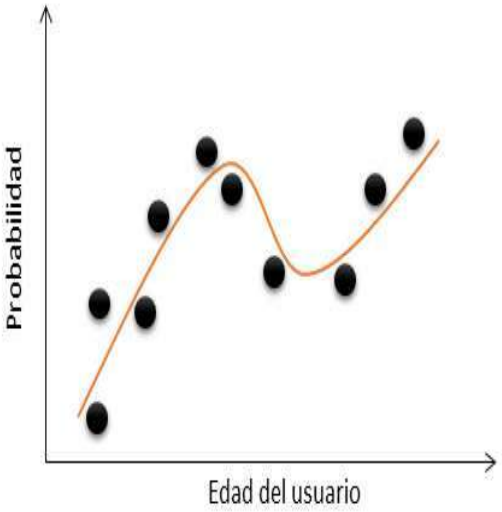


Figura 2.13: Usando una función kernel para un problema de regresión no lineal, para MSV [7].

2.4.3. Funciones Kernel

Si nos encontráramos en un caso en el que los datos no pudieran ser separados por un hiper-plano, podría ser que una superficie no-lineal pudiera separar los conjuntos, como en el ejemplo mostrado en la Figura 2.14 [7].

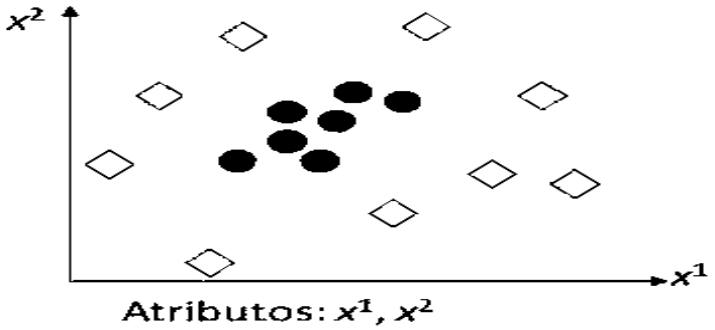


Figura 2.14: Distribución de datos no lineal [7].

Lo que se hace en MSV (y en muchas otras técnicas), es transformar el espacio de los atributos (lo que llaman el kernel). Esto suena complicado, pero si nos fijamos en el ejemplo, podemos ver que una elipse podría resolver el problema, mostrado en la Figura 2.15 [7]

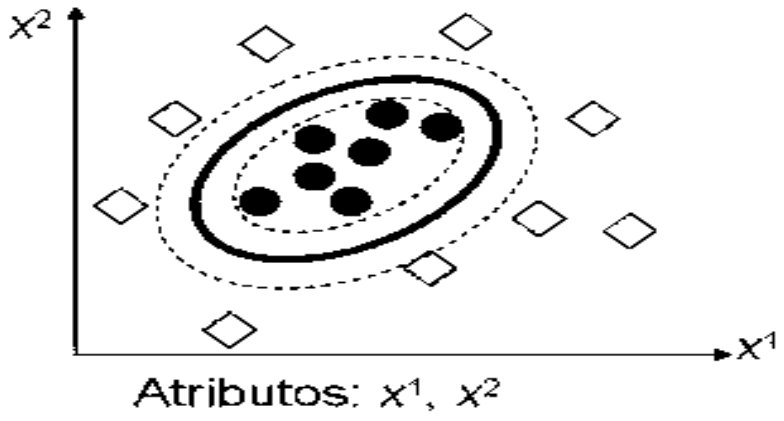


Figura 2.15: Función kernel de tipo elipse usada para un caso de distribución de datos no lineal [7].

Esa sería la superficie no-lineal que necesitamos. Todo lo que hemos venido hablando, ha sido referido a hiper-planos, y claramente la elipse no es un hiper-plano. Sin embargo, sabemos que la elipse es una figura "cónica", expresada más o menos así (en nuestro eje cartesiano del ejemplo) $a(x1+b)^2+c(x2+d)^2 = e$ donde $\{a, b, c, d, e\}$ es un conjunto de constantes, y $\{x1, x2\}$ nuestras variables $\{x, y\}$. En general, cualquier superficie cónica, termina siendo algo como esto $a1(x1)^2 + a2(x1) + a3(x1)(x2) + a4(x2) + a5(x2)^2 = a6$ Ahora, esto ni de casualidad es lineal en un espacio definido por las variables $\{x1, x2\}$. Pero si nos imaginamos un espacio donde las variables son esas dos, mas tres variables nuevas (tres dimensiones) extra $\{x3, x4, x5\}$, donde cada una de ellas representa a los términos cuadráticos de la expresión de arriba, se tiene que: $x3 = x1^2$ $x4 = x2^2$ $x5 = (x1)(x2)$ Y volviendo a escribir la ecuación cónica genérica (o cuadrática, como sería mejor llamarle), tenemos lo siguiente $a1(x3) + a2(x1) + a3(x5) + a4(x2) + a5(x4) = a6$ ¡Y acá estaremos todos de acuerdo con que se trata de una ecuación bastante lineal! Dense cuenta de que lo que se desprende de todo esto, es que un hiper-plano en este espacio de atributos ampliado, equivale a una elipse en nuestro espacio bidimensional (definido tan sólo por $\{x1, x2\}$), ver la Figura 2.16 [7].

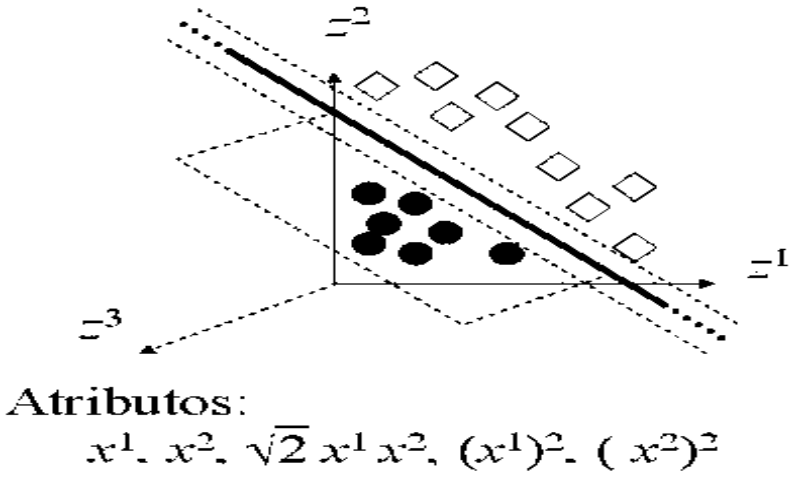


Figura 2.16: Linealidad encontrada en el espacio 5-dimensional, representada gráficamente [7].

2.4.4. Vectores de Soporte

Si asumimos que cada uno de los ejemplos de los que disponemos (círculos y rombos) es un vector en el espacio, resolver SVM es: encontrar los vectores en los que podamos apoyar los hiper-planos que definan el mayor margen de separación. Es decir, buscamos los vectores en los cuales "soportar"los hiper-planos paralelos, uno hacia un conjunto, y uno hacia el otro, para trazar justo en el medio de ambos, nuestro hiper-plano de separación. Veámoslos señalados por círculos rojos en la figura 2.17 [7].

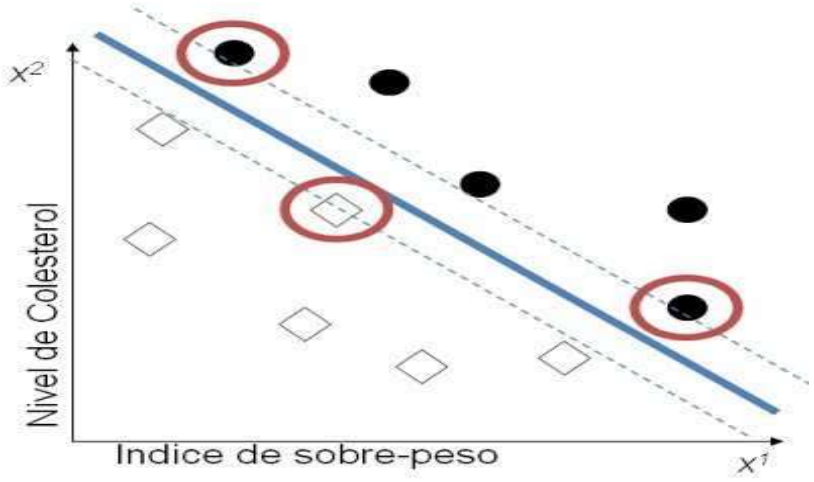


Figura 2.17: Vector de soporte [7].

2.4.5. Distancia euclídea

En matemáticas, álgebra, geometría y más específicamente, en análisis real, análisis complejo y geometría analítica, la distacia euclídea se trata de una función no negativa, usada en diversos contextos para calcular la distancia entre dos puntos, primero en el plano y luego en el espacio. También sirve para definir la distancia entre dos puntos en otros tipos de espacios, de tres o más dimensiones. Y para hallar la longitud de un segmento definido por dos puntos de una recta, del plano o de espacios de mayor dimensión. Sus bases se encuentran en la aplicación del teorema de Pitágoras sobre triángulos rectángulos, donde la distancia euclidiana viene a ser por lo general la longitud de la hipotenusa del triángulo recto conformado por cada punto y los vectores proyectados sobre los ejes directores al nivel

de la hipotenusa.

En el plano cartesiano, sean los puntos $A = (xA; yA)$ $B = (xB; yB)$, se define la distancia euclidiana entre dichos puntos por la ecuación (2.8)

$$d(A, B) = \sqrt{(X_B - x_A)^2 + (y_B - y_A)^2} \tag{2.8}$$

En el espacio, sean los puntos $A = (xA; yA; zA)$ y $B = (xB; yB; zB)$, se define la distancia euclidiana mediante la ecuación (2.9).

$$d(A, B) = \sqrt{(X_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2} \tag{2.9}$$

Y de manera más general en un espacio de N dimensiones, la distancia euclidiana entre dos puntos $A = (a1; a2; ..; aN)$ y $B = (b1; b2; ..; bN)$, se ajusta a la ecuación (2.10)

$$d(A, B) = \sum_{i=1}^{N}(b_i - a_i^2) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + ... + (b_N - a_N)^2} \tag{2.10}$$

De manera general, la métrica euclidiana entre dos puntos se define como: la longitud del segmento de recta que une a dichos puntos [8].

Importancia

Además del evidente resultado de la determinación de la longitud de un segmento de recta o la distancia entre dos puntos, pueden citarse otras muchas aplicaciones de la distancia euclidiana. Tiene sus bases en el teorema de Pitágoras, como se ve en la figura 2.18 [8].

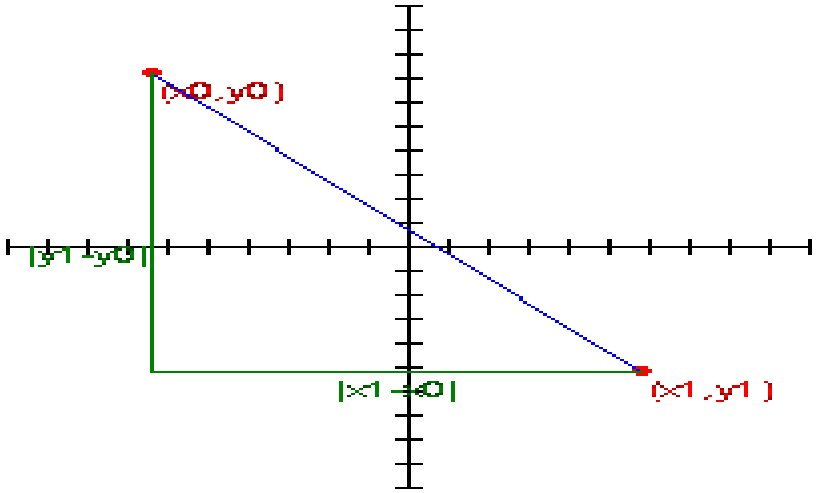


Figura 2.18: Bases de la distancia eucídea en el teorema de Pitágoras [8].

La distancia misma es la longitud de la hipotenusa (marcada en azul en la figura) y sus catetos (trazados en verde), que serían las proyecciones sobre los ejes coordenados de dicha recta, trasladados hasta los puntos en cuestión (marcados en rojo). El teorema se expresa en la ecuación (2.11), que luego queda en la conocida fórmula de la distancia euclidiana.

$$(A, B)^2 = \sqrt{(xB - xA)^2 + (yB - yA)^2} \tag{2.11}$$

Inconvenientes

La métrica euclidiana, pese a ser la más simple de las distancias a determinar y calcular por su relación con otros resultados bien conocidos de las matemáticas; presenta evidentes inconvenientes de aplicación fuera de espacios donde la línea recta sea la menor distancia que conecta a dos puntos. Un ejemplo más que evidente, es nuestra propia Tierra. La forma esférica de la misma y la incapacidad de viajar en línea recta porque habría que hacerlo por debajo del suelo, impiden el uso de este tipo de distancia sobre el planeta. Para ello existen otras métricas como la esférica e incluso, otras más abstractas para casos más complejos. En el particular de una esfera, la distancia más corta entre dos puntos es el arco que los une [8].

2.4.6. La regla de los K vecino más cercanos

La regla de los k vecino más cercanos K-NN del inglés K-Nearest Neighbors, es otro clasificador supervisado basado en reconocimiento de patrones y criterios de vecindad. También se conoce como algoritmo de clasificación K-NN. Parte de la idea de que una nueva muestra será clasificada a la clase a la cual pertenezca, la mayor cantidad de vecinos más cercanos (reconocimiento de patrones) del conjunto de entrenamiento más cercano a ésta [9].

Regla K-NN

Al aplicar la regla Nearest Neighbors (NN), se explora todo el conocimiento almacenado en el conjunto de entrenamiento para determinar cuál será la clase a la que pertenece una nueva muestra, pero únicamente tiene en cuenta el vecino más próximo a ella, por lo que es lógico pensar que es posible que no se esté aprovechando de forma eficiente toda la información que se podría extraer del conjunto de entrenamiento. Con el objetivo de resolver esta posible deficiencia surge la regla de los K-vecinos más cercanos (K-NN). La regla K-NN es una extensión de la regla NN, en la que se utiliza la información suministrada por los K prototipos del conjunto de entrenamiento más cercanos de una nueva muestra para su clasificación [9].

Definición de vecindad

Sea un conjunto de entrenamiento de N prototipos pertenecientes a M clases distintas, el reconocimiento de patrones espacio de representación de los objetos y una muestra x. La regla K-NN determina la clase a la cual pertenece una nueva muestra x, siendo la más votada por sus K vecinos más cercanos.

Ejemplo:

En la Figura 2.19 se ilustra el funcionamiento de esta regla de clasificación. En ella se encuentran representadas 12 muestras pertenecientes a dos clases distintas: la clase 1 está formada por 6 cuadrados de color azul y la clase 2 esta formada por 6 círculos de color rojo. En este ejemplo, se han seleccionado tres vecinos, es decir, (K=3) [9].

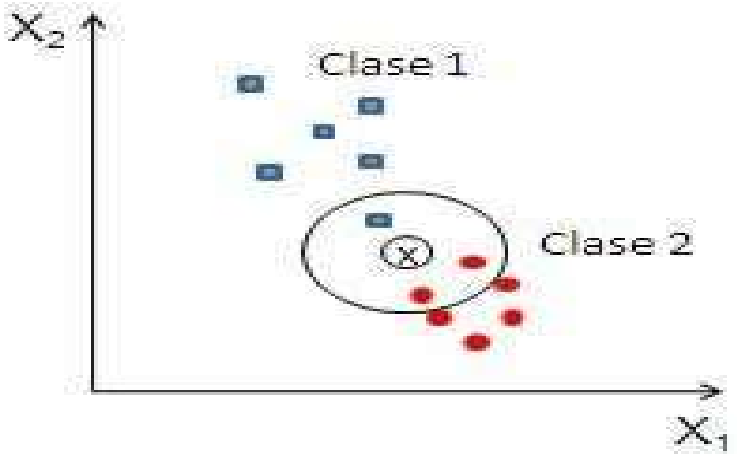


Figura 2.19: Funcionamiento de K-NN, el círculo es la vecindad, los circulos y cuadrados son datos de dos clases diferentes [9].

De los 3 vecinos más cercanos a la muestra x representada en la Figura 2.19 por una cruz, uno de ellos pertenece a la clase 1 y los otros dos a la clase 2. Por tanto, la regla 3-NN asignará la muestra x, a la clase 2. Es importante señalar, que si se hubiese utilizado como regla de clasificación la NN solamente, la muestra x, sería asignada a la clase 1, pues el vecino más cercano de la muestra x, pertenece a la clase 1 [9].

2.4.7. Ventajas y limitantes

En problemas prácticos, donde se aplica esta regla de clasificación, se acostumbra a tomar un número K de vecinos impar para evitar posibles empates, aunque esta forma es cierta en problemas que poseen dos clases nada más. También, los empates pueden ser resueltos decidiendo aleatoriamente la clasificación de la muestra entre las clases empatadas o la clase donde la distancia media de sus vecinos sea inferior. Para determinados problemas reales (es decir, con un número finito de muestras e incluso, en muchas ocasiones, un número relativamente pequeño), la aplicación de esta regla podría entenderse como una solución poco apropiada, debido a los pobres resultados que pudieran obtener, es decir, a su baja tasa de aciertos en el correspondiente proceso de clasificación. Este problema también está presente, cuando el número de muestras de que se dispone puede considerarse pequeño com-

parado con la dimensionalidad intrínseca del espacio de representación, lo cual corresponde a una situación bastante habitual [9].

2.5. Teorema de Takens

Varios fenómenos complejos a menudo se modelan como una secuencia de estados. Esta secuencia se conoce como estado espacial. Una serie de tiempo es una secuencia finita de estados que se midieron, directa o indirectamente, a partir de un sistema dinámico. El análisis de series de tiempo es relevante debido al teorema de incrustación de Takens, que establece que a partir de una secuencia de estados $S = y_{t_1}, y_{t_2}, \ldots, y_{t_n}$ (es decir, series de tiempo) de un sistema dinámico, es posible generar el estado espacial de todos los sistemas. Más detalladamente, para una secuencia de observaciones x de la dimensión m (dimensión de incrustación) y una constante τ (retraso de tiempo), existe una función f tal como [29]:

$$S = y(t) = f(x) = f[y_{(t-\tau)} \cdots y_{(t-(m-1)\ tau)}] \tag{2.12}$$

De la ecuación (2.12) podemos inferir que, dada una serie temporal S, es posible predecir el estado en el tiempo t (en adelante, y_t) usando m observaciones anteriores muestreadas en frecuencia τ. El principal problema es que la función $f(x)$ suele ser demasiado compleja para ser analizada, aquí es donde entran en juego los algoritmos de aprendizaje automático

2.5.1. Resumen del capítulo

Este capítulo, concentra el contenido teórico que se utilizó en el presente trabajo de tesis, contenido importante para entender de que trata el tema o trabajo de investigación. El contenido teórico de este trabajo de tesis comprende conceptos y fórmulas de los sistemas fotovoltaicos, series de tiempo, modelos probabilísticos, métodos de aprendizaje de máquina y el teorema de Takens; en el caso del sistema fotovoltaico interconectado a la red se describieron sus partes, también se explicó en que consistia y donde pueden ser instalados; se describieron las series de tiempo, también se describieron los modelos probabilísticos y los métodos de aprendizaje de máquina y su importancia en las series de tiempo, además

se hizo una mención del teorema de Takens el cual nos puede ser útil a la hora de optimizar una predicción; con toda esta información podemos comprender mejor este trabajo.

Capítulo 3

Pruebas y resultados del modelo ARIMA, K-NN y MSV

3.1. Introducción

Existen diversos modelos para hacer predicción, tales como: AR(p), ARMA, ARIMA, etc; también existen métodos basados en aprendizaje de máquina para hacer predicción, tales como: máquina de soporte vectorial y K-vecinos más cercanos, etc. En este Capítulo, se utilizaron los métodos ARIMA, K-vecinos más cercanos y la máquina de soporte vectorial para hacer predicción del voltaje generado por un sitema fotovoltaico, así como también la predicción de la temperatura y la radiación solar en el ambiente, provistas por una estación meteorológica. Se realizó predicción para diferentes horizontes, tales como, 1 hora, 2 horas y hasta un día para hacer esto, se utilizaron las funciones de los software Phyton y Matlab, además de hacer predicción usando estos métodos, también se hizo una comparación entre ellos, con el objetivo de determinar cual puede ser mejor para predecir usando series de tiempo relacionadas con la generación fotovoltaica, además de poder conocer mejor cuales pueden ser sus ventajas y desventajas al usar pocos y muchos datos en las series de tiempo. En este trabajo de tesis usaremos el teorema de Takens, con el fin de mejorar los resultados de predicción al emplear este teorema en conjunto con el método de aprendizaje de máquina con mejores resultados en las pruebas establecidas.

3.2. Descripción del sistema de prueba

El diagrama unifilar del sistema de prueba y la simbología de este se muestran en las Figura 3.1 y 3.2, la Figura 3.1 corresponde a un sistema fotovoltaico interconectado a la red, el cual proporciona energía al laboratorio de ingeniería eléctrica del departamento de posgrado, pertenenciente a la facultad de ingeniería eléctrica, de la UMSNH en la ciudad de Morelia, Michoacán, México. El sistema está compuesto de 36 paneles fotovoltaicos monocristalinos, donde cada panel contiene 60 celdas que en total generan 250 W. Los paneles se encuentran divididos en 3 grupos de 12, conectados en serie cada uno de ellos, como se puede apreciar en la Figura 3.1, generando en promedio, una cantidad de 3000 W a 400 V y una corriente de 8.62 A por grupo. Cada grupo se conecta a un fusible interruptor, y después se conecta a su respectivo convertidor de 400 VCD/220 VCA, las cuales se conectan a sus respectivas protecciones de CD y de ahí a la alimentación del laboratorio y a CFE mandando el excedente de energía a este último.

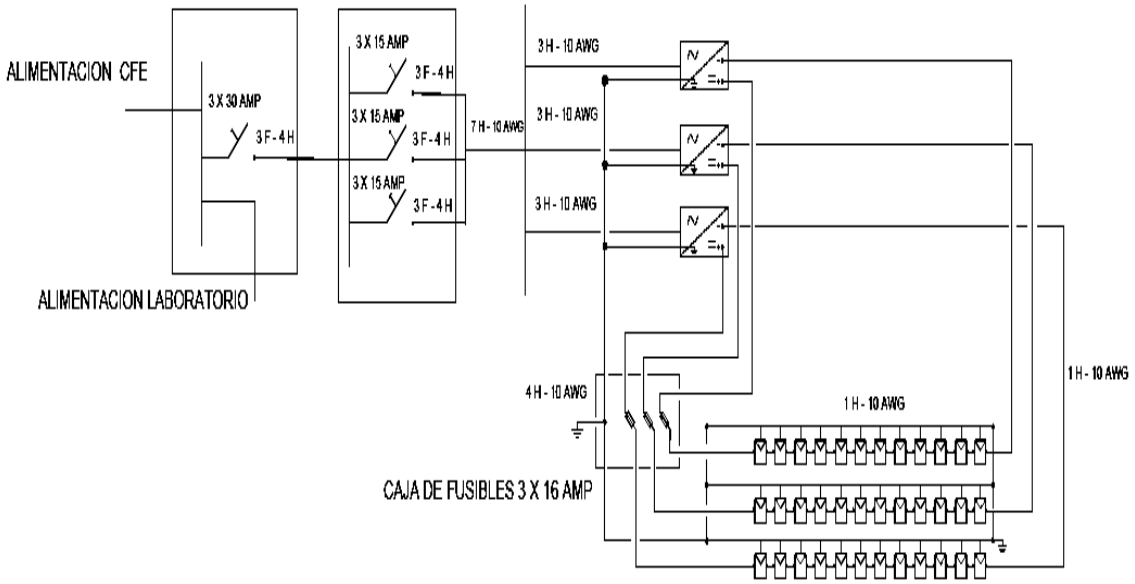


Figura 3.1: Diagarama unifilar de los paneles instalados en el laboratorio de posgrado de eléctrica (sistema de prueba).

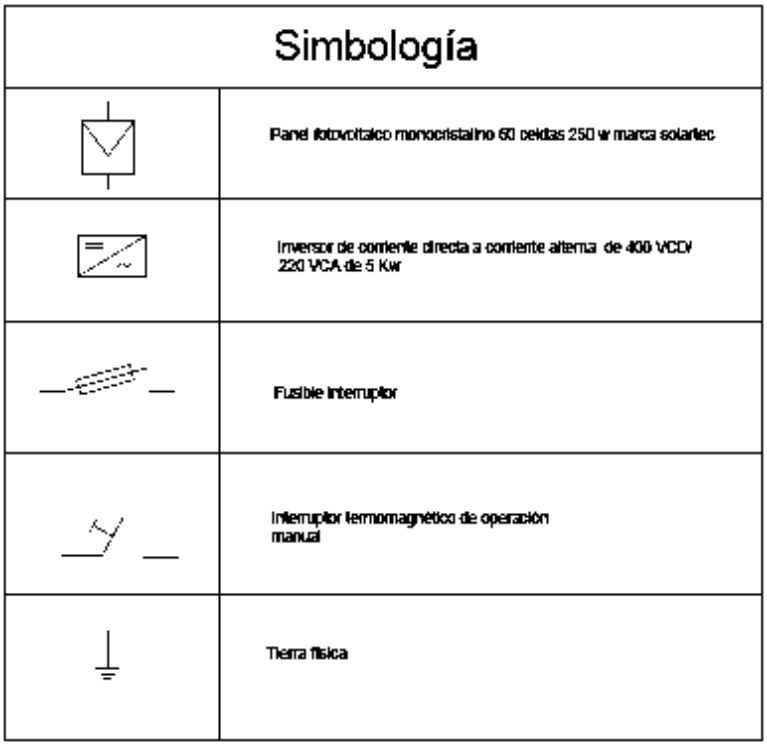


Figura 3.2: Simbología utilizada en el sistema de prueba.

3.2.1. Bases de datos proporcionadas por el sistema fotovoltaico y por el centro meteorológico

Nuestras bases de datos fueron proporcionadas por el centro meteorológico y por el sistema fotovoltaico mostrado en la Figura 3.1. Algunos de los datos proporcionados por el centro meteorológico son: radiación solar, velocidad del viento, humedad, presión, etc; algunos de los datos proporcionados por el sistema fotovoltaico son: temperatura, voltaje generado, potencia generada, frecuencia, nivel de CO2, etc. Cabe destacar que los datos del sistema fotovoltaico fueron proporcionados por la página web del fabricante de los inversores, también que los datos proporcionados por el sistema fotovoltaico se encuentran espaciados uno de otro cada 5 minutos mientras que los datos de la estación meteorológica se encuentran espaciados uno del otro cada 1 minuto.

Los datos utilizados para el estudio fueron:

- Voltaje generado.
- Temperatura.
- Radiación solar.

Estos datos o series de tiempo fueron seleccionados debido a su comportamiento e importancia dentro de un sistema fotovoltaico. El voltaje generado es importante ya que es la generación del sistema fotovoltaico, la radiación solar y la temperatura también son importantes debido a su influencia con la potencia generanerada y el voltaje generado. Los datos de voltaje generado y de temperatura, proporcionados por los tres inversores en la página del fabricante, no presentaban diferencias significativas, más sin embargo se requería tomar un conjunto de datos de alguno de los tres inversores y se eligieron los datos del tercer inversor.

3.2.2. Comportamiento de las series de tiempo: voltaje, temperatura y radiación solar

Los datos de voltaje generado, temperatura y radiación solar, se graficáron de acuerdo a como se usaron en las pruebas(mes y año.) establecidas, así que tales datos se graficaron en los siguientes conjuntos:

- Mes.
- Año.

El conjunto mes corresponde a febrero del 2017, y el conjunto año corresponde al año 2017. Se tomó el mes de febrero por la cantidad y continuidad de sus datos, ya que para otros meses los datos de temperatura y voltaje generado no estaban completos (en los datos de radiación solar principalmente). La Figura 3.3 muestra el voltaje generado durante el mes de febrero del 2017 el cual se comprende de 1978 muestras, mientras que las Figuras 3.4 y 3.5 muestran la temperatura y radiación solar registrada durante todo el año 2017,

la serie de tiempo de temperatura se comprende de 17680 muestras, mientras que la serie de tiempo de radiación solar se comprende de 450173 muestras. La Figura 3.6 presenta la radiación solar sin tomar en cuenta los ceros, es decir, sin tomar en cuenta las horas en las cuales no hay generación, que viene siendo relativamente de las 7 am hasta las 7 pm, esta serie de tiempo comprende de 236204 muestras.

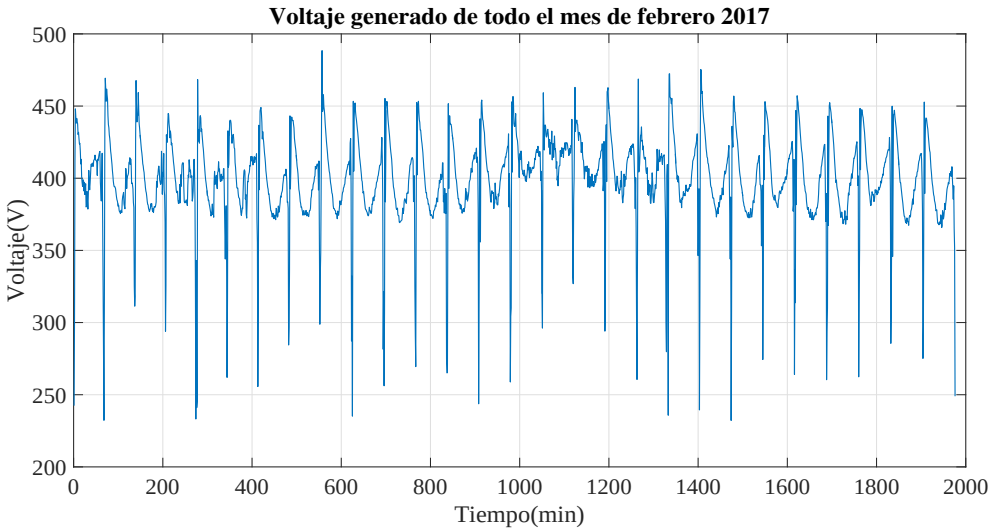


Figura 3.3: Voltaje generado de todo el mes de febrero del año 2017.

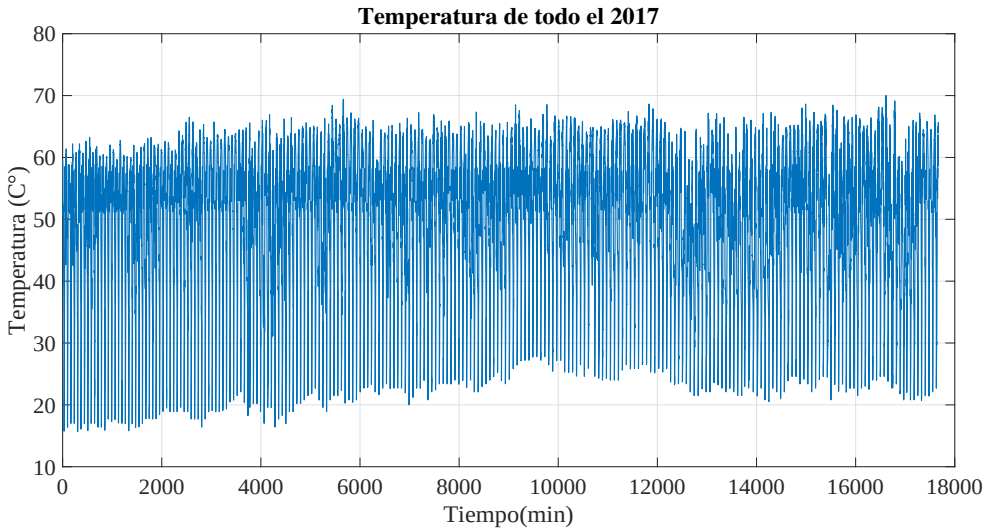


Figura 3.4: Temperatura de todo el año 2017.

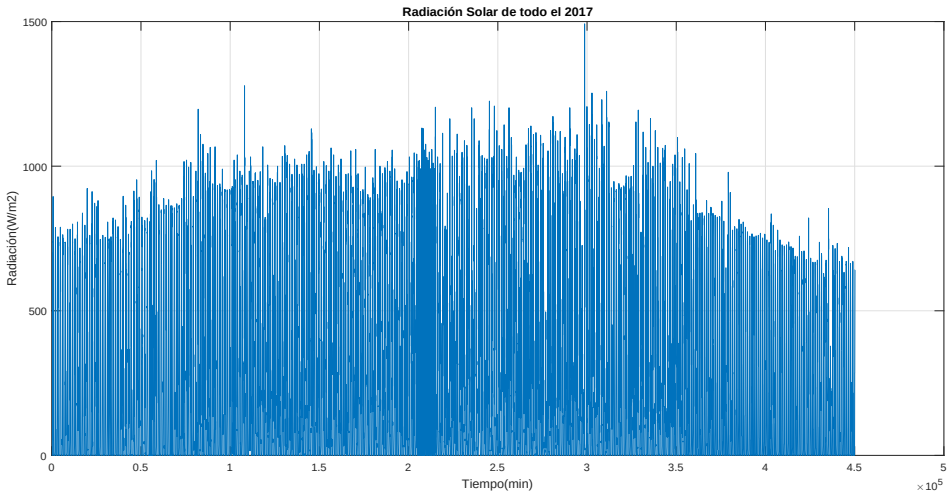


Figura 3.5: Radiación solar de todo el año 2017.

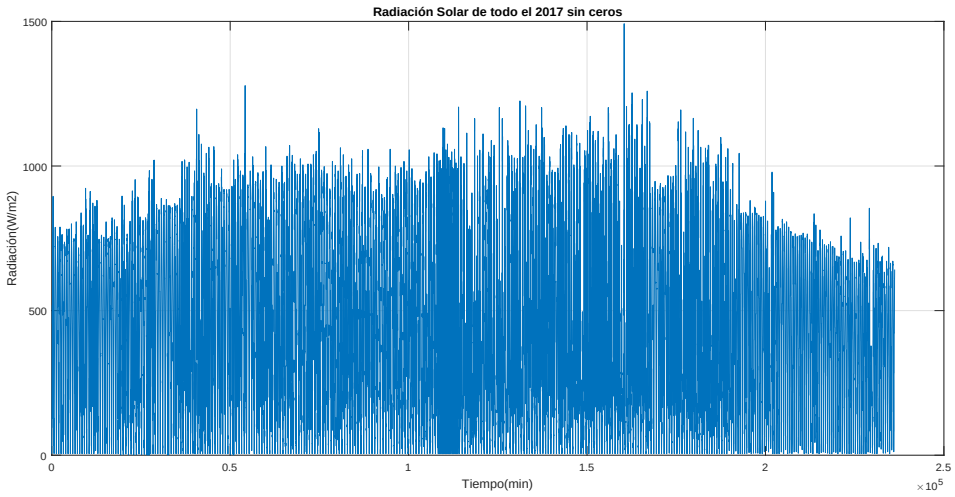


Figura 3.6: Radiación solar de todo el año 2017 (sin ceros).

3.3. Pruebas para encontrar la mejor configuración de los métodos K-NN, MSV y ARIMA para las diferentes series de tiempo.

Se realizaron una serie de pruebas a algunas configuraciones propias de los métodos basados en aprendizaje de máquina (K-NN y MSV), así como también del modelo ARIMA. Para dichas pruebas se usó la serie de tiempo correspondiente a la temperatura de todo el año 2017, la cual comprende de 17680 muestras, con el objetivo de obtener la mejor configuración de cada uno de estos métodos de predicción. Una vez obtenidas las mejores configuraciones, se utilizarán para hacer unas pruebas más especificas, las cuales se mostraran más adelante. En estas primeras pruebas se usaron 17668 muestras para el entrenamiento de los 3 métodos de predicción, lo cual equivale a un 99.9322 % de la serie de tiempo, con esto se predijo una hora, que equivale a 12 muestras de la serie de tiempo, el 0.0678 % restante de la serie de tiempo original (12 muestras), no lo vieron los 3 métodos de predicción, estas 12 muestras faltantes se usarón para comparar las 12 muestras predecidas, esto con el fin de hacer predicción y poder comparar resultados y con esto calcular un error de predicción, cabe destacar que se tomaron 12 muestras debido a que cada dato se encuentra espaciado uno de otro cada 5 minutos, por lo tanto, 12 muestras equivalen a una hora de la serie de tiempo, también se usó un rango de una hora, debido a que es un tiempo de predicción bueno, abre la posibilidad para estar preparado y tomar alguna acción en el sistema o externa a el. Más adelante se hará predicción con rangos de tiempo más grandes.

3.3.1. Pruebas para el modelo ARIMA (p,d,q) con diferentes ordenes

Para las siguientes pruebas se utilizaron como configuraciones los siguientes ordenes para ARIMA: (1,1,1), (3,3,3) y (5,5,5), no se decidió aumentar a más de 5 el orden para p,d y q, debido a que el costo computacional sería muy alto, admás que el software comenzaba a tener problemas, también se decidió dejarlos igual. Las Figuras 3.7, 3.8 y 3.9 muestran los resultados del modelo ARIMA con ordenes (1,1,1), (3,3,3) y (5,5,5), respectivamente.

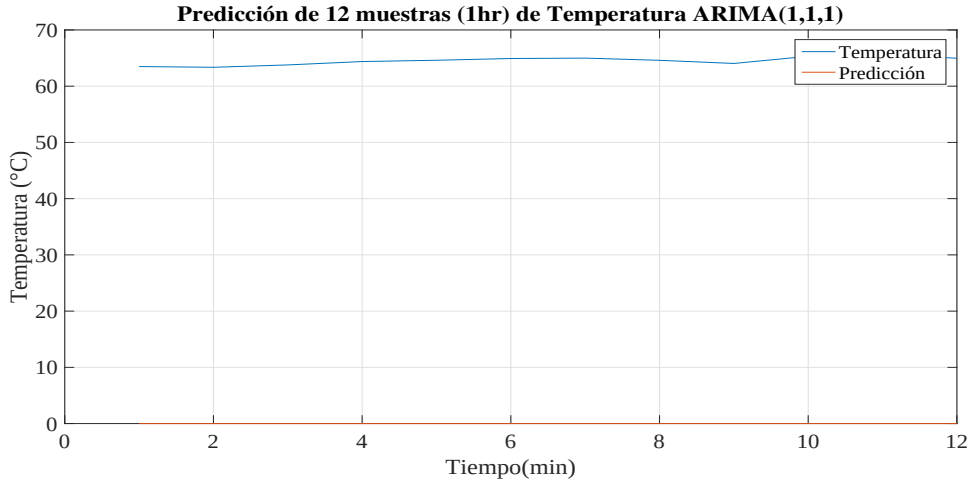


Figura 3.7: Resultados de un ARIMA (1,1,1) prediciendo una hora de un año (2017) de temperatura.

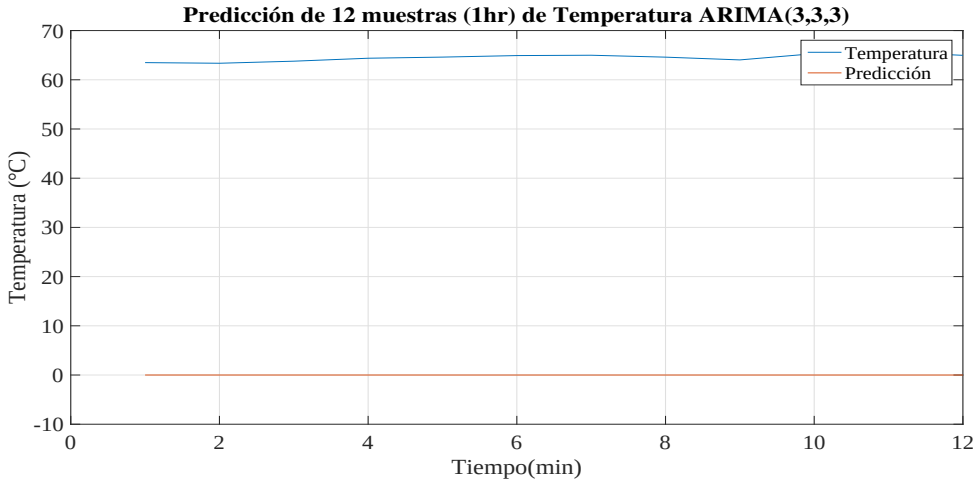


Figura 3.8: Resultados de un ARIMA (3,3,3) prediciendo una hora de un año (2017) de temperatura.

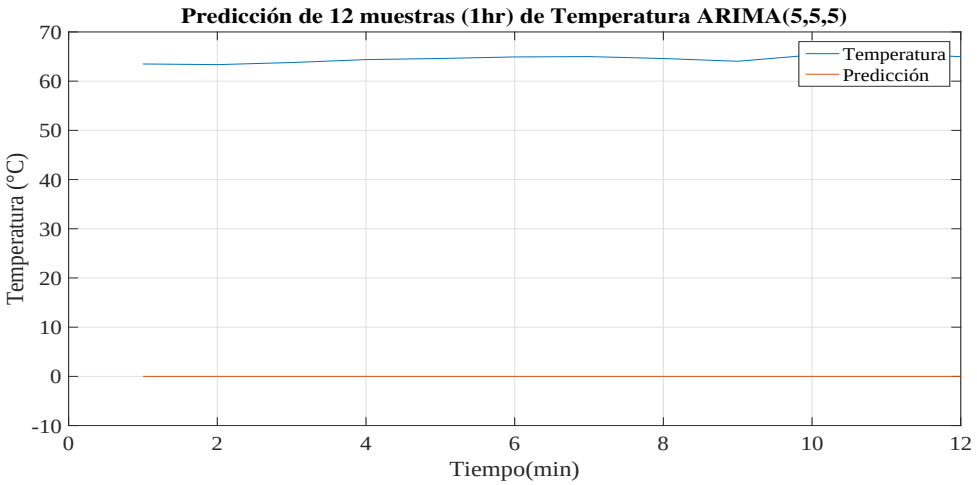


Figura 3.9: Resultados de un ARIMA (5,5,5) prediciendo una hora de un año (2017) de temperatura.

Los resultados para las 3 diferentes configuraciones del modelo ARIMA utilizadas no son tan buenos, puesto que los errores de predicción son muy grandes, (los errores de predicción para las diferentes configuraciones usadas del modelo ARIMA se encuentran en la tabla 3.9) además no sigue la forma de onda de las muestras originales, cabe destacar que la configuración de ARIMA (1,1,1) tiene un menor error que las otras configuraciones usadas, también que no es mucha la diferiencia pero aun así se tuvo que escoger esta configuración de ARIMA para las siguientes pruebas, más sin embargo se espera que mejoren los resultados de este modelo en las siguientes pruebas. Cabe mencionar que el error se calculó haciendo una diferencia entre las muestras originales con las muestras predecidas y después de eso se calculó un promedio con esas diferencias, por lo tanto el error no contempla si la predicción sigue a las muestras originales.

Rangos	ARIMA(1,1,1)	ARIMA(3,3,3)	ARIMA(5,5,5)
1 hora	64.50	64.51	64.51

Tabla 3.1: Errores en la predicción de diferentes configuraciones del modelo ARIMA.

3.3.2. Pruebas para K-NN para dos tipos de peso y diferentes tipos de K

Para las siguientes pruebas se utilizaron como configuraciones del método K-NN dos tipos de pesos, el peso distancia y el peso uniforme, también se utilizarón tres valores diferentes de K los cuales fueron: 7, 25 y 50, el método utiliza la distancia euclidea para relacionar los datos en base a un valor de K (K vecinos más cercanos). Las figuras 3.10, 3.11 y 3.12 muestran los resultados de las pruebas al método K-NN con peso igual a distancia con los diferentes valores de K utilizados (7, 25 y 50), las Figuras 3.13, 3.14 y 3.15 muestran los resultados de las pruebas al método K-NN con peso igual a uniforme con los diferentes valores de K utilizados (7,25 y 50).

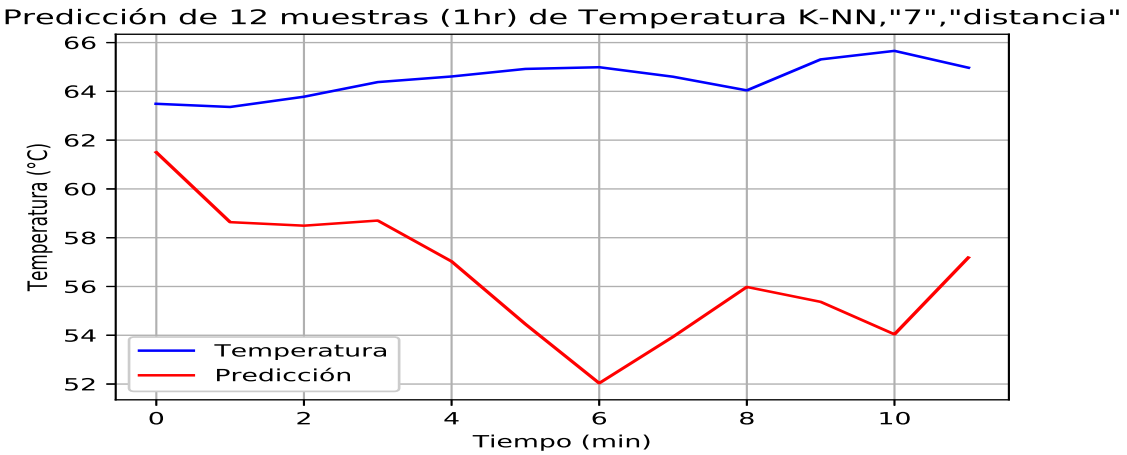


Figura 3.10: Resultados de K-NN con peso = distancia y K=7 prediciendo una hora de un año (2017) de temperatura.

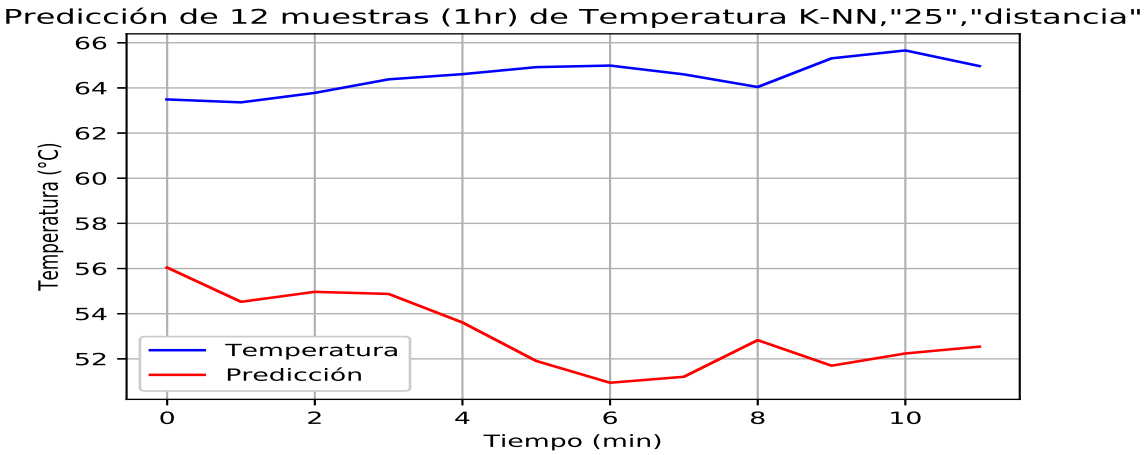


Figura 3.11: Resultados de K-NN con peso = distancia y K=25 prediciendo una hora de un año (2017) de temperatura.

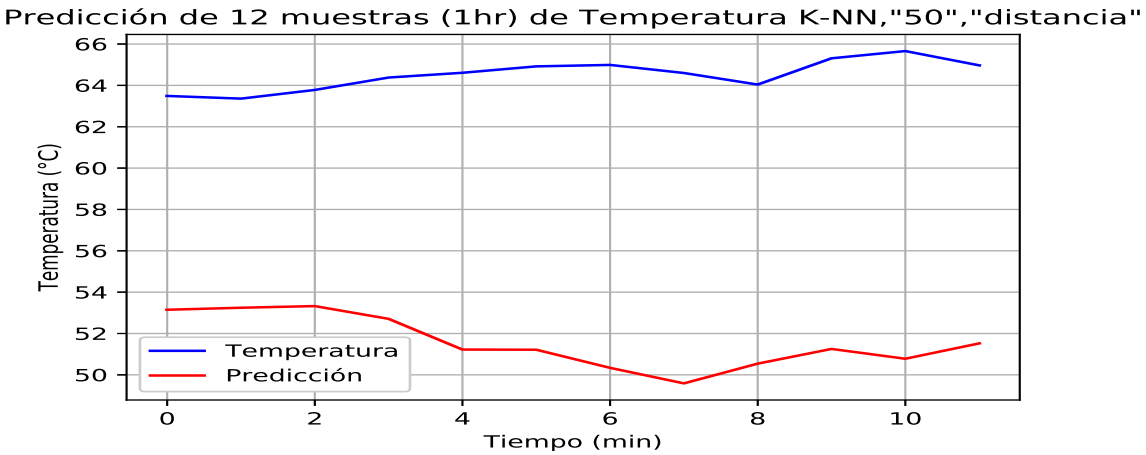


Figura 3.12: Resultados de K-NN con peso = distancia y K=50 prediciendo una hora de un año (2017) de temperatura.

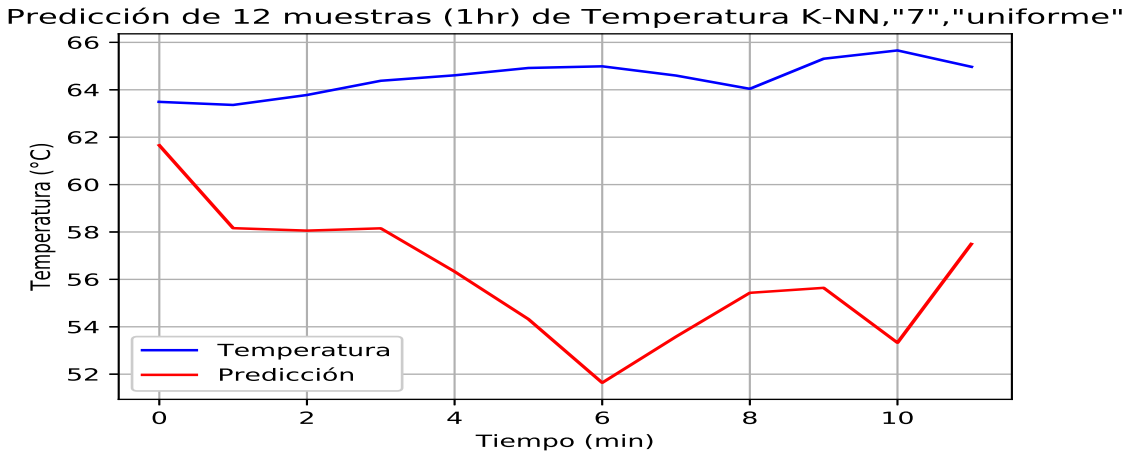


Figura 3.13: Resultados de K-NN con peso = uniforme y K=7 prediciendo una hora de un año (2017) de temperatura.

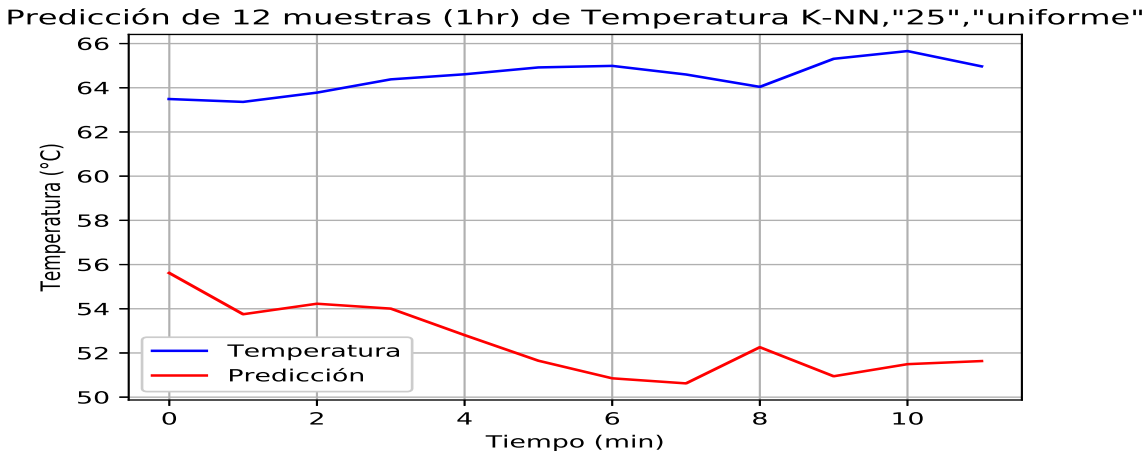


Figura 3.14: Resultados de K-NN con peso = uniforme y K=25 prediciendo una hora de un año (2017) de temperatura.

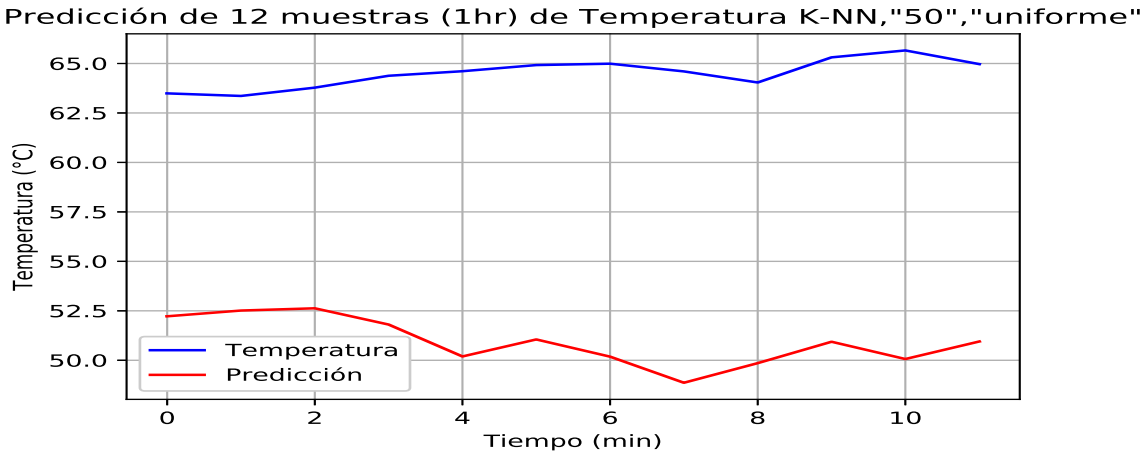


Figura 3.15: Resultados de K-NN con peso = uniforme y K=50 prediciendo una hora de un año (2017) de temperatura.

Los resultados de todas las configuraciones del método K-NN seleccionadas, así como en sus diferentes valores de K, son buenos, podemos ver en las gráficas anteriores que las muestras predecidas siguen muy bien la forma de onda de las muestras de la serie original, pero cabe destacar que los errores de predicción son menores en ambas configuraciones cuando se usa una k=7, (los errores de predicción para las configuraciones usadas del método K-NN se encuentran en la tabla 3.2) también usando el tipo de peso distancia, por lo tanto para las siguientes pruebas se usó un K-NN con tipo de peso distancia y con un valor de k=7.

Rangos	Valor de K	Uniforme	Distancia
1 hora	7	8.35	8.06
1 hora	25	12.02	11.39
1 hora	50	13.56	12.93

Tabla 3.2: Errores en la predicción de diferentes configuraciones de K-NN.

3.3.3. Pruebas para la MSV usando dos tipos de Kernel (gausiano y polinomial)

Para las siguientes pruebas se utilizaron como configuraciones del método MSV dos tipos de funciones kernel, la función Kernel del tipo gaussiano y la función Kernel del tipo polinomial (de grado 2), esto debido a que son unas de las funciones Kernel más utilizadas de las que más conincidien con nuestras series de tiempo. Las Figuras 3.16 y 3.17 muestran los resultados de las pruebas al método MSV utilizando un kernel de tipo gaussiano y un kernel de tipo polinomial (grado 2), respectivamente.

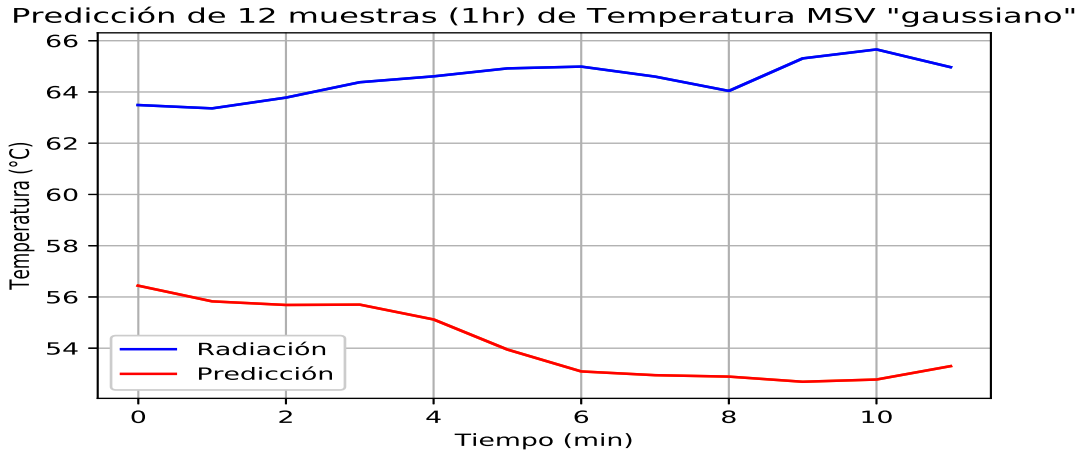


Figura 3.16: Resultados de la MSV con kernel gaussiano prediciendo una hora de un año (2017) de temperatura.

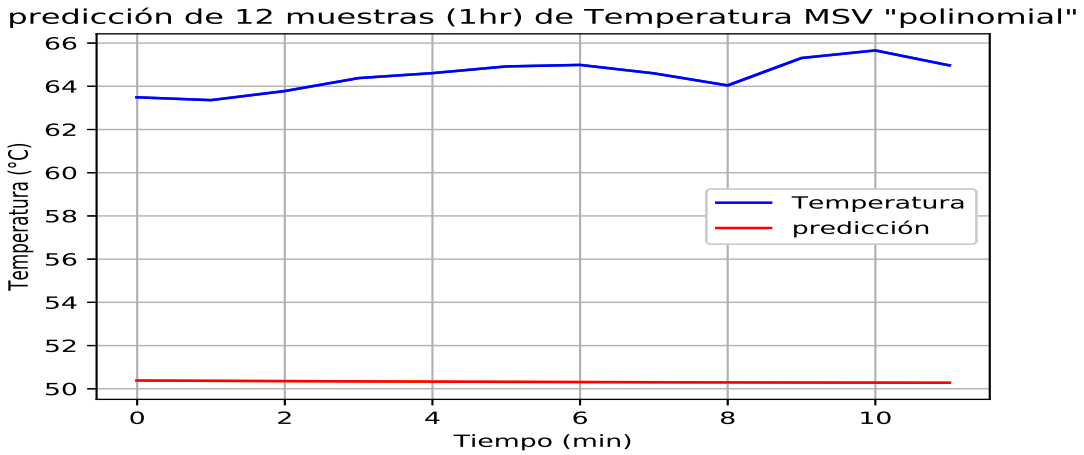


Figura 3.17: Resultados de la MSV con kernel polinomial prediciendo una hora de un año (2017) de temperatura.

Los resultados de las configuraciones usadas para el método MSV muestran que ambas funciones kernel predicen bien, llegando a obtener errores de predicción peque nos(los errores de predicción para las configuraciones usadas del método MSV se encuentran en la tabla 3.3), tamién podemos observar en las gráficas anteriores que la predicción obtenida con la función del tipo gaussiano sigue el comportamiento de las muestras de la serie original, por lo tanto se utilizará para las siguientes pruebas un kernel de tipo gaussiano.

Rangos	Kernel gaussiano	Kernel polinomial
1 hora	10.30	14.18

Tabla 3.3: Errores en la predicción de diferentes configuraciones de MSV.

3.4. Pruebas para las configuraciones seleccionadas de ARIMA, K-NN y MSV

Para las pruebas siguientes se utilizaron las mejores configuraciones de los métodos de predicción ARIMA, K-NN y la MSV con los mejores resultados en las pruebas anteriores, las cuales fueron, un modelo ARIMA de orden (1,1,1), un método K-NN con una K=7 y un peso igual a distancia y la MSV con una función Kernel de tipo gaussiano. Para cada método se hicieron predicciones equivalentes a 1 hora, 2 horas y un día. Las pruebas consistieron en hacer predicción en dos tipos de casos, el caso A y el caso B, el caso A se refiere al uso de pocos datos en la predicción (serie de tiempo de voltaje generado del mes de febrero del 2017) y el caso B se refiere al uso de muchos datos en la predicción (serie de tiempo de radiación solar de todo el año 2017).
Estas pruebas tienen el objetivo de:

- Conocer que tan efectiva puede ser la predicción de estos métodos, si se usan muchos o pocos datos(relativamente ya que la cantidad de datos es considerable), también de ver que tanto influye en la predicción, el comportamiento de las series de tiempo empleadas.
- Conocer que tanto se puede estirar el rango de predicción.
- Conocer que tan conveniente puede ser realizar predicción a series de tiempo relacionadas con la generación fotovoltaica como lo son la temperatura, el voltaje generado y la radiación solar.
- Conocer que método puede ser mejor.

3.4.1. Caso A (Pocos datos)

En este caso se realizaron pruebas de predicción al modelo ARIMA de orden (1,1,1), al método K-NN con peso igual a distancia y una K=7 y al método MSV con una función kernel de tipo gaussiano, se realizaron predicciones de 1 hora, 2 horas y un día. Se utilizó una serie de tiempo de voltaje generado durante todo el mes de febrero del 2017, la cual consta

de 1978 muestras, las cuales son menos con respecto a la serie de tiempo de temperatura usada en las primeras pruebas, la cual constó de 17668 muestras; de la serie de tiempo de un mes de voltaje generado se usaron para 1 hora un 99.4 % para entrenamiento y un 0.6 % para validar la predicción, que vienen siendo 12 muestras, para 2 horas un 98.79 % para entrenamiento y un 1.21 % para validar la predicción, que vienen siendo 24 muestras y para 1 día un 85.44 % para entrenamiento y un 14.56 % para validar la predicción, que vienen siendo 288 muestras, cabe destacar que los datos de voltaje generado se encuentran espaciados 5 minutos uno del otro, por lo tanto, 1 hora serían 12 muestras, 2 horas serían 24 muestras y un día serían 288 muestras.

ARIMA caso A

Las siguientes pruebas con el modelo ARIMA se hicieron con un orden (1,1,1) esto debido a que fue el orden que presentó mejores resultados. Las Figuras 3.18, 3.19 y 3.20 muestran los resultados del modelo ARIMA de orden (1,1,1) en el caso A, prediciendo 1 hora, 2 horas y un día, respectivamente.

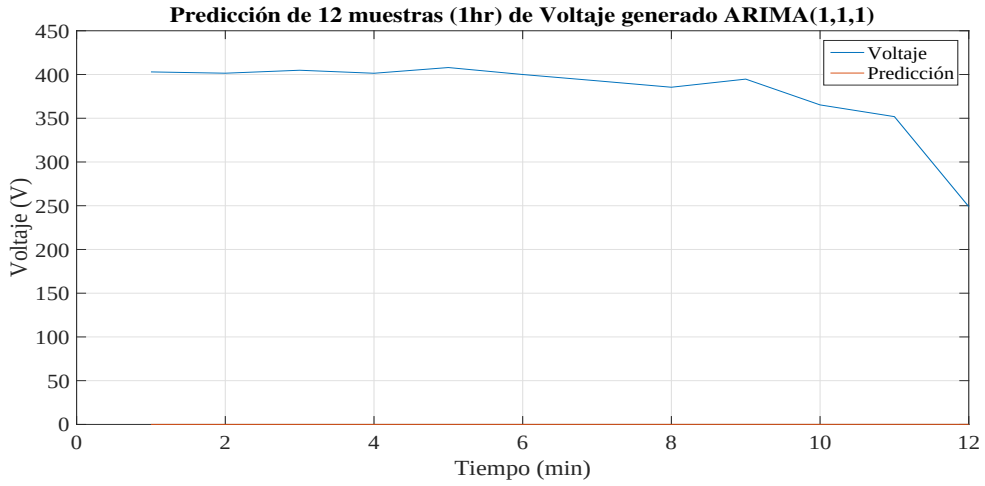


Figura 3.18: Resultados de la predicción de una hora de voltaje generado del mes de febrero del año 2017 con modelo ARIMA de orden (1,1,1), caso A.

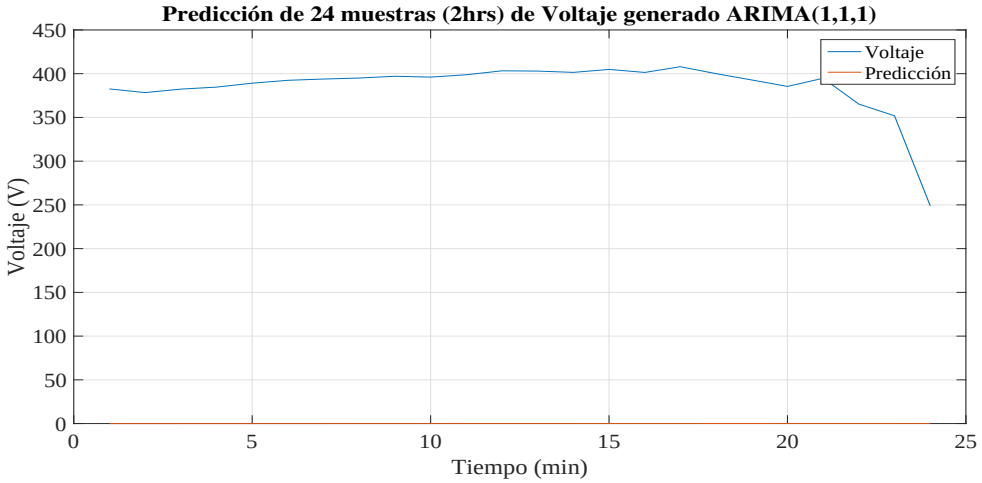


Figura 3.19: Resultados de la predicción de dos horas de voltaje generado del mes de febrero del año 2017 con el modelo ARIMA de orden (1,1,1), caso A.

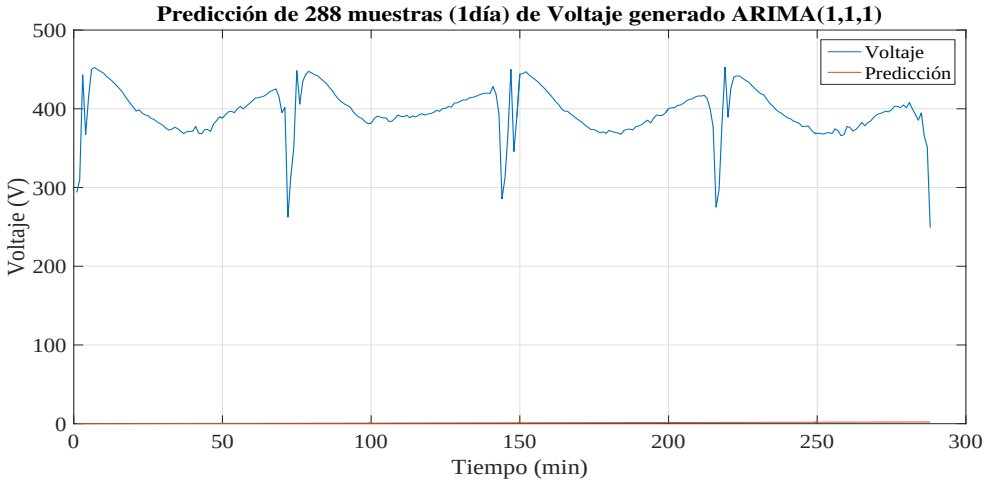


Figura 3.20: Resultados de la predicción de un día de voltaje generado del mes de febrero del año 2017 con el modelo ARIMA de orden (1,1,1), caso A.

Los resultados de estas pruebas obtuvieron errores de predicción de hasta 400 unidades (los errores de predicción para el modelo ARIMA con orden (1,1,1) para el caso A, se encuentran en la tabla 3.4), lo cual nos indica que las predicciones no son muy buenas puesto que los errores de predicción son muy grandes, además en las gráficas anteriores podemos observar que las muestras predecidas no siguen el comportamiento de las muestras

de la serie original. Los errores de predicción y el nulo seguimiento de las muestras predecidas se puede deber a que son demasiadas muestras para que el modelo de orden (1,1,1) pueda calcular sus parámetros para ajustar bien a la serie de tiempo empleada y con ello poder hacer una buena predicción, pero aun así veremos más adelante los resultados del caso B y reforzaremos esta idea del porqué una mala predicción.

Rangos	ARIMA (1,1,1)
1 hora	379.79
2 horas	391.11
1 día	412.77

Tabla 3.4: Errores en la predicción del modelo ARIMA (1,1,1) para el caso A.

K-NN caso A

Las siguientes pruebas con el método K-NN se hicieron con un valor de K=7 y un peso igual a distancia, esto debido a que fue la configuración que presentó mejores resultados. Las Figuras 3.21, 3.22 y 3.23 muestran los resultados de las pruebas al método K-NN con peso igual a distancia y un valor de K=7, en el caso A prediciendo 1 hora, 2 horas y un día, respectivamente.

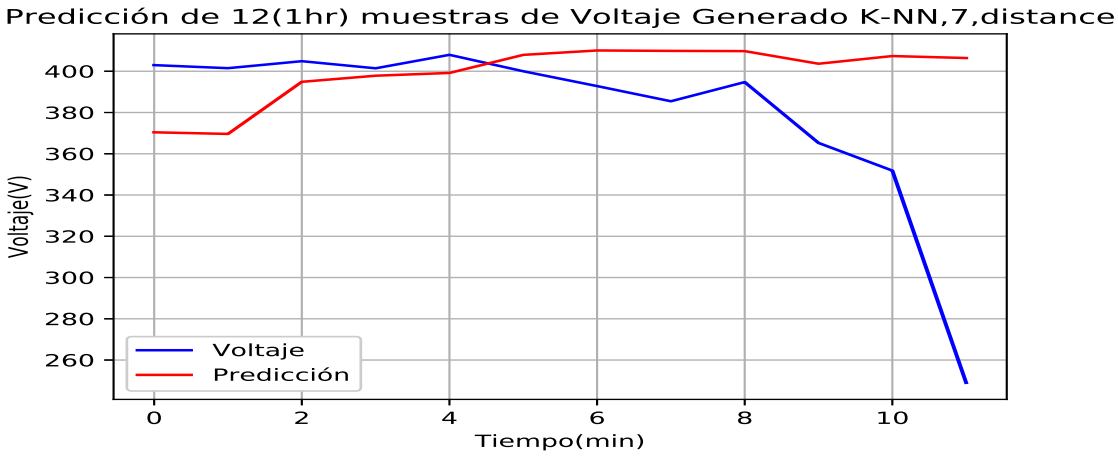


Figura 3.21: Resultados de la predicción de 1 hora de voltaje generado del mes de febrero del año 2017 con K-NN, con peso = distancia y K=7, caso A.

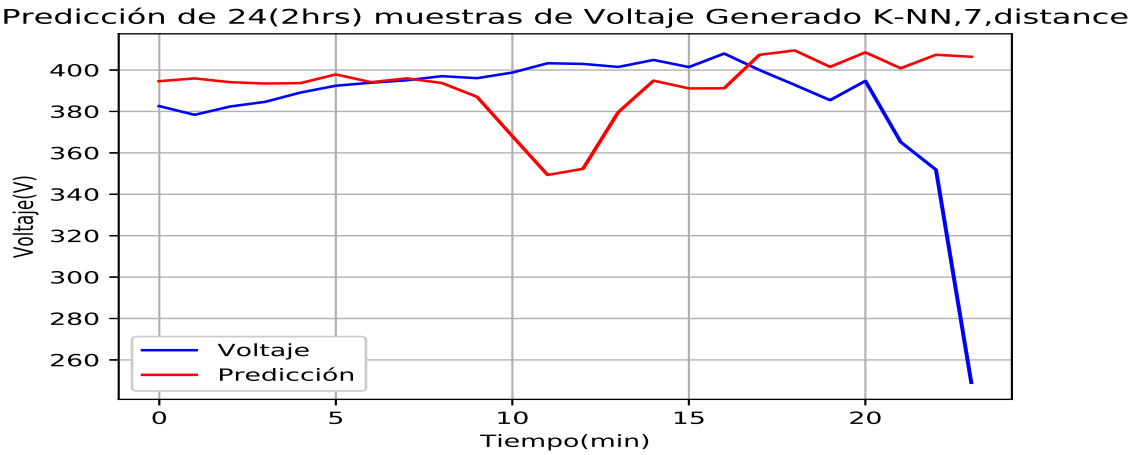


Figura 3.22: Resultados de la predicción de 2 horas de voltaje generado del mes de febrero del año 2017 con K-NN, con peso = distancia y K=7, caso A.

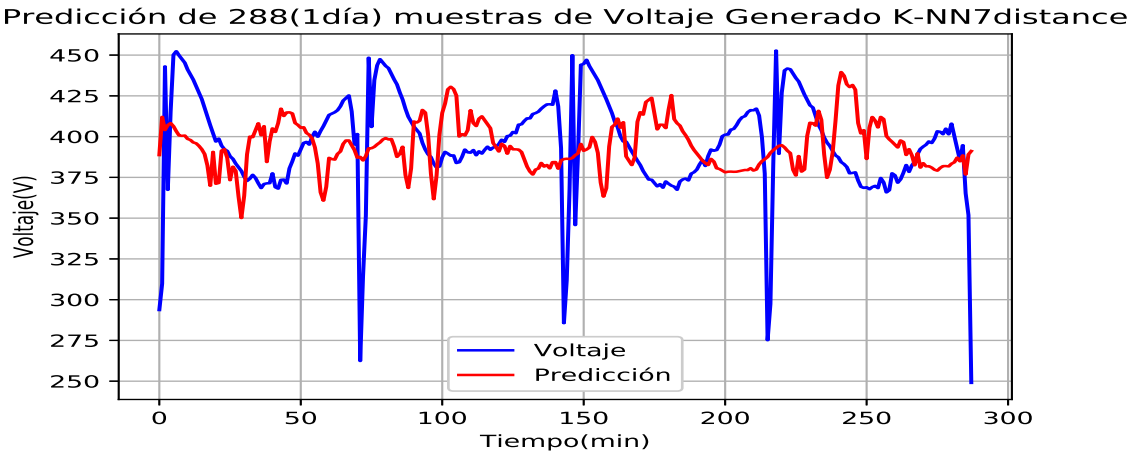


Figura 3.23: Resultados de la predicción de 1 día de voltaje generado del mes de febrero del año 2017 con K-NN, con peso = distancia y K=7, caso A.

Los resultados de estas pruebas obtuvieron errores de predicción de alrededor de 30 unidades (los errores de predicción para un método K-NN con K=7 y peso igual a distancia para el caso A, se encuentran en la tabla 3.5), lo cual es bajo, además podemos observar en las gráficas anteriores que las muestras predecidas siguen regularmente bien a las muestras originales la mayor parte del tiempo, por lo tanto, podemos observar que los resultados de

predicción son muy buenos. Ahora veremos en las pruebas del caso B que tanto se mantienen los buenos resultados de predicción de este método.

Rangos	K-NN (K=7,peso=distancia)
1 hora	33.56
2 horas	23.77
1 día	28.89

Tabla 3.5: Errores en la predicción del método K-NN con k=7 y peso igual a distancia para el caso A.

3.4.2. MSV caso A

Las siguientes pruebas con el método MSV se hicieron con una función kernel del tipo gaussiano, esto debido a que fue la configuración que presentó mejores resultados. Las Figuras 3.24, 3.25 y 3.26 muestran los resultados de la MSV con el kernel gaussiano, en el caso A, prediciendo 1 hora, 2 horas y un día, respectivamente.

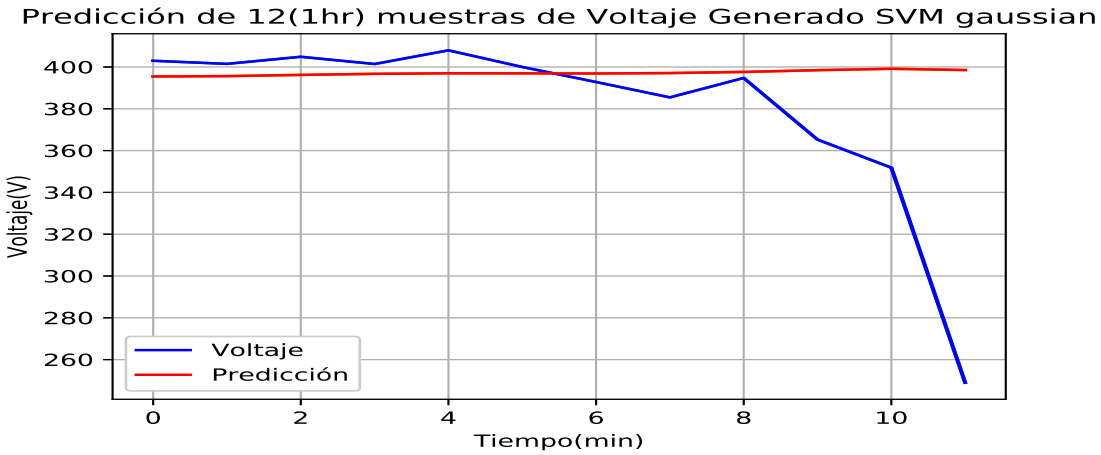


Figura 3.24: Resultados de la predicción de 1 hora de voltaje generado del mes de febrero del año 2017 con la MSV con kernel gaussiano, caso A.

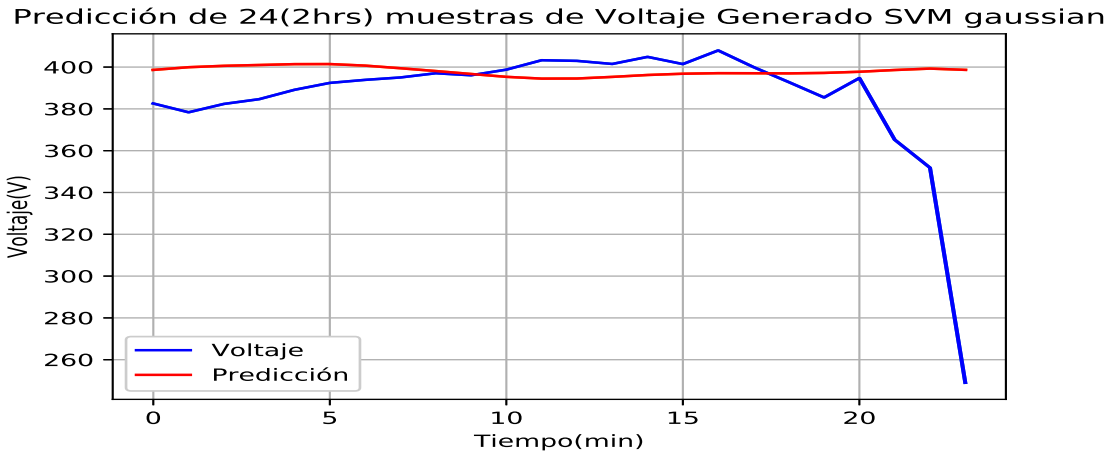


Figura 3.25: Resultados de la predicción de 2 horas de voltaje generado del mes de febrero del año 2017 con la MSV con kernel gaussiano, caso A.

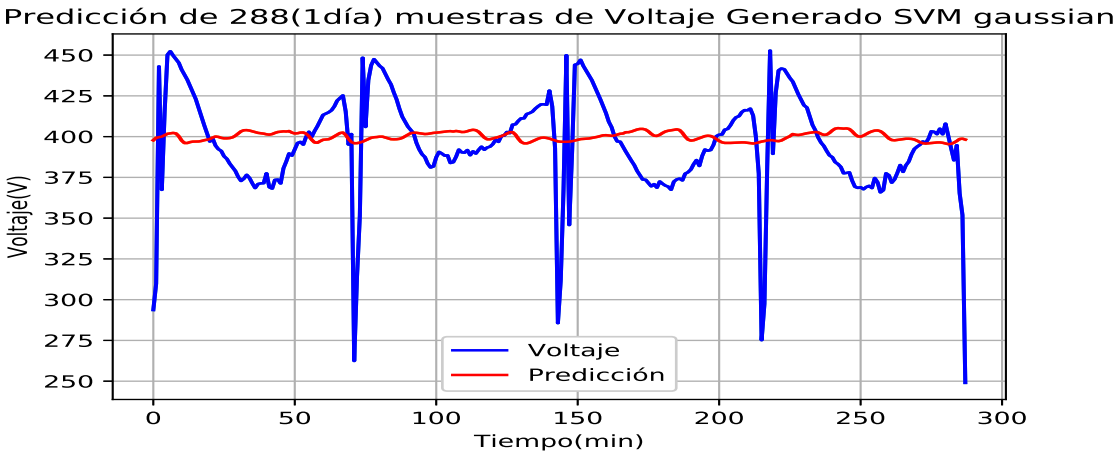


Figura 3.26: Resultados de la predicción de un día de voltaje generado del mes de febrero del año 2017 con la MSV con kernel gaussiano, caso A.

Los resultados de estas pruebas obtuvieron errores de predicción de alrededor de 25 unidades(los errores de predicción para un método MSV con kernel gaussiano para el caso A, se encuentran en la tabla 3.6), lo cual es bajo, además podemos observar en las gráficas anteriores que las muestras predecidas no siguen regularmente bien a las muestras originales la mayor parte del tiempo, por lo tanto, podemos observar que los resultados de

predicción son regulares, esto se puede deber a que la función kernel no se pueda ajustar bien a nuestra serie de tiempo empleada, por lo tanto, más adelante veremos en las pruebas del caso B, si realmente puede ser este el factor de un mal seguimiento o hasta de un aumento en los errores de predicción.

Rangos	MSV (kernel gaussiano)
1 hora	24.12
2 horas	17.07
1 día	22.90

Tabla 3.6: Errores en la predicción del método MSV con kernel gaussiano para el caso A.

3.4.3. Caso B

En este caso se realizaron pruebas de predicción al modelo ARIMA de orden (1,1,1), al método K-NN con peso igual a distancia y una K=7 y al método MSV con una función kernel de tipo gaussiano, se realizaron predicciones de 1 hora, 2 horas y un día. Se utilizó una serie de tiempo de radiación solar durante todo el año 2017, la cual constó de 236204, las cuales son mucho más con respecto a la serie de tiempo de temperatura usada en las primeras pruebas, la cual constó de 17668 muestras; de la serie de tiempo de un año de radiación solar se usaron para 1 hora un 99.97 % para entrenamiento y un 0.03 % validar la predicción, que vienen siendo 60 muestras, para 2 horas un 99.94 % para entrenamiento y un 0.06 % para validar la predicción, que vienen siendo 120 muestras y para 1 día un 99.39 % para entrenamiento y un 0.61 % para validar la predicción, que vienen siendo 1440 muestras, cabe destacar que los datos de la serie de tiempo de radiación solar se encuentran espaciados 1 minuto uno del otro, por lo tanto, 1 hora serán 60 muestras, 2 horas serán 120 muestras y un día serían 1440 muestras.

ARIMA caso B

Las siguientes pruebas con el modelo ARIMA se hicieron con un orden (1,1,1). Las Figuras 3.27, 3.28 y 3.29 muestran los resultados del modelo ARIMA de orden (1,1,1) en el caso B, prediciendo 1 hora, 2 horas y un día, respectivamente.

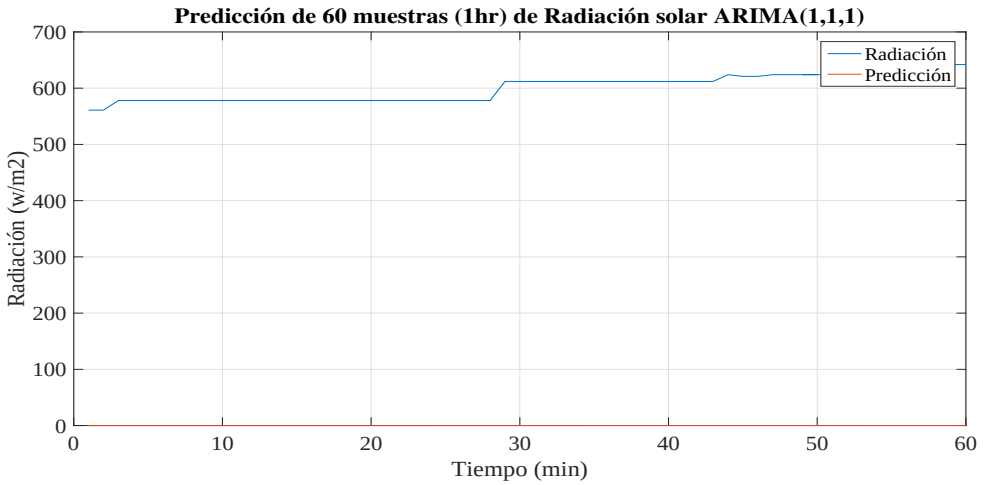


Figura 3.27: Resultados de la predicción de una hora de radiación solar del año 2017 con el modelo ARIMA de orden (1,1,1), caso B.

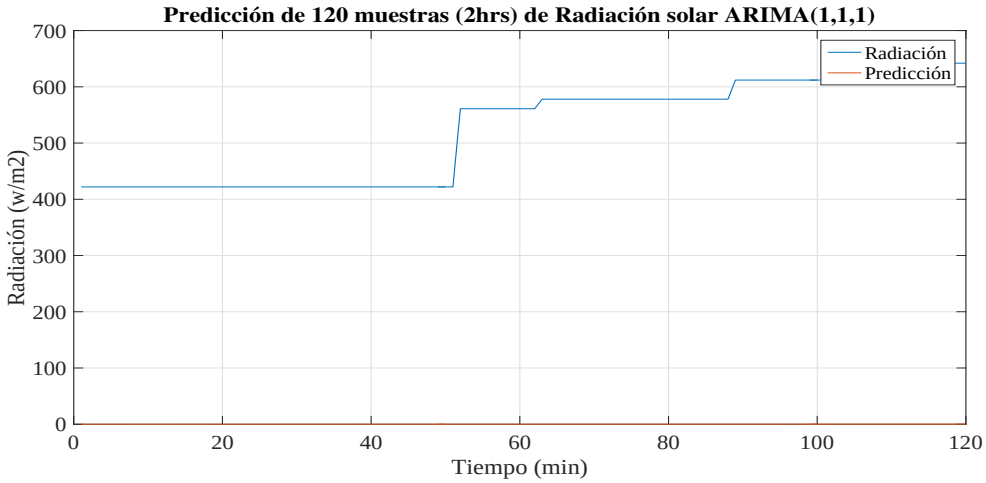


Figura 3.28: Resultados de la predicción de dos horas de radiación solar del año 2017 con el modelo ARIMA de orden (1,1,1), caso B.

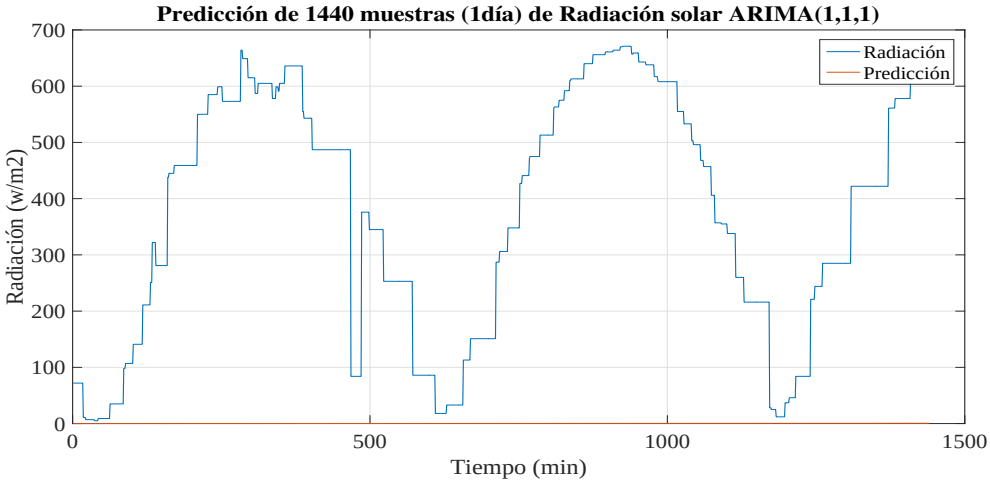


Figura 3.29: Resultados de la predicción de un día de radiación solar del año 2017 con el modelo ARIMA de orden (1,1,1), caso B.

Los resultados de estas pruebas obtuvieron errores de predicción de hasta 800 unidades (los errores de predicción para el modelo ARIMA con orden (1,1,1) para el caso B, se encuentran en la tabla 3.7), lo cual nos indica que las predicciones siguen sin ser buenas, puesto que los errores de predicción aumentaron bastante, además, en las gráficas anteriores podemos observar que las muestras predecidas continuan sin poder seguir el comportamiento de las muestras de la serie original. En base a los resultados obtenidos en ambos casos, podemos estar más seguros que la mala predicción puede deberse a la forma de las series de tiempo empleadas en estas pruebas y a la longebidad de estas, puesto que al modelo le cuesta ajustarse a ellas y por lo tanto realizar una buena predicción.

Rangos	ARIMA (1,1,1)
1 hora	575.25
2 horas	421.99
1 día	863.98

Tabla 3.7: Errores en la predicción del modelo ARIMA (1,1,1) para el caso B.

K-NN caso B

Las siguientes pruebas con el método K-NN se hicieron con un valor de K=7 y un peso igual a distancia. Las Figuras 3.30, 3.31 y 3.32 muestran los resultados de las pruebas al método K-NN con peso igual a distancia y un valor de K=7 en el caso A, prediciendo 1 hora, 2 horas y un día, respectivamente.

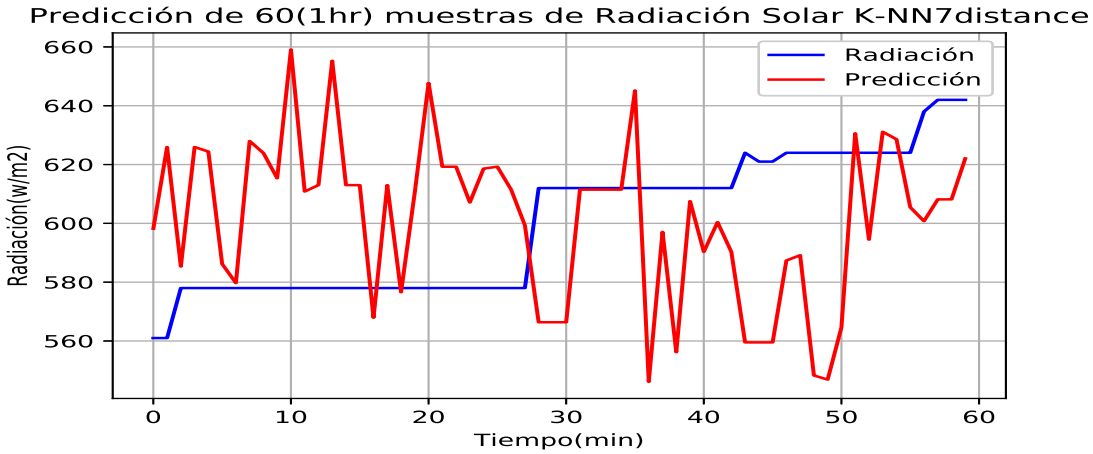


Figura 3.30: Resultados de la predicción de una hora de radiación solar del año 2017 con el método K-NN con peso= distancia y K=7, caso B.

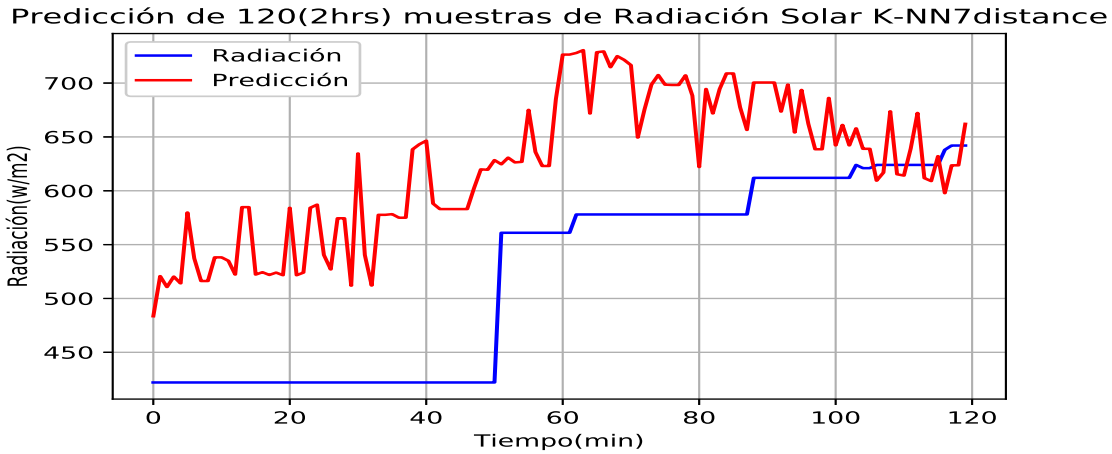


Figura 3.31: Resultados de la predicción de 2 horas de radiación solar del año 2017 con el método K-NN con peso= distancia y K=7, caso B.

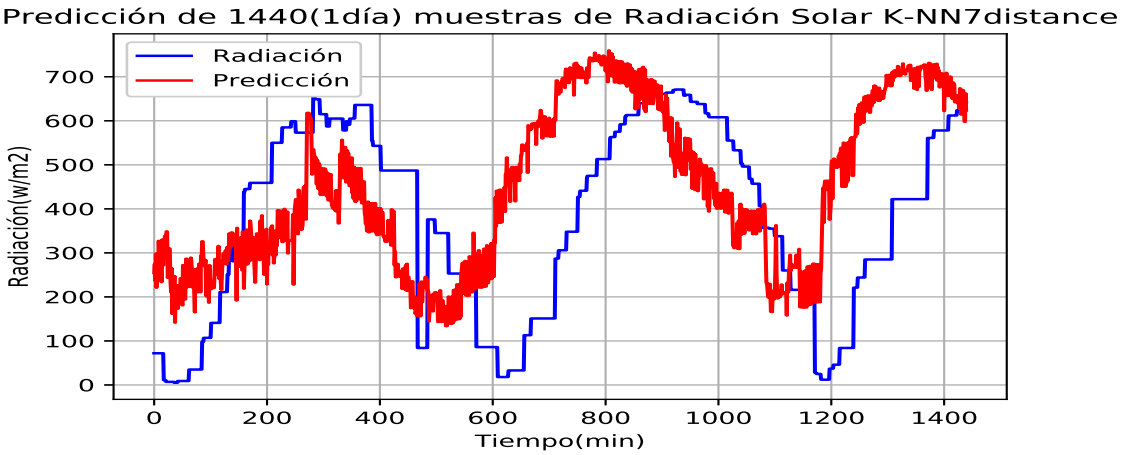


Figura 3.32: Resultados de la predicción de un día de radiación solar del año 2017 con el método K-NN con peso= distancia y K=7, caso B.

Los resultados de estas pruebas obtuvieron errores de predicción de alrededor de 200 unidades (los errores de predicción para el método K-NN con K=7 y peso igual a distancia para el caso B, se encuentran en la tabla 3.8), lo cual es muy elevado en comparación a los obtenidos en el caso A pero podemos observar en las gráficas anteriores que las muestras predecidas continuan siguiendo regularmente bien a las muestras originales la mayor parte del tiempo, por lo tanto, podemos observar que los resultados de predicción no son muy buenos a excepción de la predicción de una hora, ya que presenta un error de 34.51 unidades, más sin embargo el seguimiento de las muestras predecidas sigue siendo regularmente bueno.

Rangos	K-NN (K=7,peso=distancia)
1 hora	34.51
2 horas	104.09
1 día	198.71

Tabla 3.8: Errores en la predicción del método K-NN con k=7 y peso igual a distancia para el caso B.

MSV caso B

Las siguientes pruebas con el método MSV se hicieron con una función kernel del tipo gaussiano. Las Figuras 3.33, 3.34 y 3.35 muestran los resultados de la MSV con el kernel gaussiano en el caso B, prediciendo 1 hora, 2 horas y un día, respectivamente.

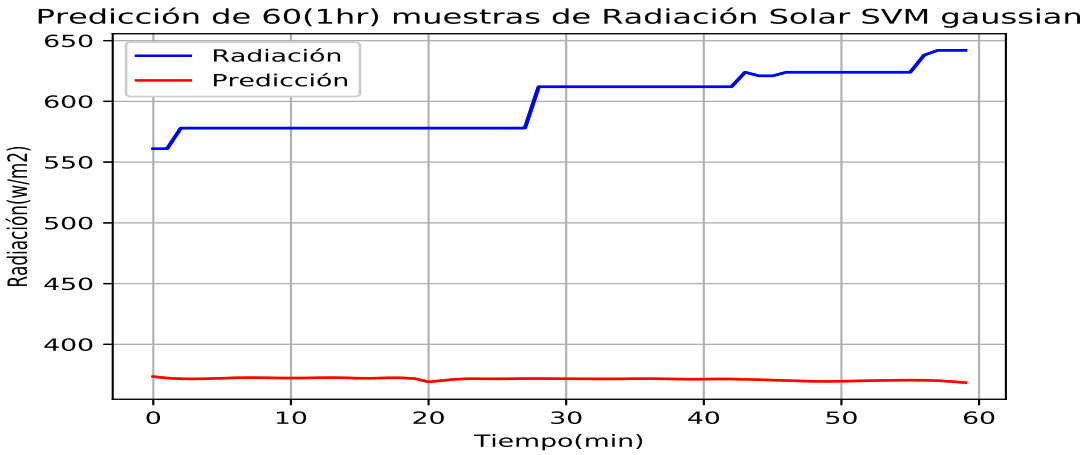


Figura 3.33: Resultados de la predicción de una hora de radiación solar del año 2017 con el método MSV con kernel tipo gaussiano, caso B.

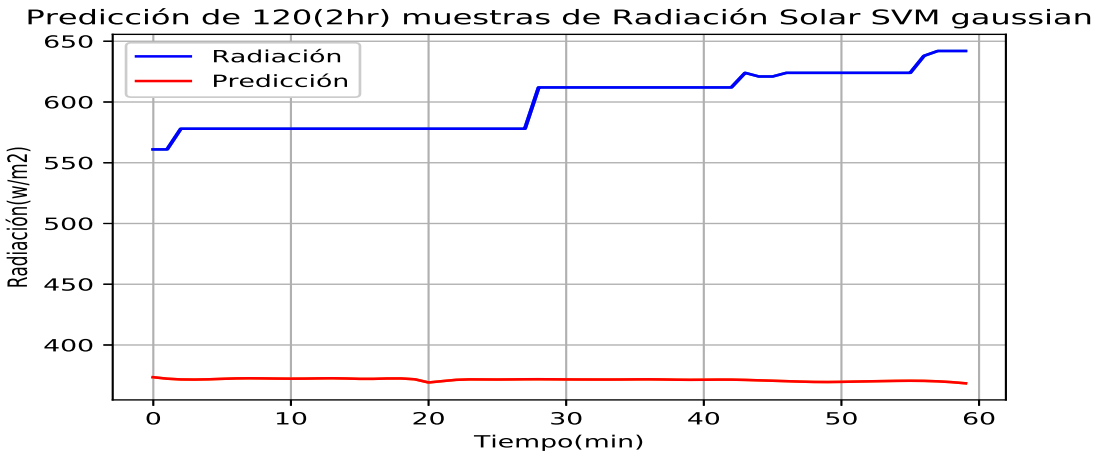


Figura 3.34: Resultados de la predicción de dos horas de radiación solar del año 2017 con el método MSV con kernel tipo gaussiano, caso B.

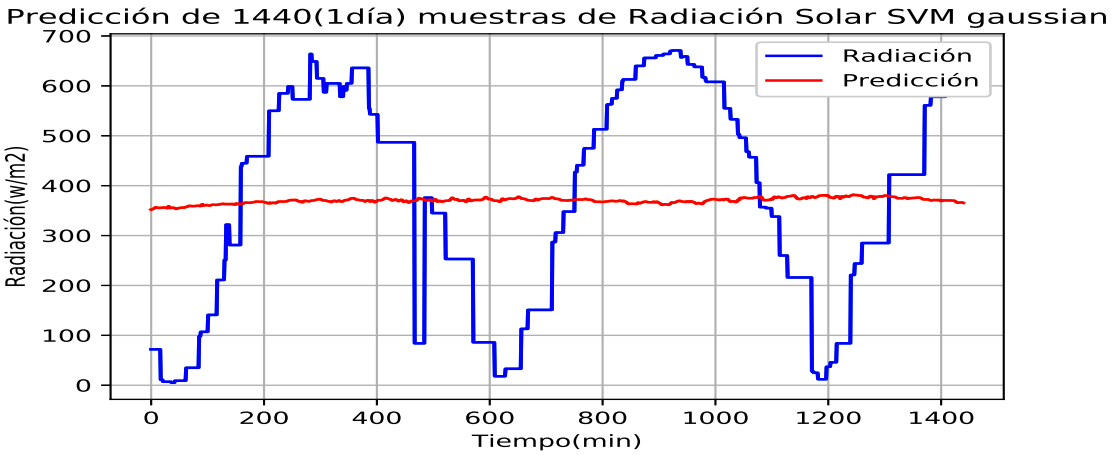


Figura 3.35: Resultados de la predicción de un día de radiación solar del año 2017 con el método MSV con kernel tipo gaussiano, caso B.

Los resultados de estas pruebas obtuvieron errores de predicción de alrededor de 230 unidades(los errores de predicción para un método MSV con kernel gaussiano para el caso B, se encuentran en la tabla 3.9), lo cual es un error muy elevado con respecto a los obtenidos en el caso A, además podemos observar en las gráficas anteriores que las muestras predecidas continuan sin poder siguir a las muestras originales pero esta vez no las siguen en ningun momento, por lo tanto, podemos observar que los resultados de predicción no son nada buenos, por lo tanto, observando los resultados en el caso A y el caso B, esto se puede deber a que la función kernel no se ajusta bien a las series de tiempo empleadas en las pruebas.

Rangos	MSV (kernel gaussiano)
1 hora	228.67
2 horas	151.40
1 día	191.18

Tabla 3.9: Errores en la predicción del método MSV con kernel gaussiano para el caso B.

Discución de los resultados obtenidos de los modelos y métodos en los casos A y B

En esta parte vamos a hacer una comparación de los resultados obtenidos en las pruebas realizadas en los casos A y B, también veremos cual de estos métodos de predicción obtuvo mejores resultados y fue relativamente el mejor.

En la tabla 3.10 podemos ver los resultados en cuanto a errores de predicción en las pruebas.

Métodos	Rangos de predicción	Errores caso A	Errores caso B
ARIMA	1 hora	379.79	575.15
ARIMA	2 horas	391.11	421.99
ARIMA	1 día	412.77	863.98
K-NN	1 hora	33.56	34.51
K-NN	2 horas	23.77	104.09
K-NN	1 día	28.89	198.71
MSV	1 hora	24.12	228.67
MSV	2 horas	17.07	151.40
MSV	1 día	22.90	191.18

Tabla 3.10: Errores en la predicción de los métodos para los casos A y B.

En las pruebas del caso A podemos observar que los mejores resultados en cuanto a errores de predicción los tuvieron los métodos K-NN y la MSV, teniendo esta última un menor error de unidades, sin embargo, si observamos las gráficas obtenidas, el método K-NN siguió mucho mejor a la serie original que la MSV, por otra parte el modelo ARIMA no obtuvo buenos resultados en ninguno de los dos puntos.

En las puebas del caso B podemos observar que los mejores resultados en errores de predicción y seguimiento de las muestras originales los obtuvo el método K-NN.

Conclusiones de los resultados de los casos A y B

Podemos concluir que el médo K-NN fue mejor, seguido del método MSV y por último el modelo ARIMA, sin embargo, no quiere decir que, estos últimos no sean buenos, si no que les afectó mucho la forma de las series empleadas, además de la longebidad de

los datos recibidos, el modelo ARIMA pude ser mejor si se calculan mejor sus parametros o se calculara que orden podría ser mejor, pero en consecuencia se tendría que usar un mejor equipo de computo, (algo de lo que no disponiamos) la MSV puede tener mejores resultados si se calcula la función kernel necesaria para ajustarse mejor a las series de tiempo empleadas en estas pruebas, y el método K-NN, es muy bueno debido a la simplicidad con la que se puede usar, además, como ya se vio en las pruebas, puede funcionar bien incluso si se le da una inmensa cantidad de datos, además este método es muy bueno empleando series de tiempo relacionadas con la generación fotovoltaica en el.

- El modelo ARIMA puede ser mejor si se calcularan mejor los parametros del modelo y se tuviera un mejor equipo de computo para las corridas.
- El método MSV puede ser mejor si se genera una función kernel más apropiada para las series de tiempo que estamos empleando en estas pruebas.
- El método K-NN puede usarse casi sin importar la cantidad de datos o por lo menos puede utilizar una inmensa cantidad de datos tales como los de la serie de tiempo de radiación solar, teniendo resultados regulares (como se vio en los resultados de las pruebas).
- El método K-NN es muy bueno debido a la simplicidad con la que se puede usar.
- El método K-NN funciona muy bien con series de tiempo relacionadas a la generación fotovoltaica.
- El método K-NN fue el que mejor resultados tuvo en las pruebas.

3.5. Teorema de Takens, predicción

Mediante el teorema de Takens se buscará mejorar la predicción usando las series de tiempo usadas anteriormente, tomando en cuenta lo siguiente:

Dada la ecuación (3.1), el teorema de Takens nos dice que:

$$S = y(t) = f(x) = f[y_{(t-\tau)} \cdots y_{(t-(m-1)\ tau)}] \tag{3.1}$$

Se puede inferir que dada una serie temporal S, es posible predecir el estado en el tiempo t (en adelante, y_t) usando m observaciones anteriores muestreadas en frecuencia τ. Entonces, una vez visto esto, se tiene como principal problema que la función $f(x)$ suele ser demasiado compleja para ser analizada (como se vio con los resultados en ARIMA y en la misma forma de las series de tiempo empleadas, así como su longitud de datos), por lo tanto es aquí es donde entran en juego los algoritmos de aprendizaje, por lo tanto, se decidió aplicar el teorema de Takens en conjunto con un método de aprendizaje que obtuviera los mejores resultados en las pruebas de los casos A y B, con el fin de que trabaje en conjunto con el teorema de Takens y ver que tanto se puede mejorar la predicción, además de observar que tan efectivo es dicho teorema para predecir. Se decidió aplicar el teorema de Takens en conjunto con el método K-NN, ya que este último fue el método de aprendizaje que obtuvo los mejores resultados en las pruebas del caso A y el caso B, así como tambíen en las pruebas para encontrar las configuraciones de los métodos que se usarían en las pruebas de los casos A y B.

Para estas pruebas se usaron los datos de la serie de tiempo de un mes de voltaje generado usada en las pruebas del caso A, también la serie de tiempo de un año de temperatura, usada en las pruebas para obtener la mejor configuración de los tres métodos de predicción y la serie de tiempo de un año de radiación solar sin ceros, usada en las pruebas del caso B, nuevamente se hará predicción de 1 hora, 2 horas y un día para cada serie de tiempo empleada en estas últimas pruebas; cuando se hagan las pruebas para la serie de tiempo de un mes de voltaje generado y para la serie de tiempo de un año de radiación solar se usarán las mismas muestras para entrenar el método K-NN en conjunto con el teorema de Takens y para validar la predicción que en las pruebas del caso A y B, y para la serie

de tiempo de un año de temperatura se usaran para 1 hora, 17668 muestras para entrenamiento, lo cual equivale a un 99.9322 % de la serie de tiempo y el 0.0678 % para validar la predicción lo que equivale a 12 muestras para 2 horas, 17656 muestras para entrenamiento, lo cual equivale a un 99.8643 % de la serie de tiempo y el 0.1357 % para validar la predicción lo que equivale a 24 muestras y para 1 día, 17392 muestras para entrenamiento, lo cual equivale a un 98.3711 % de la serie de tiempo y el 1.6289 % para validar la predicción lo que equivale a 288 muestras, se decidió usar la serie de tiempo de temperatura para ver los efectos del teorema de Takens en un rango intermedio de datos para el teorema de Takens se experimento con sus parámetros m y τ de: m=3 y $\tau = 1$, lo cual se representa en la Figura 3.36

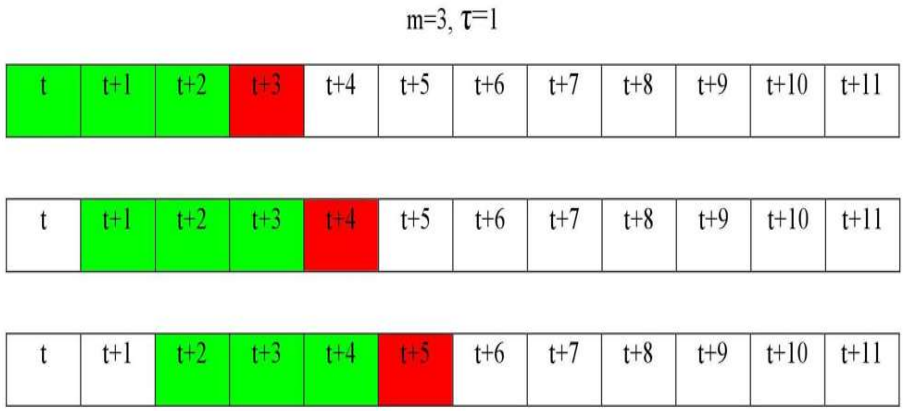


Figura 3.36: Imagen que muestra la configuración usada del teorema de Takens en las pruebas y cómo funciona.

Pruebas con la serie de tiempo de un mes de voltaje generado, aplicando el teorema de Takens en conjunto con K-NN

Las siguientes pruebas se hicieron con el método K-NN en conjunto con el teorema de Takens con configuraciónes K=7 y un peso igual a distancia, una $\tau = 1$ y m=3, respectivamente, se usó otra vez la serie de un mes de voltaje generado, simulando el caso A. Las figuras 3.37, 3.38 y 3.39 muestran los resultados de la predicción de 1 hora, 2 horas y un día, respectivamente, utilizando el teorema de Takens en conjunto con el método K-NN en

la serie de tiempo de un mes de voltaje generado.

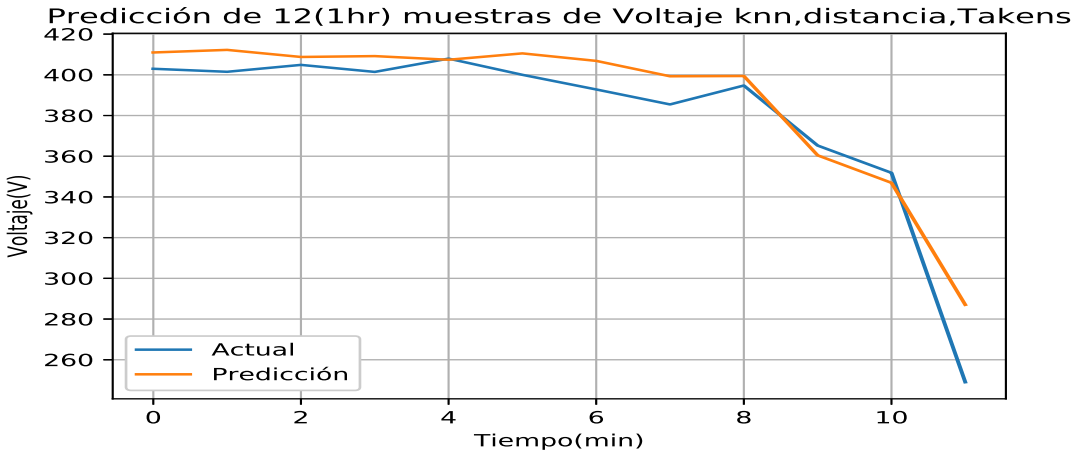


Figura 3.37: Resultados del teorema de Takens en conjunto con el método K-NN con configuraciones, K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 1 hora de la serie de tiempo un mes de voltaje generado.

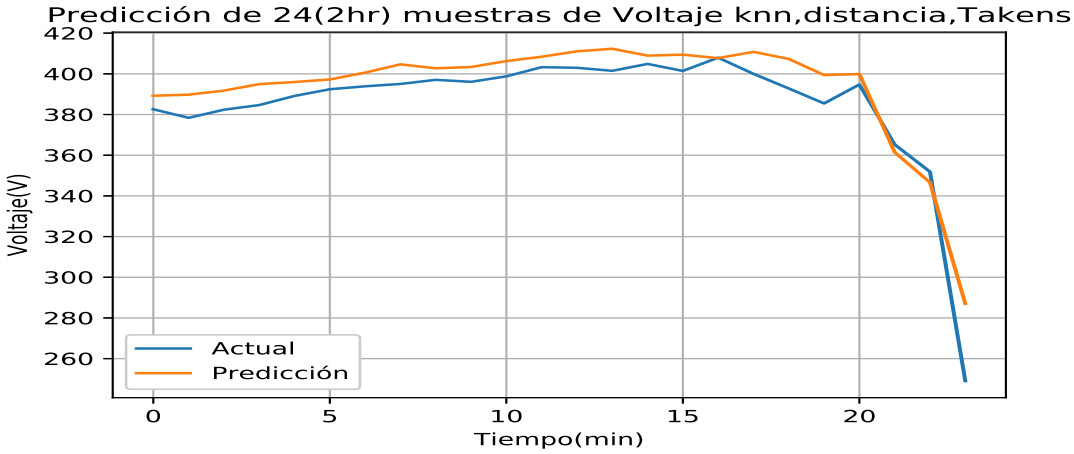


Figura 3.38: Resultados del teorema de Takens en conjunto con el método K-NN con configuraciones, K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 2 horas de la serie de tiempo un mes de voltaje generado.

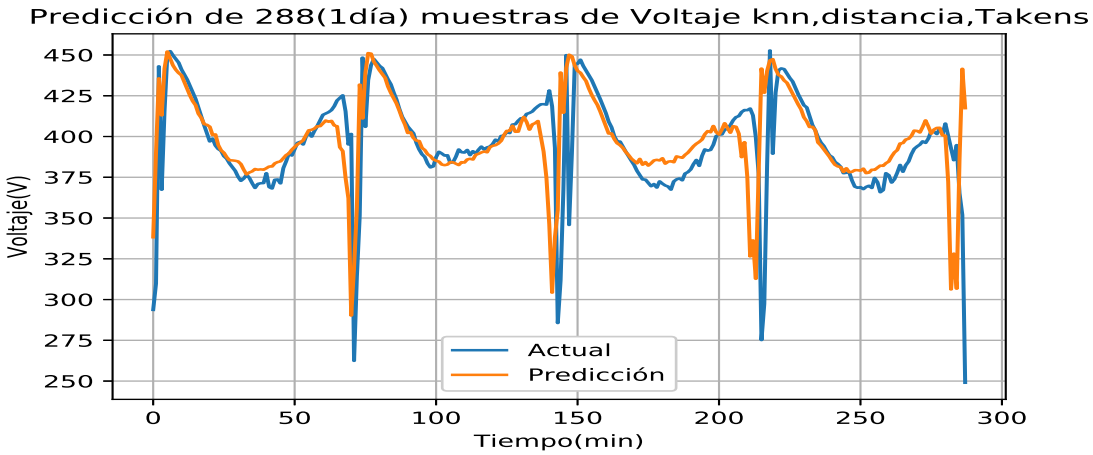


Figura 3.39: Resultados del teorema de Takens en conjunto con el método K-NN con configuraciones, K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 1 día de la serie de tiempo un mes de voltaje generado.

Los resultados de las predicciones para nuestra serie de tiempo de un mes de voltaje generado, aplicando el teorema de Takens en conjunto con el método K-NN, obtuvo errores de hasta 15 unidades logrando reducir el error considerablemente (los errores de predicción para un método K-NN con K=7 y peso distancia en conjunto con el teorema de Takens con $\tau = 1$ y $m = 3$, se encuentran en la tabla 3.11), además podemos observar que las muestras predecidas siguen muy bien a las muestras originales, logrando incluso seguir los pricos altos y bajos (como se aprecia en la Figura 3.39), por lo tanto, podemos decir que los resultados son muy buenos.

Rangos	K-NN con Takens (K=7, peso=distancia, $\tau = 1$, $m = 3$)
1 hora	10.15
2 horas	8.91
1 día	14.72

Tabla 3.11: Errores en la predicción del método K-NN en conjunto con el teorema de Takens, usando la serie de un mes de voltaje generado.

Pruebas con la serie de tiempo de un año de temperatura, aplicando el teorema de Takens en conjunto con el método K-NN

Antes que nada, se realizaron pruebas para predecir 2 horas y un día a la serie de tiempo de un año de temperatura puesto que sólo se había realizado predicción de 1 hora, además se captarón los errores de predicción para esas nuevas pruebas, las cuales tienen el siguiente comportamiento gráfico y los siguientes errores, nota: se anexó nuevamente la gráfica de predicción de una hora. Las Figuras 3.40, 3.41 y 3.42 muestran los resultados de la predicción de 1 hora, 2 horas y un día, respectivamente, utilizando el método K-NN con peso = distancia y K=7, para la serie de tiempo de un año de temperatura.

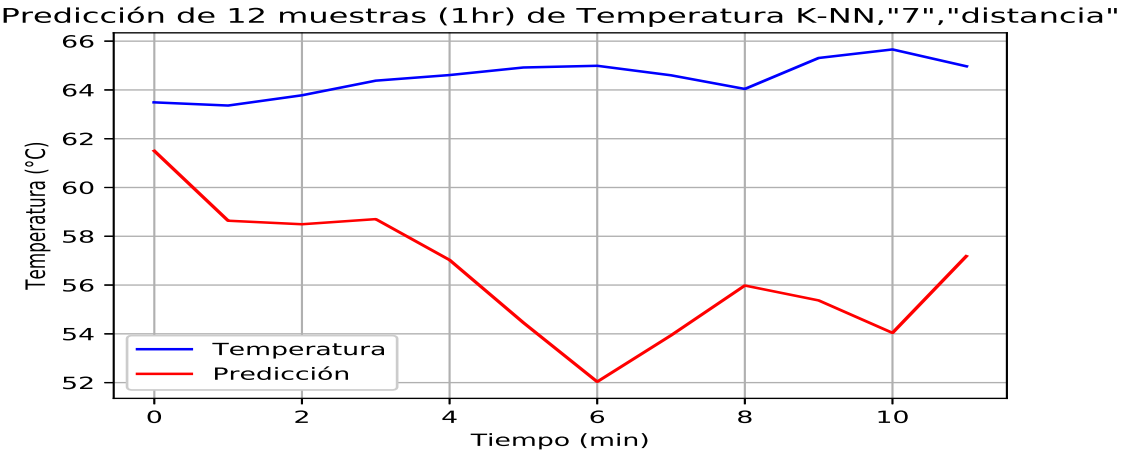


Figura 3.40: Resultados del método K-NN con K=7 y peso=distancia, prediciendo 1 hora de la serie de tiempo de un año de temperatura.

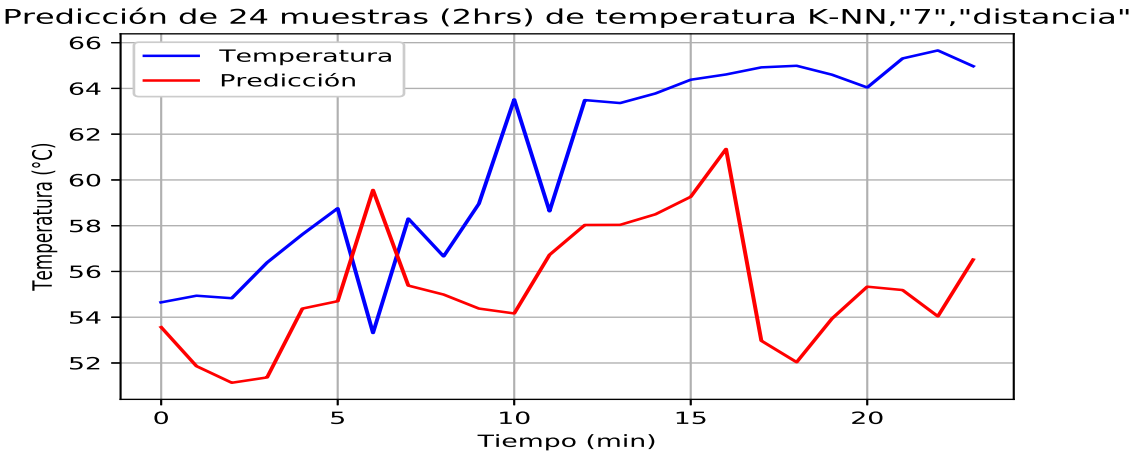


Figura 3.41: Resultados del método K-NN con K=7 y peso=distancia, prediciendo 2 horas de la serie de tiempo de un año de temperatura.

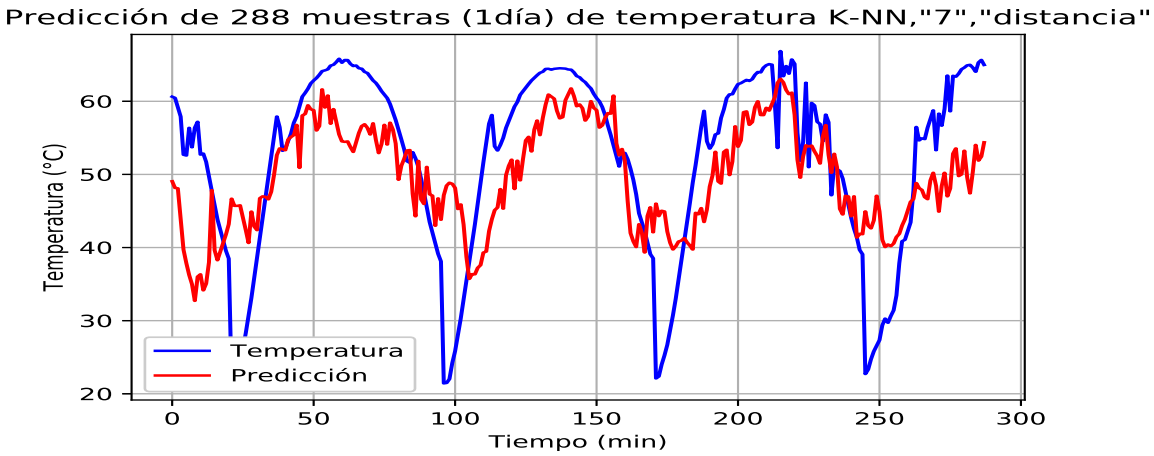


Figura 3.42: Resultados del método K-NN con K=7 y peso=distancia, prediciendo 1 día de la serie de tiempo de un año de temperatura.

Rangos	K-NN (K=7, peso=distancia)
1 hora	8.06
2 horas	6.07
1 día	7.67

Tabla 3.12: Errores en la predicción del método K-NN con K=7 y peso igual a distancia utilizando la serie de tiempo de un año de temperatura.

Pruebas a la serie de tiempo de un año de temperatura con Takens

Las siguientes pruebas se hicieron con el método K-NN y el teorema de Takens en conjunto, con configuraciónes, K=7 y un peso igual a distancia, una $\tau = 1$ y $m = 3$. Las Figuras 3.43, 3.44 y 3.45 muestran los resultados de la predicción de 1 hora, 2 horas y un día respectivamente, utilizando el teorema de Takens en conjunto con K-NN usando la serie de tiempo de un año de temperatura.

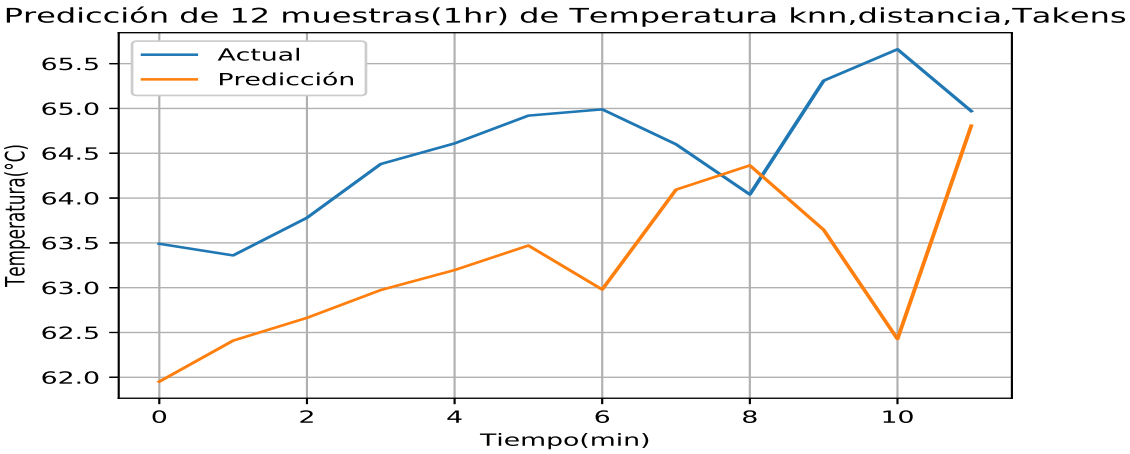


Figura 3.43: Resultados del teorema de Takens en conjunto con K-NN con un valor de K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 1 hora de la serie de tiempo de un año de temperatura.

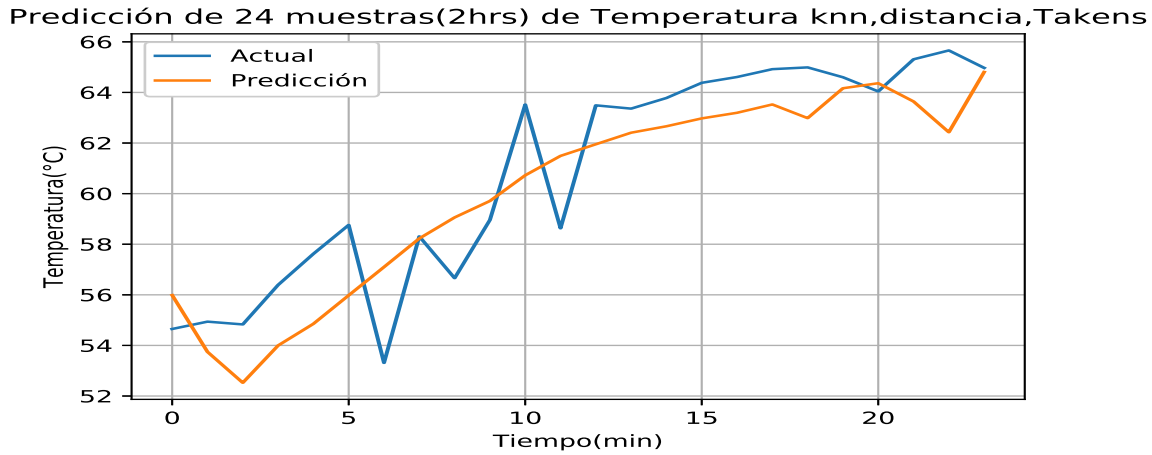


Figura 3.44: Resultados del teorema de Takens en conjunto con K-NN con un valor de K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 2 horas de la serie de tiempo de un año de temperatura.

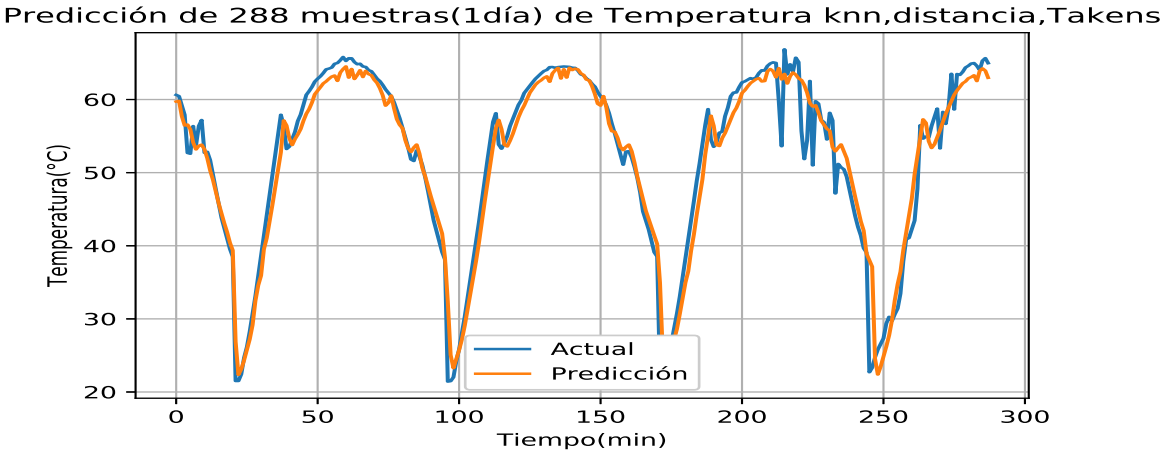


Figura 3.45: Resultados del teorema de Takens en conjunto con K-NN con un valor de K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 1 día de la serie de tiempo de un año de temperatura.

Los resultados de las predicciones para nuestra serie de tiempo de un año de temperatura, aplicando el teorema de Takens en conjunto con el método K-NN, obtuvo errores de predicción de hasta de 2 unidades, logrando reducir el error considerablemente (los errores de predicción para un método K-NN con K=7 y peso igual a distacia para la serie de un año de temperatura, se encuentran en la tabla 3.12 y en conjunto con el teorema de Takens con $\tau = 1$ y $m = 3$ se encuentran en la tabla 3.13), además podemos observar en las gráficas anteriores que las muestras predecidas siguen muy bien a las muestras originales, solo en la Figura 3.43) no logra seguir bien los altos y bajos de las muestras originales pero en la Figura 3.45) si logra seguirlos muy bien la mayor parte del tiempo, por lo tanto, podemos decir que los resultados son muy buenos.

Rangos	K-NN con Takens (K=7, peso=distancia, $\tau = 1$, $m = 3$)
1 hora	1.31
2 horas	1.71
1 día	2

Tabla 3.13: Errores en la predicción del método K-NN en conjunto con el teorema de Takens para la serie de un año de temperatura.

Pruebas con la serie de tiempo de un año de radiación solar, aplicando el teorema de Takens en conjunto con el método K-NN

Las siguientes pruebas se hicieron con el método K-NN en conjunto con el teorema de Takens con configuraciónes K=7 y un peso igual a distancia, $\tau = 1$ y $m = 3$, respectivamente, usando nuevamente la serie de un año de radiación solar simulando el caso B. Las figuras 3.46, 3.47 y 3.48 muestran los resultados de la predicción de 1 hora, 2 horas y un día respectivamente, utilizando el teorema de Takens en conjunto con K-NN, en la serie de tiempo de un año de radiación solar.

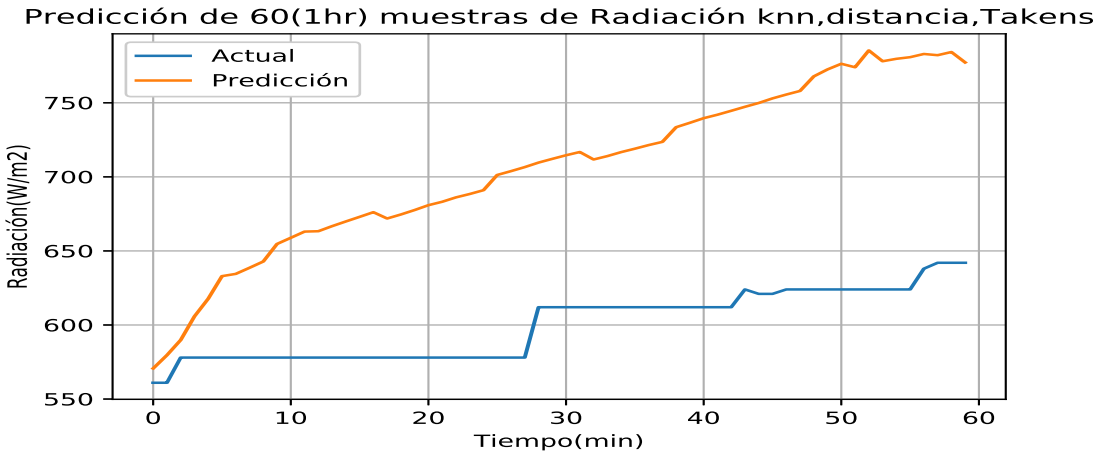


Figura 3.46: Resultados del teorema de Takens en conjunto con K-NN con un valor de K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 1 hora de la serie de tiempo de un año de radiación solar.

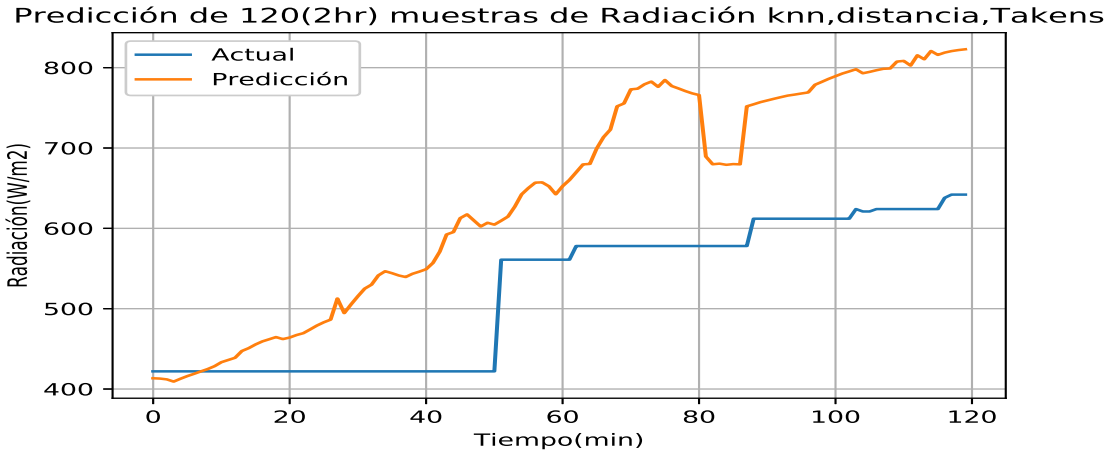


Figura 3.47: Resultados del teorema de Takens en conjunto con K-NN con un valor de K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 2 horas de la serie de tiempo de un año de radiación solar.

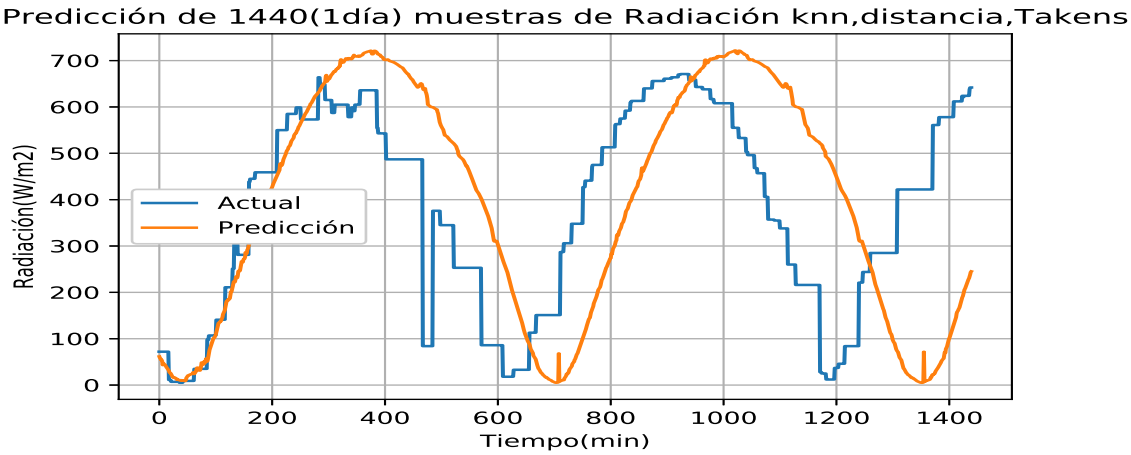


Figura 3.48: Resultados del teorema de Takens en conjunto con K-NN con un valor de K=7 y peso= distancia, $\tau = 1$ y $m = 3$, prediciendo 1 día de la serie de tiempo de un año de radiación solar.

Los resultados de las predicciones para nuestra serie de tiempo de un año de radiación solar, aplicando el teorema de Takens en conjunto con el método K-NN, obtuvo errores de predicción de hasta 180 unidades, lo que son errores muy grandes (los errores de predicción para el método K-NN en conjunto con el teorema de Takens con K=7 y peso =distancia, $\tau = 1$ y $m = 3$ se encuentran en la tabla 3.14), tambíen podemos observar en las gráficas anteriores que las muestras predecidas intentan seguir y ajustarse a las muestras originales, como se muestra en el inicio de la Figura 3.48, la cual intenta seguirlas, ahí podemos ver los efectos del teorema de Takens, sin embargo, la predigción no mejoró realmente, solo una parte del ajuste, tal vez para este caso se tendrían que ajustar los parámetros utilizados, para ver un mejor efecto y obtener mejores resultados, puesto que esta serie de tiempo sigue siendo muy grande y bastante compleja.

Rangos	K-NN con Takens(K=7, peso=distancia, $\tau = 3$, $m = 1$)
1 hora	106.07
2 horas	119.93
1 día	186.23

Tabla 3.14: Errores en la predicción del método K-NN en conjunto con el teorema de Takens usando la serie de tiempo de un año de radiación solar.

3.5.1. Discución de resultados con el teorema de Takens y sin el teorema de Takens

En esta parte haremos una comparación de los resultados usando el teorema de Takens y sin usar el teorema de Taken, con el método K-NN y veremos que tanto se corrige la predicción al usar el teorema de Takens y hasta que longitud de datos puede hacerlo bien.

En la tabla 3.15 podemos ver los resultados en cuanto a errores de predicción, en las pruebas con el teorema de Takens y sin el teorema de Takens.

Rangos	Vol.	Vol. Takens	Temp.	Temp. Takens	Rad.	Rad. Takens
1 hora	33.56	10.15	8.06	1.31	34.51	106.07
2 horas	23.77	8.91	6.07	1.71	104.09	119.93
1 día	28.89	14.72	7.67	2	198.71	186.23

Tabla 3.15: Errores en la predicción de generación de voltaje, caso A, sin el teorema de Takens y con el teorema de Takens, temperatura sin el teorema de Takens y con el teorema de Takens, la radiación solar, caso B, sin el teorema de Takens y con el teorema de Takens.

En los resultados de estas últimas pruebas, usando el teorema de Takens en conjunto con el método K-NN, podemos ver que los errores de predicción, así como el seguimiento a las muestras originales mejora bastante, hasta en un 300 %, si no es que más. El teorema de Takens dio resultados muy positivos para las series de temperatura y voltaje generado, en especial la serie de temperatura, lo cual indica que los resultados dependen en gran parte del comportamiento de los datos, dado que para el caso B la serie de tiempo de radiación solar no hubo realmente una mejora en los resultados, si no que el error aumento, por lo tanto,

se tendría que modificar los parametros del teorema de Takens para casos con cantidades muy grandes de datos como lo es la serie de radiación solar empleada en estas pruebas y en las series de tiempo de generación fotovoltaica, en general el teorema de Takens dio muy buenos resutados. Es importante conocer este tipo de teoremas, ya que con ellos podemos reducir los errores en las predicciones de nuestros métodos, que para una planta grande de generación fotovoltaica resultaría algo escencial, ya que si se tiene conocimiento de que mi método de predicción genera errores que representan 90 mega watts, se concluiría que estos resultados no servirían para tomar en cuenta alguna acción correctiva, sin embargo, si ese error es muy bajo y aparte las muestras predecidas siguen muy bien a las muestras originales, entonces se tendrá la certeza y la confianza de poder realizar alguna acción y así generar cambios positivos y muy importantes en la generación y forma de trabajar, y de abastacer la demanda energética.

Conclusiones de los resultados usando el Teorema de Takens

Podemos concluir que el médo K-NN en conjunto con el teorema de Takens, dio buenos resultados para las series de tiempo de voltaje y temperatura, logrando reducir ampliamente el error de predicción y ayudando también a mejorar ampliamente el seguimiento de las muestras predecidas a las muestras originales, en la serie de tiempo de radiación solar no pudo hacer mucho el teorema de Takens en conjunto con el método K-NN, esto debido a la inmensa cantidad de datos, sin embargo en la Figura 3.48 si se ajustó bien el inicio, aunque esto no significa que el teorema de Takens en conjunto con el método K-NN no pueda funcionar con esta serie de tiempo, si no que posiblemente no pudo hacer mucho debido a que para una serie de tiempo tan compleja si se necesitammría ajustar los parametros del teorema de Takens y tal vez los del método K-NN también, pero en general, si funcionaron los métodos K-NN y el teorema de Takens, si se cumplió lo que dice el teorema de Takens y ayudó a mejorar las predicciones.

- El teorema de Takens redujó bastante el error que hay entre las muestras predecidas y las muestras originales, hasta en un 300 porciento o más en la serie de generación de voltaje y la serie de temperatura, en especial esta última.

- El teorema de Takens mejoro en gran medida el seguimiento de las muestras predecidas hacia las muestras originales en la serie de generacion de voltaje y la serie de temperatura, en especial esta ultima.

- El teorema de Takens no pudo mejorar la predicción del caso B(radiación solar), habrían que ajustarse los parámetros del teorema y tal vez del método K-NN.

- El Teorema de Takens en conjunto con el método K-NN, resultó muy bueno para reducir el error de predicción en gran medida.

Capítulo 4

Conclusiones

4.1. Introducción

En esta sección nos encontraremos todo lo aprendido durante el desarrollo de este tema de tesis, dando las conclusiones particulares y generales sobre el sistema fotovoltaico estudiado, los métodos empleados a lo largo de este trabajo de tesis, sus resultados, los resultados del teorema de Takens el cual se uso para mejorar la predicción y además se mencionan los trabajos futuros.

4.2. Conclusiones particulares

1. La importancia de conocer distintos métodos de predicción, así como también de saber cómo funcionan resulta ser esencial en el tema de generación de energías limpias, esto debido al comportamiento aleatorio de sus fuentes de generación.

2. Los resultados de los métodos de predicción ARIMA, K-NN y la MSV demostraron depender de la forma y las propiedades de las series de tiempo empleadas, en especial el método MSV y el modelo ARIMA.

3. El método K-NN funcionó muy bien utilizando las series de tiempo empleadas, a excepción de la serie detiempo de radiación solar.

4. Es importante conocer métodos que nos permitan mejorar nuestra predicción, métodos

como el teorema de Takens, esto debido a que si las diferencias son grandes, nos podrían generar desconfianza a la hora de implementar alguna acción correctiva, lo cual no sería bueno en granjas solares donde nuestra generación se encuentra en el orden de los gigawatts.

5. El teorema de Takens se desempeñó muy bien en conjunto con el método K-NN, debido a que logró reducir el error de predicción bastante y además mejoró el seguimiento de la predicción.

4.3. Conclusión general

Los resultados de los métodos de predicción usados nos dan a concluir lo siguiente, el modelo ARIMA presenta problemas para predecir cuando la serie de tiempo es muy compleja, así como las series de tiempo empleadas en las pruebas, por supuesto esto no significa que el modelo ARIMA sea malo, más bien nos dice que se tiene que usar una serie menos compleja o bien usar un mejor equipo de computo y así calcular a que orden sería conveniente usar el modelo ARIMA para tales series de tiempo, además de emplear técnicas de optimización para facilitar el trabajo del modelo ARIMA; el método MSV dio resultados regularmente buenos en las primeras pruebas (las que se realizaron con la serie de tiempo de generación fotovoltaica y la serie de tiempo de temperatura), solo con la serie de tiempo de radiación solar presentó problemas, esto debido a que su función kernel no era la más adecuada, lo cual nos dice que para aprovechar más el método MSV se necesita generar la función kernel adecuada para las series de tiempo empleadas en las pruebas, entonces tomando en cuenta esto, el método MSV es muy bueno; el método K-NN fue el que mejor resultados obtuvo en las pruebas, demostrando que puede ser muy bueno al predecir usando series de tiempo complejas, como lo son las series de tiempo de voltaje generado y temperatura, pero para la serie de tiempo de radiación solar habría que jugar más con los parametros del método o usar algun otro método para mejorar la predicción, un método como el teorema de Takens, el cual funciono muy bien en conjunto con el método K-NN, logrando reducir los errores de predicción hasta en un 400 %, logrando también mejorar ampliamente el seguimiento de las muestras originales, en especial las series de

voltaje generado y temperatura, cabe destacar que en la serie de radiación solar, en la predicción de un día, logró ajustarse un poco a ella en el inicio, como se puede ver en la Figura 3.48m más sin embargo si tomamos en cuenta que se usaron parámetros chicos, el resultado fue muy bueno, por lo tanto el teorema de Takens en conjunto con el método K-NN dio muy buenos resultados.

4.4. Trabajos Futuros

1. Además de ya tener como calcular los coeficientes AR, también calcular los coeficientes para el modelo MA.

2. Crear funciones Kernel que se adecuen mejor a las series de tiempo (temperatura, voltaje generado y radiación solar) para aprovechar al máximo este método.

3. Implementar este trabajo de tesis en un despacho económico de generación fotovoltaica.

4. Enfocar este trabajo de tesis para generación eólica.

Apéndice A

Principios de funcionamiento de los paneles solares y conceptos básicos del comportamiento estadístico en series de tiempo

A.1. Principios de funcionamiento de los paneles solares

Introducción

En este primer apartado, hablaremos acerca de los principios de los paneles solares y aprenderemos un poco de como afectan las variables de radiación solar y temperatura en la generación fotovoltaica; abordaremos lo que es un sistema interconectado a la red y las partes que lo componen.

A.1.1. Paneles solares

Los paneles solares están formados de muchas celdas solares, las cuales son pequeñas células hechas de silicio cristalino y/o arseniuro de galio, ambos, materiales semiconductores. Esto quiere decir que son materiales que pueden comportarse como conductores

de electricidad o como aislante dependiendo del estado en que se encuentren. Generalmente, los paneles solares comerciales están hechos con silicio [10].

Estos dos materiales se mezclan con otros, como por ejemplo el fósforo o el boro, el objetivo es entregar una carga positiva y una carga negativa, con lo cual se logra que las celdas tengan las dos cargas y puedan generar eléctricidad, las celdas solares al exponerse a la luz del sol directamente producen corriente. La energía del sol mueve los electrones de la parte de la celda que le sobran hacia la parte de la celda que le faltan. Este movimiento de electrones es justamente la corriente eléctrica, por lo tanto, de esta forma se consigue generar corriente eléctrica de un punto a otro. Todas las celdas solares trabajando en conjunto hacen que se produzca un campo eléctrico en el panel solar y es así como los paneles solares pueden generar energía que posteriormente podemos utilizar como electricidad [10].

Partes que conforman un panel fotovoltaico

Un módulo fotovoltaico está compuesto de:

- **Cubierta frontal:**

  Suele ser de vidrio templado entre 3 y 4 mm de espesor, con muy buena transmisión de la radiación solar, proporciona protección contra los agentes atmosféricos y los impactos (granizo, actos vandálicos, etc.). La superficie exterior del vidrio es anti reflexiva y esta tratada para impedir la retención del polvo y la suciedad. La superficie interior generalmente es rugosa, lo que permite una buena adherencia con el encapsulante de las celdas, además de facilitar la penetración de la radiación solar [10].

- **Encapsulante:**

  En la mayoría de los módulos se emplea Etil-Vinil-Acetato (EVA). En contacto directo con las celdas, protege las conexiones entre las mismas. Además, proporciona el acoplamiento con la cubierta frontal y la protección posterior. Al igual que la cubierta frontal permite la transmisión de la radiación solar y no se degrada con la radiación ultravioleta [10].

- **Cubierta posterior:**

  Se utiliza una capa de polivinilo fluoruro (PVF, comercialmente denominado TEDLAR) o de poliéster. Junto con la cubierta frontal, protege al módulo de la humedad y otros agentes atmosféricos y lo aísla eléctricamente. De naturaleza opaca, es habitual que sea de color blanco para reflejar la luz solar que no recogen las celdas sobre la cara posterior rugosa de la cubierta frontal, que la refleja de nuevo hacia las celdas. Algunos fabricantes ponen está cubierta de vidrio para aprovechar la radiación sola reflejada que puede recogerse por la parte posterior del módulo. Para ello las celdas solares incluyen capas de silicio amorfo que recoge esta radiación [10].

- **Marco:**

  La mayoría de los fabricantes utilizan aluminio anodizado. Proporciona rigidez y resistencia mecánica al módulo, además de un sistema de fijación. Puede incorporar una conexión para la toma de tierra. Nunca se debe mecanizar, porque las vibraciones pueden romper el cristal de la cubierta frontal como se aprecia en la figura A.1 [10].

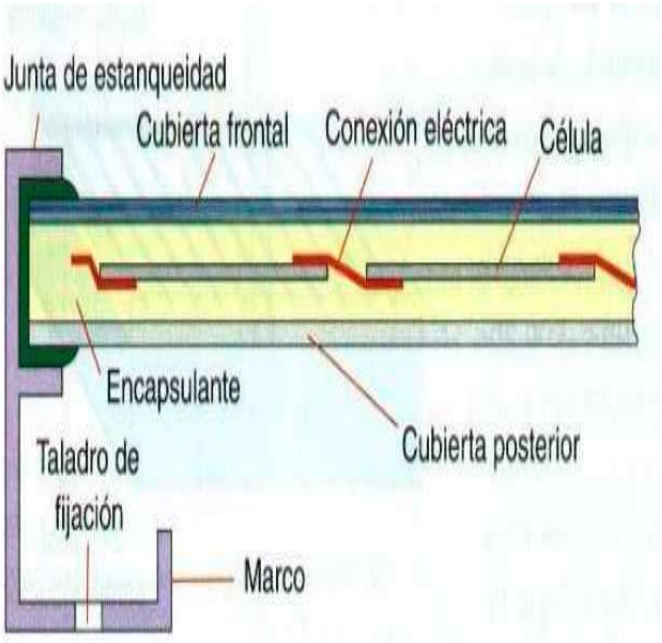
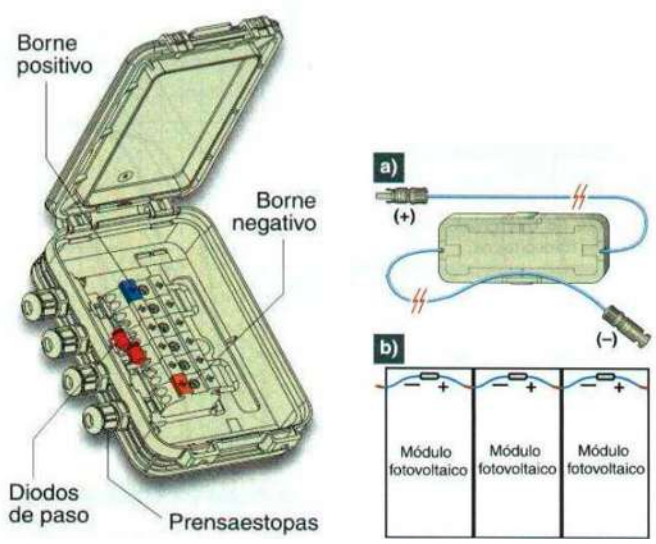


Figura A.1: Partes que conforman un módulo fotovoltaico [10].

Conexiones:

Situadas en la parte superior del módulo, habitualmente consiste en una caja con una protección recomendada contra el polvo y el agua IP-65, fabricada con materiales plásticos resistentes a las temperaturas elevadas, que su interior incorpora los bornes de la conexión positivo y negativo del módulo y los diodos de paso (diodos by-pass). El uso de prensaestopas para el paso de cables, mantiene la protección contra el polvo y el agua. Para su conexión, el fabricante suministra el módulo fotovoltaico con dos cables, finalizados con conectores, diferentes para el positivo y negativo, con la longitud suficiente para permitir una rápida conexión serie entre paneles consecutivos como se aprecia en la figura A.2 [10].



(a) Caja de conexiones. (b) Diagrama de conexiones.

Figura A.2: Caja y diagrama de conexión de un módulo fotovoltoaico [10].

- **Celdas:**

La conexión de las celdas de un módulo fotovoltaico se realiza con cintas metálicas soldadas o incrustadas sobre la rejilla de conexión eléctrica de la cara frontal de cada celda. La interconexión entre celdas se realiza uniendo las contas de la cara frontal (negativo) de una celda con la cara posterior (positivo) de la celda siguiente como se

aprecia en la figura A.3 [10].

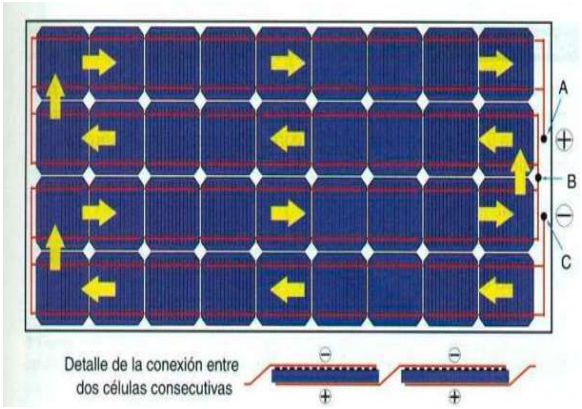


Figura A.3: conexión de un módulo fotovoltaico [10].

A.1.2. Influencia de la irradiación y temperatura sobre una placa fotovoltaica

Las condiciones de funcionamiento de una celda fotovoltaica, tales como la irradiación y la temperatura afectan directamente a la tensión, intensidad y potencia generada por la misma, por lo que es conveniente saber cómo afectan estas condiciones en el comportamiento de una celda solar [11].

Para esto, es necesario introducir conceptos fundamentales tales como:

- Radiación solar.
- Temperatura.
- Tensión de circuito abierto VOC.
- Corriente de cortocircuito ISC.

Radiación solar

La radiación solar es el conjunto de radiaciones electromagnéticas emitidas por el Sol. La radiación solar se distribuye desde el infrarrojo hasta el ultravioleta. No toda la radiación alcanza la superficie de la Tierra, porque las ondas ultravioletas más cortas son

absorbidas por los gases de la atmósfera. La magnitud que mide la radiación solar que llega a la Tierra es la irradiancia, que mide la potencia que por unidad de superficie alcanza a la Tierra. Su unidad es el W/m^2 [30], en el continente americano, México es el país con mayor radiación [31].

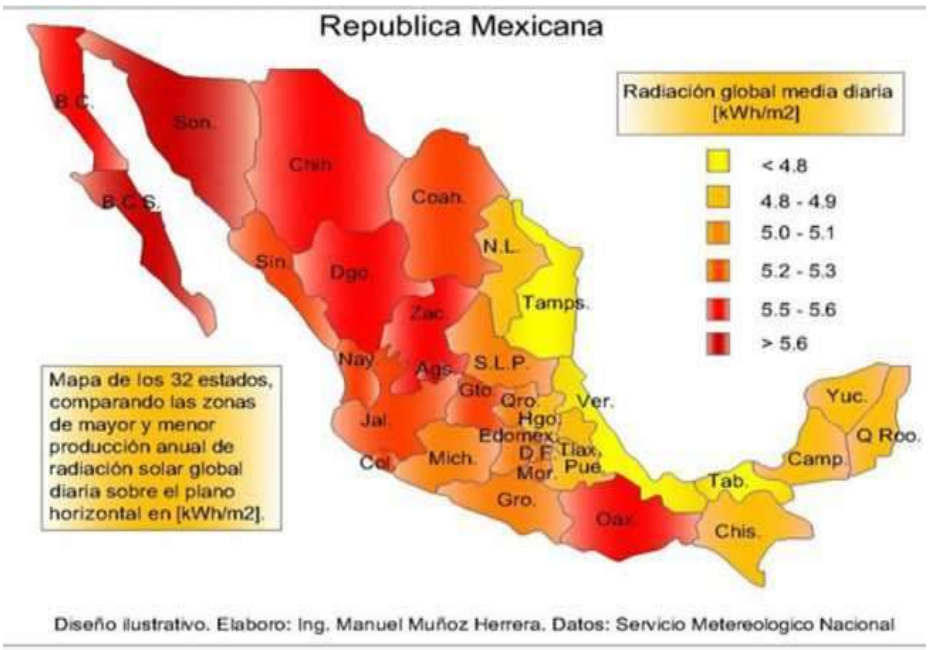


Figura A.4: Mapa de irradiación solar en México

Temperatura

La Temperatura es una magnitud que mide el nivel térmico o el calor que un cuerpo posee. La temperatura atmosférica es el grado de calor que posee el aire en un momento y lugar determinado. Su origen se encuentra fundamentalmente en la influencia de los rayos solares sobre la atmósfera[32].

Actualmente se utilizan tres escalas de temperatura; grados Fahrenheit ($°F$), Celsius $°C$) y Kelvin ($°K$). En la escala Fahrenheit, que es la más utilizada en Estados Unidos, se definen los puntos de congelación y de ebullición normales del agua en 32 y 212 °F, respectivamente. La escala Celsius divide en 100 grados el intervalo comprendido entre el punto de congelación (0 °C) y el punto de ebullición del agua (100 °C) [32].

La tensión de circuito abierto

La tensión de circuito abierto (V_{OC}) es la diferencia de potencial que se alcanza cuando una celda fotovoltaica es iluminada, sin estar conectadas las regiones P y N, siendo proporcional a la iluminación recibida. Es el máximo valor de tensión de la celda [11].

La corriente de cortocircuito

La corriente de cortocircuito (I_{SC}) es aquella que se genera cuando las regiones P y N están unidas por un conductor exterior con una resistencia nula y es proporcional a la iluminación recibida. Es el máximo valor de intensidad de la celda [11]. La figura A.5 muestra el V_{OC} e I_{SC} para una celda solar.

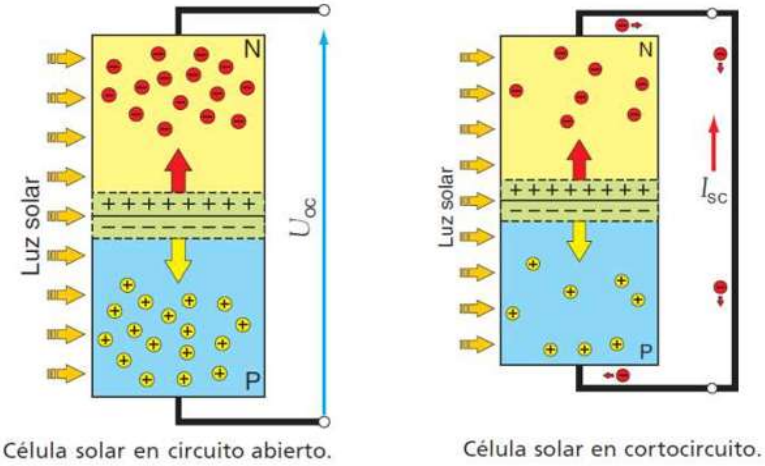


Figura A.5: V_{OC} e I_{SC} en una celda solar [11].

Una situación intermedia entre las dos imágenes de la Figura A.5, es la de un circuito donde las regiones P y N estuvieran unidas mediante un conductor y se encontra con una resistencia receptora. En tal caso, la tensión proporcionada por la celda se obtiene mediante la siguiente ecuación, la ley de Ohm [11]:

A mayor resistencia, el circuito se comportaría como un circuito abierto ($R = \infty, I = 0$), y con una resistencia muy pequeña, se comporta como si estuviera en cortocircuito ($R = 0, I = \infty$). La potencia suministrada por la celda se expresa mediante la ecuación (A.1) [11]

$$P_L = (V_L)(I_L) \tag{A.1}$$

Donde:

P_L: Potencia suministrada por la celda.

V_L: Voltaje suministrado por la celda.

I_L: Corriente suministrada por la celda.

Se cumple siempre que la intensidad I_L y la tensión V_L en el receptor, son inferiores a la intensidad de cortocircuito y a la tensión de circuito abierto, respectivamente. Cuando se habla de la potencia máxima capaz de suministrar una celda, se utiliza la ecuación (A.2) [11]

$$Pmax = (Vmp)(Imp) \tag{A.2}$$

Donde:

P_{max}: Es la potencia máxima.

V_{mp}: Es el voltaje de máxima potencia.

I_{mp}: Es la corriente de máxima potencia.

Si representamos la intensidad y la potencia frente a la tensión generada por una celda a temperatura e irradiación constante, se obtienen las curvas características I-V o P-V, donde se observa cuál es la potencia máxima con la cual se puede extrapolar ese punto para obtener la intensidad en el punto de máxima potencia y la tensión en el punto de máxima potencia. Por lo cual es conveniente trabajar a la celda fotovoltaica cerca de este punto [11]. La figura A.6 muestra un ejemplo de las curvas I-V o P-V.

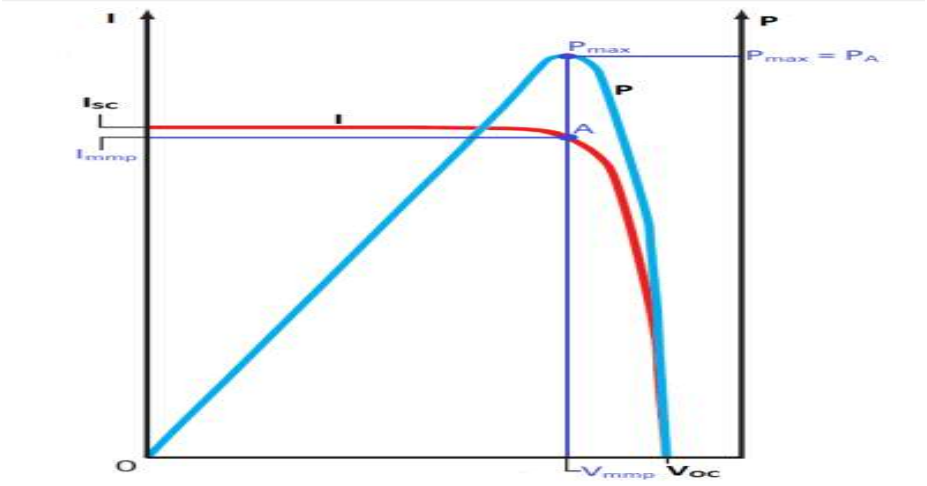


Figura A.6: Potencia máxima en condiciones estándar de medida [11].

La potencia máxima en condiciones estándar de medida (CEM) o Standard Test Conditions (STC), que son: temperatura de la celda ($25°C$), irradiancia ($1000W/m^2$) y AM (masa de aire) de 1.5, también se denomina potencia de pico de la celda. Sin embargo, los sistemas fotovoltaicos raramente operan en condiciones estándar. Las condiciones de funcionamiento son muy variables, pudiendo variar en un rango de $0a1000W/m^2$ en el caso de la irradiancia, y temperatura de la celda hasta $50°C$ superior a la temperatura ambiental [11].

Efectos de la irradiancia

La tensión y corriente generada en una celda depende directamente de la iluminación recibida. La corriente de cortocircuito de la celda es directamente proporcional a la irradiancia como se muestra en la figura A.7, disminuyendo a medida que se reduce la irradiancia. La tensión de circuito abierto varía poco con la irradiancia, aunque también decrece, a efectos prácticos se puede considerar constante [11].

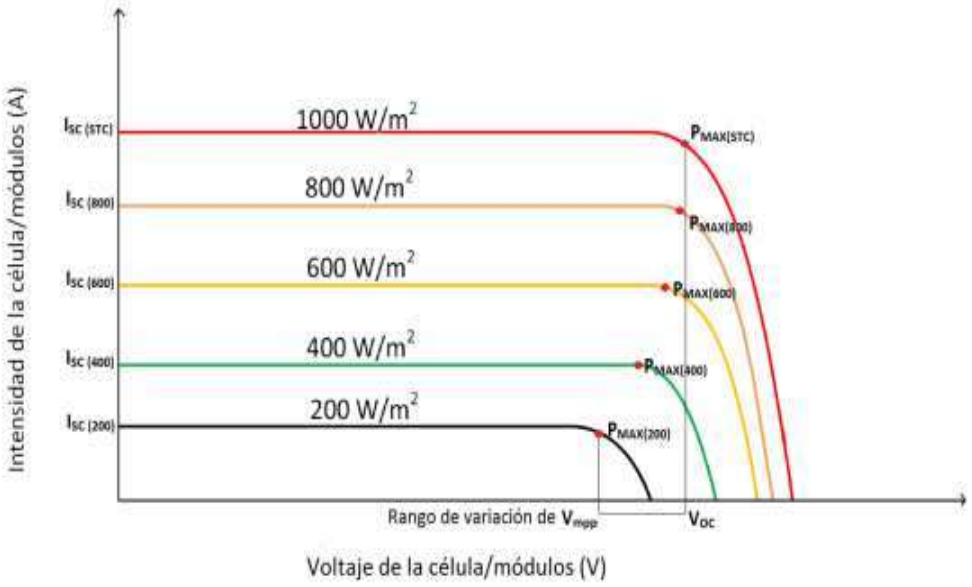


Figura A.7: Efectos de la irradiancia en una celda fotovoltaica [11].

La intensidad de cortocircuito, varía con la irradiancia, siendo esta variación lineal acorde a la ecuación (A.3) [11]

$$Pmax = G\frac{I_{sc}(CEM)}{1000} \tag{A.3}$$

Donde:

$I_{SC}(G)$: intensidad de cortocircuito para una irradiación G.

$I_{SC}(CEM)$: intensidad de cortocircuito en condiciones CEM.

G: irrandiancia (W/m^2).

CEM: Condiciones estándar de medida.

Efecto de la temperatura

Por otro lado, la temperatura afecta de manera considerable a la tensión, tal y como se muestra en la Figura A.8

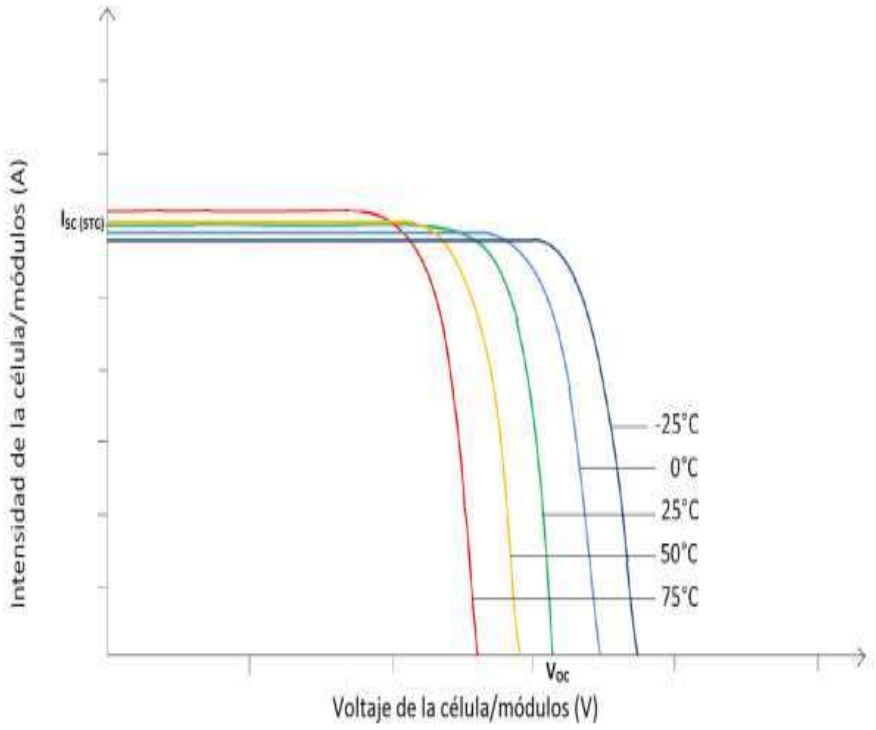


Figura A.8: Efectos de la temperatura en una celda fotovoltaica [11].

Como se aprecia, la tensión de circuito abierto disminuye cuando aumenta la temperatura. La intensidad de cortocircuito, sin embargo, aumenta cuando aumenta la temperatura, aunque la variación es muy pequeña y a efectos prácticos se considera constante. Es evidente que si la tensión de la celda disminuye cuando aumenta la temperatura, la intensidad prácticamente se mantiene constante, la potencia entregada por la celda disminuirá conforme aumenta la temperatura de la celda, tal y como muestra en la Figura A.9 [11]:

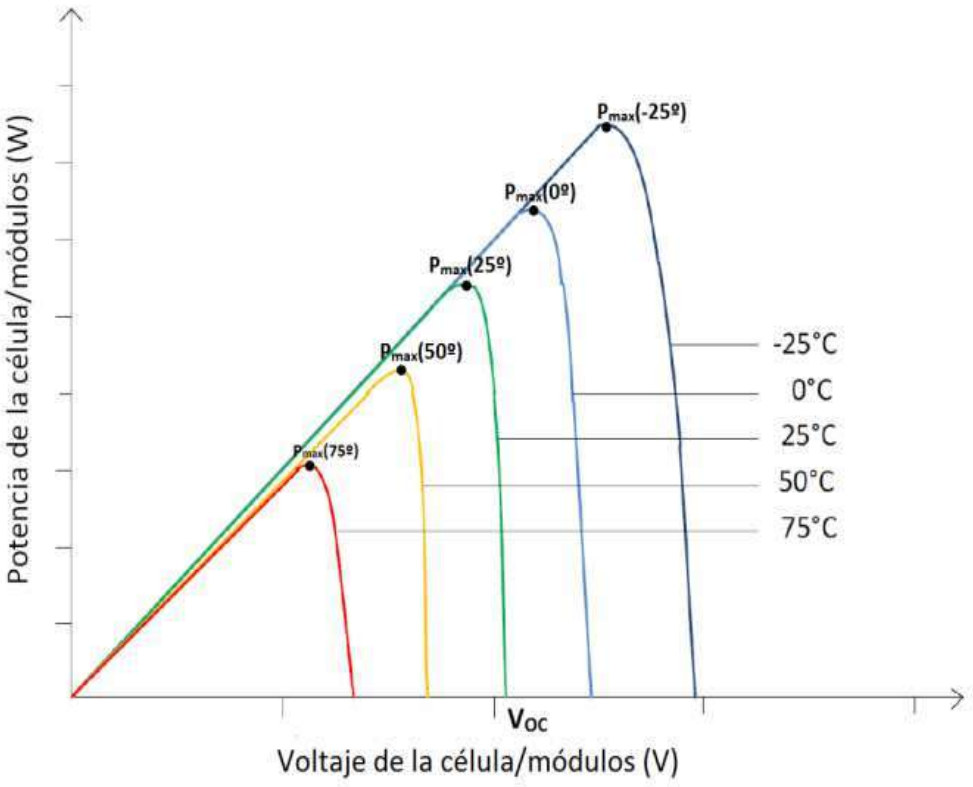


Figura A.9: Efecto de la temperatura hacia la potencia entregada en una celda fotovoltaica [11].

La temperatura de trabajo de una celda está íntimamente relacionada con la temperatura ambiente y la irradiación. Esta se puede obtener mediante la ecuación (A.4) [11]

$$T_c = T_\alpha + G\frac{TONC - 20}{800} \tag{A.4}$$

Donde:

T_c: Temperatura de trabajo de la celda ($°C$).

T_a: Temperatura ambiente ($°C$).

$TONC$: Temperatura de operación nominal de la celda ($°C$).

G: Irrandiancia (W/m^2).

El valor de la temperatura de operación nominal de la celda ($TONC$), es un parámetro que se obtiene de las hojas características de los módulos fotovoltaicos, toma valores que van de 43 a 49°C y si no se dispone de él, se puede tomar 45°C como un valor razonable. TONC o NOCT del inglés Nominal Operating Cell Temperature, corresponde a una irradiaciancia en el plano del módulo de $800W/m^2$, con orientación normal a la radiación incidente al mediodía solar, temperatura ambiente de 20°C, velocidad del viento de $1m/s$g. y funcionamiento en circuito abierto [11].

Sistemas interconectados a la red

Un sistema fotovoltaico interconectado a la red está constituido básicamente por un generador fotovoltaico y un inversor que convierte la corriente continua del generador en corriente alterna con la tensión y la frecuencia requeridas por las compañías eléctricas. Además, debe incluir las protecciones eléctricas correspondientes, como se aprecia en la figura A.10. Toda la energía eléctrica producida se envía a la red de distribución eléctrica, donde es comprada por las compañías distribuidoras de electricidad. La energía necesaria para el consumo, se tiene que extraer de la red comprándola a la compañía distribuidora de electricidad. Es necesario disponer de un sistema de medida de energía eléctrica que contabilice la energía que sale o la energía que entra. El usuario comprará la energía eléctrica que consume a la compañía distribuidora al precio establecido y además es propietario de un sistema generador de electricidad que puede facturar los kWh producidos a un precio superior, la figura A.10 muestra el esquema de conexión de un sistema interconectado a la red [10].

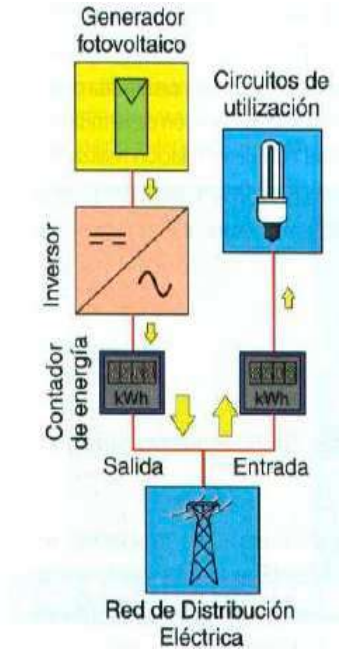


Figura A.10: Diagrama que muestra un esquema de un sistema interconectado a la red de paneles fotovoltaicos [10].

Es imprescindible para instalar un sistema fotovoltaico conectado a la red, disponer de un espacio en un edificio, nave industrial, o en cualquier otro lugar libre de sombra para ubicar el generador fotovoltaico [10]. En función del lugar de instalación del generador fotovoltaico, se pueden distinguir tres tipos de sistemas conectados a la red eléctrica:

- **Tejados de viviendas:**

  Se utiliza la superficie de un tejado para instalar sobre ella los módulos fotovoltaicos del generador. Son sistemas sencillos de instalar por su concepción modular donde el peso de los módulos no suele suponer una carga excesiva para cualquiera de los tejados existentes [10].

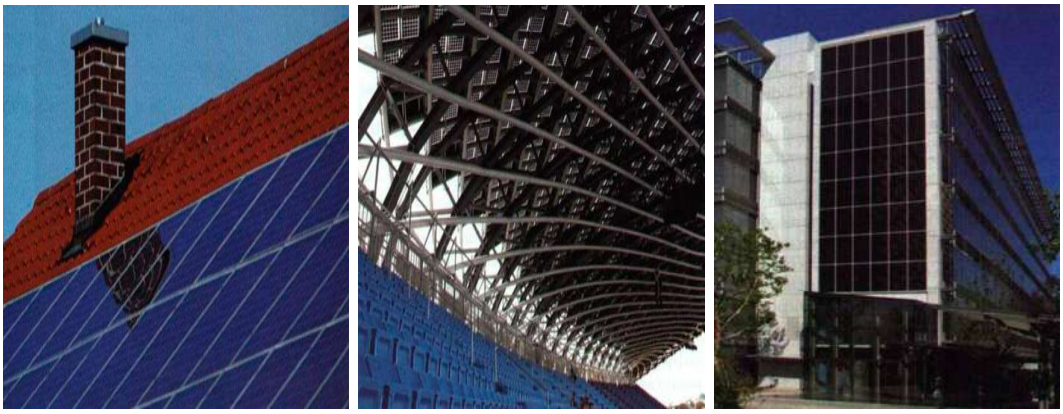
- **Plantas de generación:**

  Son aplicaciones de carácter industrial que pueden instalarse en zonas rurales no aprovechadas para otros usos o sobrepuestas en grandes cubiertas de áreas urbanas

(naves industriales, aparcamientos, zonas comerciales, áreas deportivas, etc.). En este tipo de sistemas, para aumentar la capacidad de producción se pueden utilizar sistemas de seguimiento solar [10].

- **Integración en edificios:**

Son aplicaciones donde se sustituyen a elementos arquitectónicos que incluyen el elemento fotovoltaico, y que por lo tanto son generadores de energía, aunque es prioritario el nivel de integración del elemento fotovoltaico en la estructura. A veces es necesario sacrificar parte del rendimiento energético por mantener la estética del edificio. La integración de sistemas fotovoltaicos en edificios, son aportaciones energéticas en las horas punta, contribuye a reducir la producción diurna de energía por medios convencionales [10]. Estos ejemplos se pueden ver ilustrados en la Figura A.11.



(a) Instalación en casas. (b) Instalación en estructuras. (c) Instalación en edificios.

Figura A.11: Diferentes tipos de instalación de paneles fotovoltaicos [10].

A.2. Conceptos básicos del comportamiento estadístico en series de tiempo

Introducción

En este segundo apartado, se hará énfasis en los conceptos básicos utilizados para el cálculo de los comportamientos probabilisticos de las series de tiempo.

Media aritmética

En matemáticas y estadística, la media aritmética, también llamada promedio o media, de un conjunto finito de números es el valor característico de una serie de datos cuantitativos, objeto de estudio que parte del principio de la esperanza matemática o valor esperado, se obtiene a partir de la suma de todos sus valores dividida entre el número de sumandos, se calcula mediante la ecuación (A.5). Cuando el conjunto es una muestra aleatoria recibe el nombre de media muestral, siendo uno de los principales estadísticos muestrales [33].

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \tag{A.5}$$

Dode:

$\bar{x}$: Es la media aritmética.

x_i: Son los elementos de la serie.

n: Es el índice superior de la sumatoria.

N: Son el número total de elementos.

i: Es el índice inferior de la sumatoria.

Desviación estándar

La desviación estándar, es la medida de dispersión más común que indica qué tan dispersos están los datos con respecto a la media, se calcula mediante la ecuaciónes (A.6) y (A.7). Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos [12].

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X-\mu)^2}{N}} \tag{A.6}$$

Dode:

σ: Es la desviación estándar de una población.

X: Son los elementos de la serie.

n: Es el índice superior de la sumatoria.

N: Son el número total de elementos.

i: Es el índice inferior de la sumatoria.

μ: es la media.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{N}} \tag{A.7}$$

Dode:

s: Es la desviación estándar de una muestra.

x: Son los elementos de la serie.

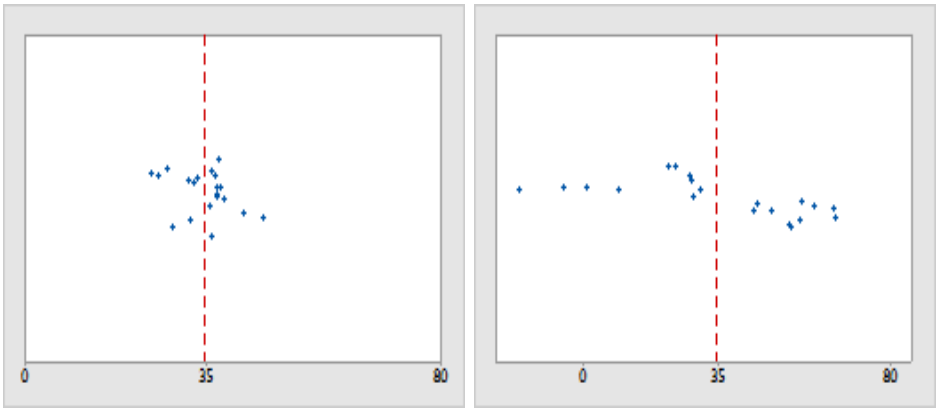
n: Es el índice superior de la sumatoria.

N: Son el número total de elementos.

i: Es el índice inferior de la sumatoria.

$\bar{x}$: es la media aritmética.

El sámbolo σ, se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra. La variación que es aleatoria o natural de un proceso se conoce comúnmente como ruido. La desviación estándar, se puede utilizar para establecer un valor de referencia para estimar la variación general de un proceso, ver A.12 [12].



(a) Hospital 1. (b) Hospital 2.

Figura A.12: Ejemplo de la desviación estándar en tiempos de egreso de dos hospitales [12].

La figura A.12, muestra donde los administradores dan seguimiento al tiempo de egreso de los pacientes tratados en las áreas de urgencia de dos hospitales. Aunque los tiempos de egreso promedio son aproximadamente iguales (35 minutos), las desviaciones estándar son significativamente diferentes. La desviación estándar del hospital 1 es de aproximadamente 6. En promedio, el tiempo para dar de alta a un paciente se desvía de la media (línea discontinua) aproximadamente 6 minutos. La desviación estándar del hospital 2 es de aproximadamente 20. En promedio, el tiempo para dar de alta a un paciente se desvía de la media (línea discontinua) aproximadamente 20 minutos [12].

Varianza

La varianza mide qué tan dispersos están los datos alrededor de la media. La varianza es igual a la desviación estándar elevada al cuadrado y se calcula con las ecuaciones (A.8) y (A.9) respectivamente, [13].

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X-\mu)^2}{N} \tag{A.8}$$

Dode:

σ^2: Es la varianza de una población.

X: Son los elementos de la serie.

n: Es el índice superior de la sumatoria.

N: Son el número total de elementos.

i: Es el índice inferior de la sumatoria.

μ: es la media.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N} \tag{A.9}$$

Dode:

s^2: Es la desviación estándar de una muestra.

x: Son los elementos de la serie.

n: Es el índice superior de la sumatoria.

N: Son el número total de elementos.

i: Es el índice inferior de la sumatoria.

$\bar{x}$: es la media aritmética.

Monitorear la varianza es esencial en las industrias de manufactura y control de calidad, porque con la reducción de la varianza del proceso, aumenta la precisión y disminuye el número de defectos [13].

Por ejemplo, la presición en la fabricación de clavos para carpintería. Una fábrica produce clavos para carpintería que miden 50 mm de largo. Un clavo cumple con las especificaciones, si su longitud no difiere en más de 2 mm del valor objetivo de 50 mm. La fábrica utiliza dos tipos de máquina para producir los clavos. Ambas máquinas producen clavos con longitudes distribuidas normalmente y una longitud media de 50 mm. Sin embargo, los clavos de cada máquina tienen varianzas diferentes: la máquina A, con la distribución de línea continua en la figura A.13, produce clavos con una varianza de 9 mm^2, mientras que la máquina B, con la distribución de la línea de puntos en la figura A.13, produce clavos con una varianza de 1 mm^2. Las distribuciones de la longitud de los clavos para cada máquina están superpuestas,junto con los límites de especificación verticales inferior y superior:

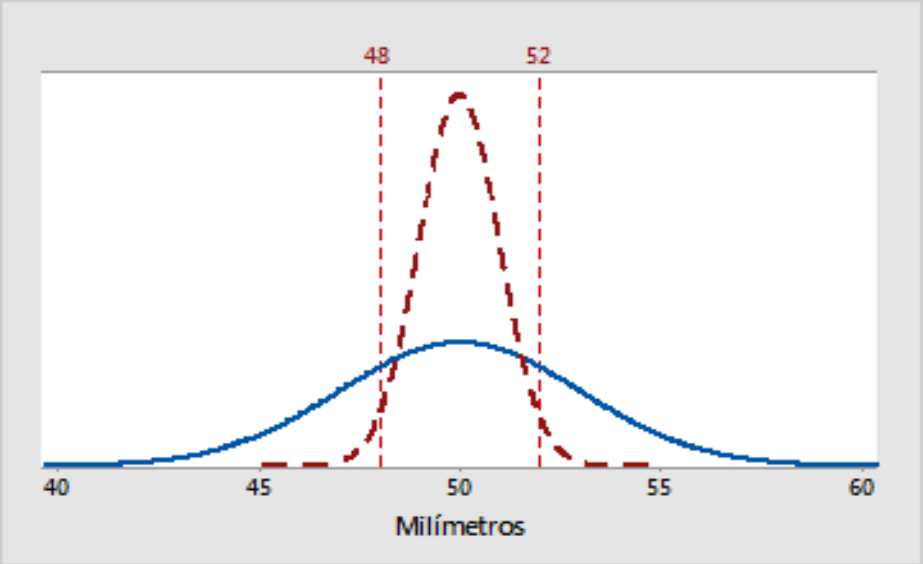


Figura A.13: Varianza en la presición de las longitudes para la fabricación de clavos para carpintería [13].

La longitud de los clavos de la máquina A tiene una variación mayor que la longitud de los clavos de la máquina B. Por lo tanto, cualquier clavo en particular de la máquina A tiene una mayor probabilidad de estar fuera de los límites de especificación que un clavo de la máquina B [13].

Diferencia entre la varianza y la desviación estándar

La desviación estándar σ mide cuánto se separan los datos, mientras que la varianza σ^2, es la media de las diferencias, con la media elevadas al cuadrado.

Ejemplo:

Tú y tus amigos han medido las alturas de sus perros (en milímetros), como se muestra en la Figura (A.14) [14]

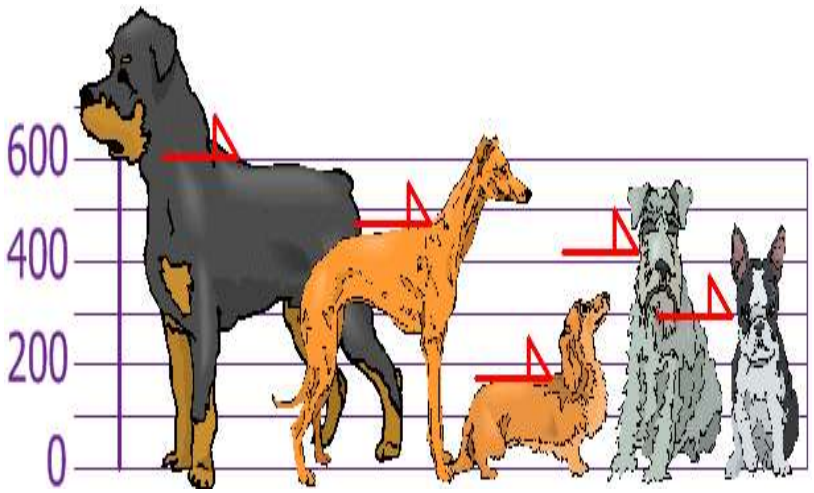


Figura A.14: Alturas de distintos perros en milímetros [14].

Las alturas (de los hombros) son: 600mm, 470mm, 170mm, 430mm y 300mm. Calcula la media, la varianza y la desviación estándar [14].

Solución:

$$\bar{x}=\frac{600+470+170+430+300}{5}=\frac{1970}{5}=394$$

Por lo que la altura media es 394 mm, ahora dibujando esto como se muestra en la figura A.15 [14], se tiene que:

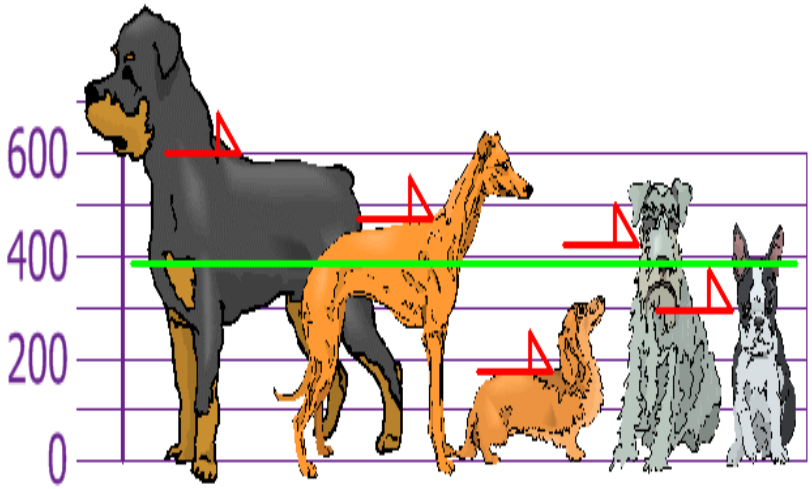


Figura A.15: La media calculada especificada en verde [14].

Ahora calculando la diferencia de cada altura con la media, como se muestra en la figura A.16

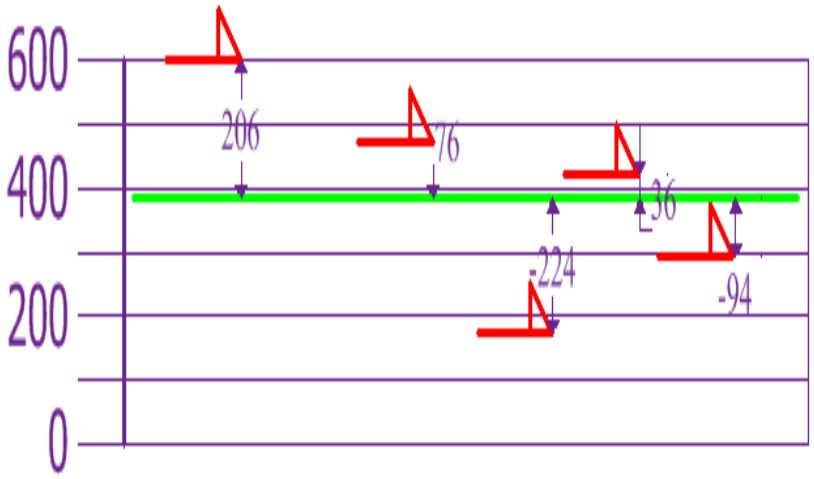


Figura A.16: Usando las diferencias para calcular la media [14].

Para calcular la varianza, tomar cada diferencia, elévarla al cuadrado y calcular la media [14]:

$$\sigma^2 = \frac{2062^2 + 762^2 + (-224)^2 + 362^2 + (-94)^2}{5} = \frac{108,520}{5} = 21704$$

Así que la varianza es 21704 y la desviación estándar es la raíz de la varianza, así que:

Desviación estándar: $\sigma = \sqrt{21704} = 147$.

Ahora eremos qué alturas están a distancias menores de la desviación estándar (147mm) y de la media [14], la figura A.17 muestra estos resultados.

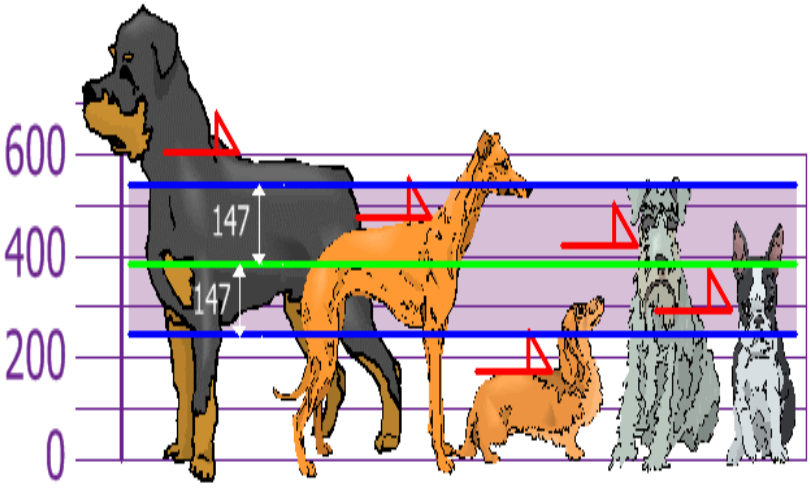


Figura A.17: Desviación estándar calculada y especificada en verde [14].

Usando la desviación estándar, se tiene una manera estándar de conocer qué es normal, extragrande o extra pequeño. Los Rottweilers son perros grandes. Y los Dachsunds son un poco más pequeños [14].

Correlación

En probabilidad y estadística, la correlación indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas. Se considera que dos variables cuantitativas están correlacionadas, cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra: si tenemos dos variables (A y B) existe correlación entre ellas si al disminuir los valores de A lo hacen también los de B y viceversa. La correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad, Fuerza, sentido y forma de la correlación La relación entre dos variables cuantitativas queda representada mediante la línea de mejor ajuste, trazada a partir de la nube de puntos. Los principales componentes elementales de una línea de ajuste y, por lo tanto, de una correlación, son la fuerza, el sentido y la forma. La fuerza extrema según el

caso mide el grado en que la línea representa a la nube de puntos; si la nube es estrecha y alargada, se representa por una línea recta, lo que indica que la relación es fuerte; si la nube de puntos tiene una tendencia elíptica o circular, la relación es débil. El sentido mide la variación de los valores de B con respecto a A; si al crecer los valores de A, lo hacen los de B, la relación es directa (pendiente positiva); si al crecer los valores de A disminuyen los de B, la relación es inversa (pendiente negativa). La forma establece el tipo de línea que define el mejor ajuste: la línea recta, la curva monotónica o la curva no monotónica [34].

Coeficiente de correlación de Karl Pearson

Dado dos variables, la correlación permite hacer estimaciones del valor de una de ellas conociendo el valor de la otra variable. Los coeficientes de correlación, son medidas que indican la situación relativa de los mismos sucesos respecto a las dos variables, es decir, son la expresión numérica que nos indica el grado de relación existente entre las dos variables y en qué medida se relacionan. Son números que varían entre los límites +1 y -1. Su magnitud indica el grado de asociación entre las variables; el valor r = 0 indica que no existe relación entre las variables; los valores 1 son indicadores de una correlación perfecta positiva (al crecer o decrecer X, crece o decrece Y) o negativa (al crecer o decrecer X, decrece o crece Y) [15]. La correlación de Karl se calcula mediante la ecuación (A.10).

$$r = \frac{\sum_{i=1}^{n} xy}{\sqrt{(\sum_{i=1}^{n} x^2)(\sum_{i=1}^{n} y^2)}} \tag{A.10}$$

Donde:

$x = X - \bar{X}$

$y = Y - \bar{Y}$

X: Son los valores de la primer serie.

$\bar{X}$: Es la media aritmética de la primer serie.

Y: Son los valores de la segunda serie.

$\bar{Y}$: Es la media aritmética de la segunda serie.

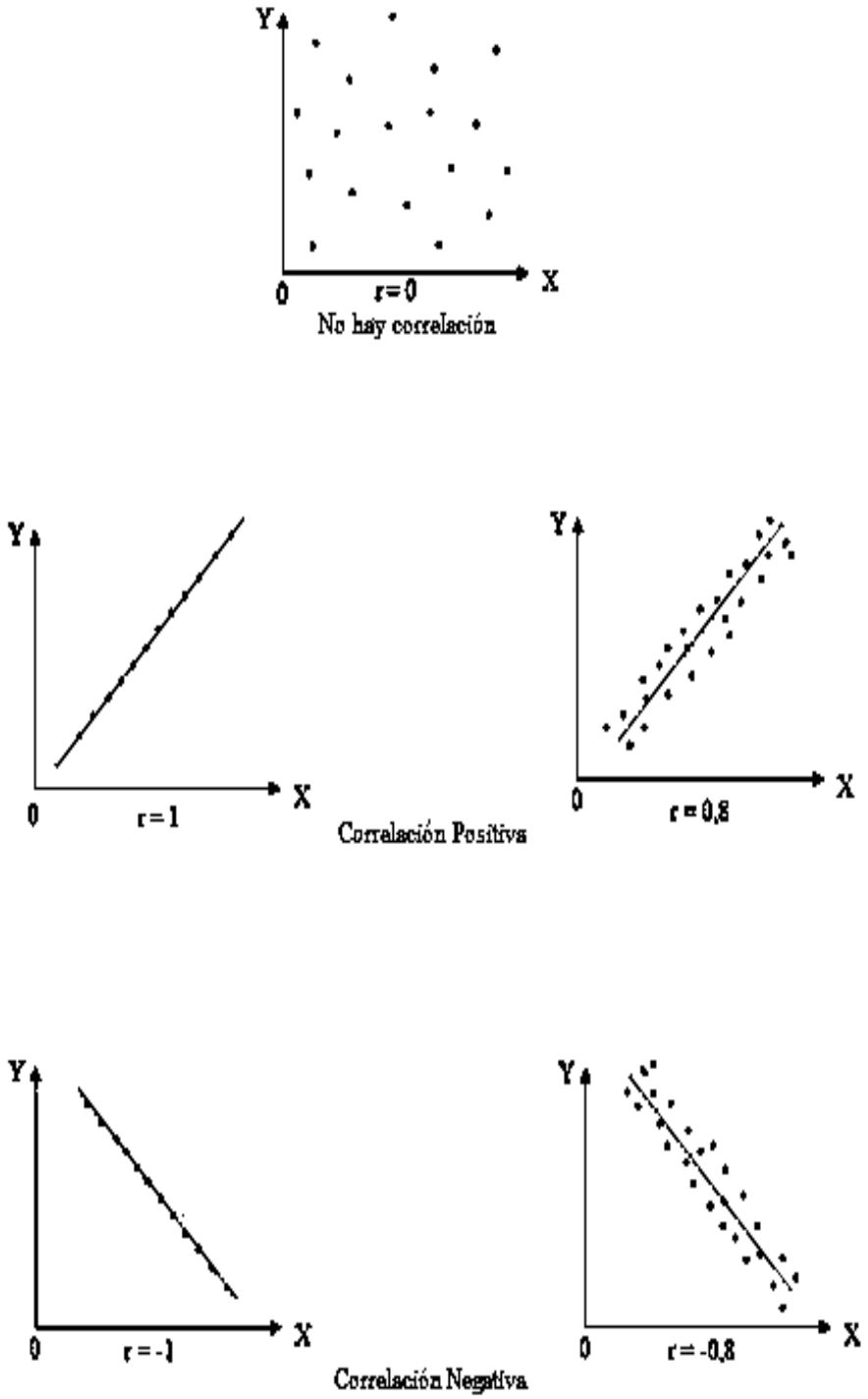


Figura A.18: Diferentes comportamientos de la correlación [15].

Para interpretar el coeficiente de correlación, utilizamos la siguiente escala mostrada en la tabla A.1 [15]:

Valor	Significado
-1	Correlación negativa grande y perfecta
-0.9 a -0.99	Correlación negativa muy alta
-0.7 a -0.89	Correlación negativa alta
-0.4 a -0.69	Correlación negativa baja
-0.2 a -0.19	Correlación negativa muy alta
-0.01 a -0.19	Correlación negativa muy baja
0	Correlación nula
0.01 a 0.19	Correlación positiva muy baja
0.2 a 0.3	Correlación positiva baja
0.4 a 0.69	Correlación positiva moderada
0.7 a 0.89	Correlación positiva alta
0.9 a 0.99	Correlación positiva muy alta
1	Correlación positiva grande perfecta

Tabla A.1: Interpretación del coeficiente de correlación [15].

Ejemplo:

Con los datos sobre las temperaturas en dos días diferentes en una ciudad (ver tabla A.2), determinar el tipo de correlación que existe entre ellas mediante el coeficiente de PEARSON [15].

x	18	17	15	16	14	12	9	15	16	14	16	18	SX=180
y	13	15	14	13	9	10	8	13	12	13	10	9	SY=138

Tabla A.2: Valores de X y Y para el ejemplo [15].

Solución:

Se calcula la media aritmética

$\bar{x} = \frac{180}{12} = 5$ $\bar{y} = \frac{138}{12} = 5$

y se llena la tabla A.3.

x	y	x= X.$\bar{X}$	$y = Y.\bar{Y}$	x^2	x.y	y
18	13	3	15	9	4.5	2.25
17	15	2	3.5	4	7	12.25
15	14	0	2.5	0	0	6.25
16	13	1	1.5	1	1.5	2.25
14	9	-1	-2.5	1	2.5	6.25
12	10	-3	-1.5	9	4.5	2.25
9	8	-6	-3.5	36	21	12.25
15	13	0	1.5	0	0	2.25
16	12	1	0.5	1	0.5	0.25
14	13	-1	1.5	1	-1.5	2.25
15	13	0	15	0	0	2.25
16	12	1	0.5	1	0.5	0.25
14	13	-1	-1.5	1	-1.5	2.25
16	10	1	-1.5	1	-1.5	2.25
18	8	3	-3.5	9	-10.5	12.25
$\Sigma = 180$	$\Sigma = 138$			72	28	63

Tabla A.3: Muestra la tabla de valores llena [15].

Se aplica la ecuación (A.10), y con esto se obtiene el resultado:

r=$\frac{28}{\sqrt{(72)(63)}} = 0.416$

Autocorrelación

La autocorrelación o dependencia secuencial, es una herramienta estadística utilizada frecuentemente en el procesado de señales. La función de autocorrelación, se define como la correlación cruzada de la señal consigo misma. La función de autocorrelación resulta de gran utilidad para encontrar patrones repetitivos dentro de una señal, como la periodicidad de una señal enmascarada bajo el ruido o para identificar la frecuencia fundamental de una señal que no contiene dicha componente, pero aparecen numerosas frecuencias armónicas de ésta. Los procesos de raíz unitaria, autorregresivos, de tendencia estacionaria y los modelos de medias móviles, son ejemplos de procesos con autocorrelación [35].

Función de autocorrelación

La función de autocorrelación (fac) y la función de autocorrelación parcial (facp), miden la relación estadística entre las observaciones de una serie temporal. Por ejemplo, el coeficiente de autocorrelación entre la variable y_t y la misma variable un período antes, $y_t - 1$, al que denominaremos coeficiente de autocorrelación de primer orden, se formula como la ecuación (A.11) [26]:

$$P_1 = \frac{cov(y_t, y_{t-1})}{\sqrt{var(y_t)var(y_{t-1})}} \tag{A.11}$$

Dado el supuesto de estacionariedad, se tiene que $var(y_t) = var(y_t - 1)$, por lo que [26]:

$$P_1 = \frac{cov(y_t, y_{t-1})}{\sqrt{var(y_t)}} \tag{A.12}$$

En general, para un desfase de k períodos se tiene que [26]:

$$P_k = \frac{cov(y_t, y_{t-k})}{\sqrt{var(y_t)}} \tag{A.13}$$

y cuando $k = 0$,

$$P_0 = 1 \tag{A.14}$$

Donde:

P: Coeficiente de autocorrelación. $cov(y_t)$: Es la covarianza que hay entre dos puntos de una serie de tiempo.

$var(y_t)$: Es la varianza de un elemento de la misma serie de tiempo.

k: Es un índice.

A efectos de la identificación del modelo, debemos comparar el valor que esta función presentaría para los distintos modelos teóricos, con una estimación de ésta para nuestra serie. Pues bien, el estimador muestral de la fac, para el que utilizaremos la expresión r_k, viene dado, con ciertas condiciones y aproximaciones que no trataremos aquí por la ecuación (A.15) [26]:

$$r_k = \frac{\sum_{t=1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{A.15}$$

Donde:

y_t: Serie de tiempo.

k: Índice.

$\bar{y}$: Media aritmética.

r_k: Coefiente de autocorrelación.

n: Índice superior de la sumatoria.

t: Índice inferior de la sumatoria.

Apéndice B

Cálculo de los coeficientes del modelo AR(P)

Introducción

En este apartedo, trataremos brevemente al modelo AR(p) con la intención de aprender a calcular sus coeficientes, aplicando y resolviendo un ejemplo.

Obtención de los coeficientes del modelo AR

Las observaciones de una serie cronológica están asociadas a los diferentes valores que alcanza una magnitud que varía en el tiempo. Tales valores se guardan en una lista de la forma $y1, y2, ..., yn$. El ajuste autorregresivo de orden p, AR(p), consiste en suponer que los valores registrados han sido generados por un modelo subyacente, tal como la ecuación (B.1) [16]

$$y_t = \theta_0 + \theta y_{t-1} + ... + \theta_p y_{t-p} + \epsilon_t = \sum_{i=1}^{p} \theta_i y_{t-i} + \epsilon_t \tag{B.1}$$

Donde:

θ_i: Son los coeficientes de auto-regresión

y_t: Es la serie bajo investigación

P: Es el orden (longitud) del modelo.

ϵ_t: Es ruido, casi siempre se supone que es ruido blanco Gaussiano.

Esto es, la lectura que se obtiene en la etapa i-ésima depende linealmente de las últimas p observaciones, más un error aleatorio representado por ϵ_t. Tal como ha quedado especificado, el modelo AR(p) tiene $p+2$ parámetros a estimar a partir de los datos observados: los $p+1$ coeficientes autorregresivos $\theta_0, \theta_1, \theta_2, ..., \theta_p$. El método de estimación adoptado aquí es el de los mínimos cuadrados, que consiste en calcular los parámetros autorregresivos de forma tal que minimicen el error cuadrático, para ello utilizamos la ecuación (B.3) [16].

$$EC = \sum_{t=p+1}^{n} (y_t - \theta_0 - \sum_{i=1}^{p} \theta_i y_{t-i})^2 \tag{B.2}$$

donde EC es el error cuadratico

$$\hat{\theta} = \begin{bmatrix} \hat{\theta_0} \\ \hat{\theta_1} \\ \vdots \\ \hat{\theta_p} \end{bmatrix} = (X^T X)^2 X^T Y \tag{B.3}$$

siendo:

$$Y = \begin{bmatrix} y_{p+1} \\ y_{p+2} \\ \vdots \\ y_n \end{bmatrix} \tag{B.4}$$

$$X = \begin{vmatrix} 1 & y_p & y_{p-1} & \cdots & y_1 \\ 1 & y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{vmatrix} \tag{B.5}$$

Ejemplo:

El siguiente ejemplo presenta en total de 1000 muestras de una suma de 4 sinusoides [16],

ver figura B.1.

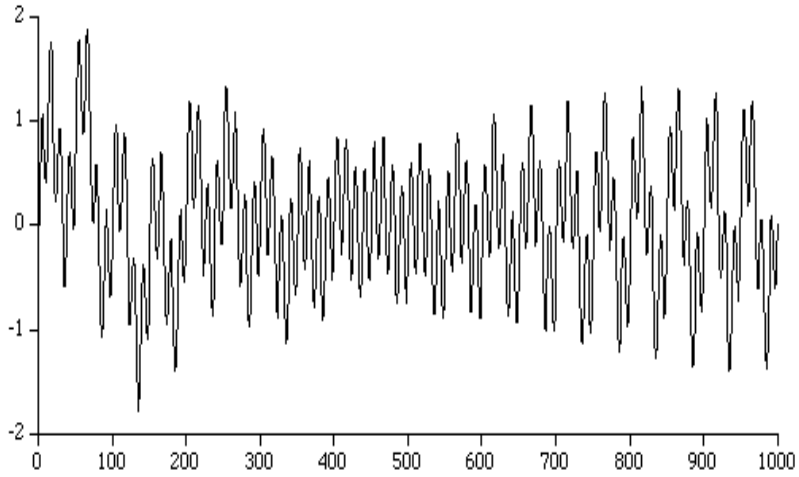


Figura B.1: Comportamiento gráfico resultante de la suma de 4 sinusoidales [16].

Al aplicar un modelo AR de orden 1 se obtiene un coeficiente de 0.941872, esto no es totalmente sorprendente, ya que está diciendo que, al mirar solo un término de la serie, el siguiente término de la serie probablemente sea casi el mismo, es decir: $xt + 1 = 0.941872xt$

La tabla B.1 proporciona los coeficientes para diferentes de órdenes en el modelo del ejemplo anterior [16].

Orden/Coeficientes	1	2	3	4	5	6
1	0.9418					
2	1.8261	-0.9388				
3	2.7532	-2.7403	0.9855			
4	3.7367	-5.4742	3.7311	-0.9967		
6	4.2590	-6.2327	2.1073	2.9697	-1.4212	-2.5918

Tabla B.1: Cálculo de coeficientes para diferente orden del modelo AR [16].

A medida que el orden aumenta, las estimaciones generalmente mejoran (esto puede no ser necesariamente así, por ejemplo para los datos ruidosos y órdenes AR grandes). A menudo es útil trazar el error RMS entre la serie estimada por los coeficientes AR y la serie real.

Apéndice C

Códigos con funciones de Matlab y phyton usados en la tesis

C.1. Código en Matlab para calcular los coeficientes del modelo AR(P)

```
yt= serie;
n=input('Dame el numero de elementos de tu serie de tiempo:')
p=input('Dame el dorden P:')
p2=p;
p3=p;
%Calculo de Y
renglon=1;
columna=1;
indice=0;
for p=p:1:n
p=p+1;
Y(renglon,columna)= yt(p,columna);
renglon=renglon+1;
if p==n
break;
end;
end;
Y;
renglon=1;
columna2=1;
indice=0;
indice2=indice;
fin=1;
resta=0;

%Calculo de X
for renglon=1:n-p2
for columna2=1:p2+1
if columna2==1
X(renglon,columna2)=1;
```

```
end;
if columna2 ~=  1
p2;
X(renglon,columna2)=yt(p2+indice,columna;
columna2=columna2 + 1;
indice= indice-1;
end;
columna2=columna2+1;
end;
p=p3;
indice=indice2+renglon;
renglon= renglon + 1;
end;

%Calculo de los coeficientes
X;
phi= inv(X'*X)*X'*Y
```

C.2. Código en Matlab para hacer predicción usando ARIMA

```
%Codigo para quitar ceros de la serie de tiempo
Y=serie;
j=1;
i=1;
y(i)=0;
for i=1:length(Y)
    if Y(i)~=0
       y(j)=Y(i);
    end;
    if Y(i)==0
       j=j-1;
    end;
    i=i+1;
    j=j+1;
end;
y=y';

%Funciones para hacer predicion
t=[1:1:length(Y)]';
mdl=arima(3,3,3);
sys=estimate(mdl,Y(1:end-1440));
predi=forecast(sys,1440);
set(0,'DefaultLineLineWidth',1);set(0,'DefaultAxesFontName','Times New Roman');

%Funciones para graficar

plot(t,Y,t(end-1440*2:end-1440),[Y(end-1440*2);predi])
title('Prediccion de 1440(1dia) muestras de Radiacion Solar ARIMA(3,3,3)')
legend('Radiacion','Prediccion')
ylabel('Radiacion(W/m2)')
xlabel('Tiempo(min)')
grid on
```

C.3. Código en Python para hacer predección usando K-NN y SVM

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import math
%matplotlib inline

data=pd.read_excel('weather (6).xlsx',names=('fecha','radiacion','velocidad','temperatura','humedad','presion'),

data.head()

plt.plot(data.radiacion)
data.dtypes

for x in data.fecha:
print(x.month, x.day, x.hour, x.minute)
break;

x= np.array([[x.month, x.day, x.hour, x.minute] for x in data.fecha])
x[-1]

y= np.array(data.radiacion)
plt.plot(y)

x[0],y[0]

%Prediccion usando K-NN
from sklearn.neighbors import KNeighborsRegressor
knn=KNeighborsRegressor(n_neighbors=7,n_jobs=2, weights='distance')

knn.fit(x,y)
n=120
xt=x[:-n,:]
yt=y[:-n]

yv=y[-n:]
xv=x[-n:,:]
yp=knn.predict(xv)

plt.plot(yv,color='blue',label='Radiacion')
plt.plot(yp,color='red',label='Prediccion')
plt.ylabel('Radiacion(w/m2)')
plt.xlabel('Tiempo(min)')
plt.title('Prediccion de 1440(1dia) muestras de Radiacion Solar K-NN7distance')
plt.grid()
plt.legend()

plt.savefig("Radiacionknndistancia(1440).png")
plt.savefig("Radiacionknndistancia(1440).eps")

%Prediccion usando SVM
from sklearn.svm import SVR

svr_rbf = SVR(kernel='rbf')
svr_rbf.fit(xt, yt)

plt.plot(yv)
plt.plot(y_pred_train)
```

```
plt.legend(loc='best');
plt.title("prediccion de 12(1hr) muestras de Temperatura SVM polynomial")
plt.grid()
plt.plot(yv,color='blue',label='Temperatura')
plt.plot(y_pred_train,color='red',label='prediccion')
plt.ylabel('Temperatura')
plt.xlabel('Tiempo(min)')
plt.legend()

plt.savefig("temp12msvpolynomial.png")
plt.savefig("temp12msvpolynomial.eps")
```

C.4. Código en python para hacer predicción usando K-NN y optimizarla usando el Teorema de Takens

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

%Teorema de Takens
import more_itertools as mit

def wsplit(data,m=3,tau=1):
l,i=len(data),0
n=l-m*tau
x,y=[[] for k in range(n)],[0 for k in range(n)]
while i+m*tau<l:
ind=[0 for k in range(m)]
for j in range(m):
val= i+j* tau
ind[j]=val
x[i]=data[ind]
y[i]=data[ind[-1]+tau]
i+=1
return np.array(x),np.array(y)

X=pd.read_excel("weather (6).xlsx", names=['fecha','radiacion','velocidad','temperatura','humedad','presion'],
skiprows=[0])
X.head(3)

data=np.array(X.radiacion[X.radiacion>0])

%Datos de entrenamiento
train=data[:-720]

%Datos de prueba/validacion
test=data[-720:]

plt.plot(test)

k=90
%Asumo que el valor en t+1 es modelada por los valores de t-k hasta t
x,y=wsplit(train,k)

%prediccion usando K-NN
from sklearn.neighbors import KNeighborsRegressor as nnr
knnr=nnr(n_neighbors=15, weights='distance')
```

```
knnr.fit(x,y)

def forecast(regressor,window,n=720):
pred=[]
window=list(window)
while len(pred)<n:
val=regressor.predict([window])
pred.append(val[0])
window.append(val[0])
window=window[1:]
return np.array(pred)

yp=forecast(knnr,y[-k:])

plt.plot(test)
plt.plot(yp)
plt.legend(labels=['Actual','Prediccion'])
plt.ylabel('Radiacion(w/m2)')
plt.xlabel('Tiempo(min)')
plt.title('Prediccion de 720(12hr) muestras de Radiacion knn,distancia,Takens')
plt.grid()
plt.savefig("Takens(90(15).png")
plt.savefig("Takens(90(15).eps")
```

Referencias

[1] "La historia de la energía solar fotovoltaica," 2014.

[2] IRENA, "China construye la granja solar más grande del planeta," 2016.

[3] J. Pareja, "Análisis de series de tiempo.," 2011.

[4] "Serie temporal," 2018.

[5] "Modelos de pronóstico," 2018.

[6] "Machine learning y support vector machines: porque el tiempo es dinero," 2016.

[7] O. Manzanilla, "Optimization & machine learning," 2008.

[8] "Distancia euclídea," 2018.

[9] "Regla de los k vecinos más cercanos," 2018.

[10] G. S. Agustín Castejón, Instalaciones solares fotovoltaicas.

[11] "Influencia de la irradiación y temperatura sobre una placa fotovoltaica," 2014.

[12] "Qué es la desviación estándar," 2018.

[13] "Qué es la varianza.," 2018.

[14] "Varianza y desviación estándar," 2017.

[15] M. O. S. Ibujes, "Coeficiente de correlación de karl pearson," 2018.

[16] P. J. Diggel, Time Series. A Biostatistical Introduction. 1992.

[17] "La demanda de energía crece mientras el mix en'ergetico continúa diversificandose," Feb. 2018.

[18] M. M. Pantoja, "China construye la granja solar más grande del planeta.," 2016.

[19] J. A. Roca, "Las 20 mayores plantas fotovoltaicas del mundo: China, india y eeuu arrasan," 2018.

[20] J. A. Resendiz, Las máquinas de soporte vectorial para la identificación en linea. PhD thesis, Instituto politecnico nacional, 2006.

[21] G. T. Serrano, "Machine learning approach to forecast global solar radiation time series," Master's thesis, The university of new Mexico, 2016.

[22] E. Klages, "Time series based forecasting of renewable power infeed for operation of microgrids," Master's thesis, Technische Universitat Berlin, 2016.

[23] J. Vermorel, "Definición series de tiempo," 2012.

[24] alvaro alvarez, "modelo probabilístico," 2013.

[25] R. H. Alex Sergejew, Nick Hawthorn, "Auto-regression analysis (ar)," 1998.

[26] J. R. V. Antonio Pulido, Ana M López, "Curso de predicción económica y empresarial," 2004.

[27] R. E. L. Briega, "Machine learning con python," 2015.

[28] Jacobsoft, "Regresión de soporte vectorial," *2019*, 2019.

[29] F. Takens, Dynamical System and Turbulence, Warwick 1980. 2019.

[30] "Radiación solar," 2018.

[31] M. Lizardi, "Desaprovecha méxico radiación solar: Miguel ángel meneses," May 2017.

[32] "Definición de temperatura," 2015.

[33] "Media aritmética," 2018.

[34] "correlación," 2018.

[35] "Autocorrelación," 2018.