Marco general para la extracción de información y estimación de radiación solar diaria

General framework for the extraction of information and estimation of daily solar radiation

D. F. Muñoz-Torres, O. D. Montoya-Giraldo; S. M. Sergio-Arturo.
DOI: https://orcid.org/0000-0003-1658-8921
Artículo de investigación científica y tecnológica

***Abstract*—This article presents a comparative study resulting from the design and simulation of a system of prediction of climatic conditions using machine learning models, in which the results obtained when using a database of environmental conditions are compared with another database generated from the treatment of the data through the analysis by main components. In the first phase of the study, metadata is generated through the subspaces created with the analysis by main components, a second phase consists of developing a system of prediction of climatic conditions using several machine learning models, which will use as a resource the original data and metadata generated in the first phase of the study, in the final phase of the study, both results are compared with the aim of observing the behavior of solar radiation inference systems. The proposed data processing strategy allows to extract information from environmental databases facilitating interpretation and observation as a time series of data, additionally, it is possible to build an experimental frame of reference for the inference of solar radiation using different supervised learning techniques on the generated databases.**

***Index Terms*— Interpretation of climate data, Principal Component Analysis, Solar Energy, Supervised learning, Weather Inference.**

***Resumen*— Este articulo presenta un estudio comparativo resultado del diseño y simulación de un sistema de predicción de condiciones climáticas usando modelos de aprendizaje automático, en el cual, se confrontan los resultados obtenidos al usar una base de datos de condiciones ambientales, con otra base de datos generada a partir del tratamiento de los datos mediante el análisis por componentes principales. En la primera fase del estudio, se generan metadatos a través de los subespacios creados con el análisis por componentes principales, una segunda fase consta de elaborar un sistema de predicción de condiciones climáticas usando varios modelos de aprendizaje de máquina, los cuales, usarán como recurso los datos originales y los metadatos generados en la primera fase del estudio, en la fase final del estudio, se comparan ambos resultados con el objetivo de observar el comportamiento de los sistemas de inferencia de la radiación solar. La estrategia de tratamiento de datos propuesta permite extraer información de las bases de datos ambientales facilitando la interpretación y observación como serie temporal de datos, adicionalmente, se logra construir un marco de referencia experimental para la inferencia de la radiación solar usando diferentes técnicas de aprendizaje supervisado sobre las bases de datos generadas.**

***Palabras claves*— Análisis por componentes principales, Aprendizaje supervisado, Energía solar, Interpretación datos climáticos, Inferencia de Condiciones climáticas.**

ACRÓNIMOS

KNN	**:**	**Clasificador K-vecinos más cercanos**
LDA	**:**	**Análisis lineal discriminante**
PCA	**:**	**Análisis por componentes principales**
PV	**:**	**Energía solar fotovoltaica**
QDA	**:**	**Análisis discriminante cuadrático**
SVM	**:**	**Máquina de soporte vectorial**

I. INTRODUCTION

Las redes eléctricas están cambiando de una conexión vertical clásica, es decir, generación, transmisión, distribución y comercialización a conexiones horizontales donde se integran nuevas fuentes de energía tales como las energías renovables [1]–[3], dentro de estas tecnologías podemos encontrar la solar fotovoltaica (PV), la cual, al ser integrada al sistema de transmisión pueden proporcionar un confiable soporte de energía [4]. El desarrollo de modelos de generación, transmisión y distribución de energía eléctrica que incluya sistemas solares se encuentra en una etapa de desarrollo, los datos de radiación solar de las mediciones en tierra o las imágenes de satélite suelen estar disponibles durante períodos de tiempo limitados, y a menudo se trabaja con datos históricos que reflejan de manera cercana pero no precisa las condiciones actuales ambientales, generando problemas de calidad de datos

This manuscript was sent on October 28, 2021 and accepted on February 23, 2022. D. F. Muñoz-Torres, Ingeniería Electrónica, Fundación Universitaria Tecnológico Comfenalco, Cartagena de Indias, Colombia. (dmunozt@tecnocomfenalco.edu.co)

O. D. Montoya-Giraldo, Facultad de Ingeniería Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia; (odmontoyag@udistrital.edu.co).

S. M. Sergio-Arturo, Ingeniería Electrónica, Fundación Universitaria Tecnológico Comfenalco, Cartagena de Indias, Colombia. (ssabach@tecnocomfenalco.edu.co)

y/o lagunas de datos [5]–[8].

El análisis por componentes principales (PCA, Principal component Analisys), es una técnica ampliamente usada en aplicaciones que asocian una reducción de dimensionalidad, también son usados para evaluar la calidad de datos y problemas de datos faltantes, mejorando la interpretación de variables y las dependencias ocultas que pueden presentarse al interior de los datos [9]. Una adecuada interpretación de las variables ambientales que influya los rango de radiación solar es esencial para un diseño adecuado de un modelo de predicción y, en futuro, un modelo de despacho energético [10].

En este trabajo el interés principal es diseñar una estrategia de tratamiento de datos ambientales mediante PCA para extraer la información inmersa en las complejas bases de datos ambientales con el objetivo de entrenar una máquina que permita estimar las condiciones climáticas para el día siguiente [11]–[17].

En este sentido, se proponen tres etapas de diseño y simulación con el fin de analizar de forma generalizada el comportamiento de los datos y establecer un marco de referencia a seguir para el tratamiento de datos ambientales: la primera etapa de diseño consiste en la creación de una base de metadatos a partir de los datos de condiciones ambientales e irradiación solar mediante PCA; analizar su comportamiento. Una segunda etapa de diseño consiste en entrenar diferentes máquinas usando los metadatos y los datos originales y observar cómo son las salidas en cada modelo dada la entrada mediante la validación cruzada. Finalmente, la tercera etapa consiste en la comparación de los resultados evaluando las ventajas y desventajas del modelo con metadatos y datos.

Las diferentes secciones que comprende este documento están organizadas de la siguiente forma: en la sección II, se presenta una revisión del estado del arte. En la sección III, se presenta los modelos matemáticos relacionados con el análisis por componentes principales al igual que las estrategias de aprendizaje supervisado usadas en el entrenamiento de modelos de clasificación e inferencia. En la sección IV, se presenta la metodología. En la sección V se presenta los resultados y discusión de resultados. Finalmente, la sección VI presente las conclusiones derivadas de este trabajo de investigación seguida de las referencias citadas en el documento.

II. Revisión del Estado del Arte.

En la literatura relacionada con las energías renovables, el aprendizaje supervisado ha sido usado en múltiples aplicaciones para la generación de modelos estadísticos, estos se encuentran estrechamente ligados a los observatorios ambientales y en sí, a los avances que se den con respecto al análisis de bases de datos de condiciones ambientales y en la minimización de los errores de interpretación y datos faltantes. Por tal motivo, esta breve revisión del estado del arte expone algunos trabajos relacionados al tratamiento de datos ambientales que tienen como objetivo una aplicación práctica en los sistemas de generación de energía eléctrica, y quienes se centran en el uso de los PCA y técnicas de aprendizaje supervisado para el análisis de datos ambientales.

Revisando diferentes tecnologías de generación de energía eléctrica mediante fuentes renovables, encontramos que estas juegan un papel esencial en la estructura energética a nivel mundial, en donde, la generación es dependiente de las condiciones ambientales, por esta razón, los estudios relacionados usan técnicas estadísticas para la inferencia de las condiciones ambientales como por ejemplo los modelos bayesianos [9], por lo cual, establecer un marco referencial para toma de decisiones precisas ha sido tarea prioritaria durante los últimos años [10], [18], y en lo cual, el problema ha sido abordado usando técnicas computacionales como lo es el PCA en combinación con técnicas heurísticas para establecer un modelo matemático adecuado, este se logra al fijar la cantidad de coeficientes adecuados para la técnica usada [11], [17].

En [12] se usa un método de cuadricula para interpolar los datos de condiciones ambientales, en donde el interpolador es diseñado con PCA. Otros estudios presentados como en [13]–[15], [19] proponen estimadores para la irradiación solar basados en PCA, en donde el estimador permite generar un modelo ajustado. Otros estudios como el presentado en [16] se enfocan en mejorar las técnicas de recolección de información por parte de los satélites a través de técnicas de análisis de múltiples bases de datos en las cuales incluyen el uso de PCA.

Finalmente, los estudios relacionados con el crecimiento sostenible que relacionan la energía PV también incluyen el uso de PCA en el análisis de datos históricos para establecer modelos que se usaran en tiempo real para monitorear las instalaciones [20]–[22]

A diferencia de los trabajos anteriores en los cuales el problema ha sido estudiado para una ubicación especifica o se ha utilizado el PCA como una herramienta de pre-análisis de los datos, en este estudio se propone la formulación de un modelo en el cual se consideran datos de diferentes continentes, la trasformación de los datos de entrada mediante PCA fijando como salida la irradiación solar, en el cual se observara, seleccionara y fijara la cantidad adecuada de componentes del PCA para describir de forma adecuada la base de datos y las variables de salida.

III. Modelos Matemáticos.

A continuación se hace una descripción de los modelos matemáticos/estadísticos usados en el artículo de investigación.

A. Análisis por componentes principales.

El Análisis por componentes principales o PCA es una técnica utilizada en aplicaciones tales como: la reducción de dimensionalidad, la compresión de datos con pérdida, la extracción de características y la visualización de datos [23]. También se conoce como la transformación de Karhunen-Loeve.

Hay dos definiciones de PCA de uso común que dan lugar al mismo algoritmo. El PCA se puede definir como la proyección ortogonal de los datos en un espacio lineal de menor dimensión, conocido como el subespacio principal, de manera que se maximiza la varianza de los datos proyectados. De forma equivalente, se puede definir como la proyección lineal que minimiza el costo promedio de proyección, definido como la distancia cuadrática media entre los puntos de datos y sus proyecciones [24], [25]

1) Formulación de varianza máxima

Considere un conjunto de datos de observaciones $\{X_n\}$ donde n = l , ... , N y X_n es una variable euclidiana con dimensionalidad D. El objetivo es proyectar los datos en un espacio que tenga dimensionalidad M<D maximizando la varianza de los datos proyectados (La proyección en un espacio unidimensional (M = 1)).

Se define la dirección de este espacio usando un vector D-dimensional $\mathbf{u_1}$, que por conveniencia (y sin pérdida de generalidad) se elige como un vector unitario de modo que $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$. Cada punto de datos $\mathbf{X}_n$ se proyecta sobre un valor escalar $\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n$. La media de los datos proyectados es $\mathbf{u}_1^{\mathrm{T}}\bar{\boldsymbol{x}}$ donde $\bar{\boldsymbol{x}}$ es la media del conjunto muestral dada por la ecuación (3.1)

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{3.1}$$

y la variación de los datos proyectados viene dada por la ecuación (3.2)

$$\frac{1}{N}\sum_{n=1}^{N}\{u_1^T x_n - u_1^T \bar{x}\}^2 = u_1^T S u_1 \tag{3.2}$$

donde **S** es la matriz de covarianza de datos definida por:

$$S = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})(X_n - \bar{X})^T \tag{3.3}$$

Al maximizar la varianza proyectada $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$ con respecto a $\mathbf{u_1}$. Restringido por la condición de normalización $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$. Para hacer cumplir esta restricción, se introduce un multiplicador de Lagrange denotado por λ_1 y luego se hace una maximización no restringida de

$$u_1^T S u_1 + \lambda_1(1 - u_1^T u_1) \tag{3.4}$$

Al establecer la derivada con respecto a $\mathbf{u}_1$ igual a cero, se tiene la ecuación (3.5)

$$S u_1 = \lambda_1 u_1 \tag{3.5}$$

La ecuación 3.5 muestra que $\mathbf{u}_1$ debe ser un vector propio de S. Si multiplicamos por la izquierda por $\mathbf{u}_1^{\mathrm{T}}$ y se usa de $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$, la varianza queda dada por

$$u_1^T S u_1 = \lambda_1 \tag{3.6}$$

Por lo que la varianza será máxima cuando se establece $\mathbf{u}_1$ igual al vector propio que tiene el valor propio más grande λ_1. Este vector propio se conoce como el primer componente principal. Si se considera el caso general de un espacio de proyección de dimensión M, la proyección lineal óptima para la cual se maximiza la varianza de los datos proyectados se define ahora por los M vectores propios $\mathbf{u}_1$,..., $\mathbf{u}_M$ extraídos de la matriz de covarianza de datos. Con el conocimiento de la covarianza se realizaran las mediciones de las componentes generadas mediante el PCA, permitiendo de esta forma elegir la cantidad adecuada de coeficientes perdiendo la minima cantidad de información de los datos, para esta investigación se considera tomar las 3 primeras componentes principales que según estudios previos encontrados como se ha mencionado anteriormente logra captar el 90% de la información.

2) Formulación de error mínimo cuadrado

Una formulación alternativa de PCA basada en la minimización del error de proyección se discute en este apartado. Para hacer esto, se introduce un conjunto orto-normal completo de vectores base de dimensión D $\{\mathbf{u}_i\}$ donde i = 1,..., D que satisfacen

$$u_i^T u_j = \delta_{ij} \tag{3.7}$$

Debido a que esta base es completa, cada punto de datos se puede representar exactamente mediante una combinación lineal de los vectores base

$$x_n = \sum_{i=1}^{D} \propto_{ni} u_i \tag{3.8}$$

Donde los coeficientes $\propto_{ni}$ serán diferentes para diferentes puntos de datos. Esto simplemente corresponde a una rotación del sistema de coordenadas a un nuevo sistema definido por $\{\mathbf{u}_i\}$, y los componentes D originales $\{x_{n1},..., x_{nD}\}$ son reemplazados por un conjunto equivalente $\{\propto_{n1},..., \propto_{nD}\}$. Tomando el producto interno con $\mathbf{u}_j$, y haciendo uso de la propiedad orto-normal, obtenemos $\propto_{nj} = x_n^T u_j$, por lo que sin pérdida de generalidad podemos escribir la ecuación 3.9.

$$x_n = \sum_{i=1}^{D}(x_n^T u_i)u_i \tag{3.9}$$

B. Técnicas de procesamiento de datos

A continuación, se hace una descripción de las técnicas de entrenamiento supervisado usadas en el artículo de investigación para la inferencia de las condiciones climáticas.

1) Análisis Linear Discriminante (LDA)

El análisis linear discriminante determina las clases entre las variables que componen una base de datos dos o más clases, el modelo generado por el LDA de igual forma también permite la clasificación y predicción de pertenencia a un grupo específico para nuevas observaciones. Para cada uno de los grupos, el LDA asume que las variables explicativas son normalmente distribuidas con matrices de covarianza iguales. El LDA más simple tiene dos grupos. A discriminar entre ellos, una función discriminante lineal que pasa a través de él[26], [27].

El modelo estándar LDA supone que la distribución condicional de X|Y es normal multivariante con vector medio μ_y y matriz de covarianza común Σ. Por lo cual, la ecuación 3.10 se definen como:

$$P(1|x) = \frac{1}{1 + (e^{\alpha+\beta_x})^{-1}} \tag{3.10}$$

Donde los coeficientes α y β son respectivamente:

$$\beta = (\mu_1 - \mu_0)^T \sum^{-1} \tag{3.11}$$

$$\alpha = -\log\frac{\pi_1}{\pi_0} + \frac{1}{2}(\mu_1 - \mu_0)^T \sum^{-1} (\mu_1 - \mu_0) \tag{3.12}$$

2) Análisis discriminante cuadrático (QDA)

QDA es similar al LDA, excepto que asume que la matriz de covarianza puede ser diferente para cada clase y, por lo tanto, se estima la matriz de covarianza por separado para cada una de las clases k, con k = 1, 2,..., K[28].

La ecuación para la función discriminante cuadrática se presenta en 3.13:

$$\begin{aligned}\delta_k(x) = &-\frac{1}{2} log \left|\sum\nolimits_k\right| \\ &-\frac{1}{2}(x - \mu_k)^T \sum\nolimits_k^{-1} (x - \mu_k) \\ &+ log\pi_k\end{aligned} \tag{3.13}$$

Esta función discriminante cuadrática es parecida a la función discriminante lineal, excepto que, debido a que la matriz de covarianza no es idéntica, no se pueden descartar los términos cuadráticos. Esta función discriminante es una función cuadrática y contendrá términos de segundo orden.

3) Clasificadores Naive Bayes

Los clasificadores Naive Bayes son altamente escalables. El entrenamiento de máxima verosimilitud se puede realizar evaluando una expresión de forma cerrada, lo cual reduce el tiempo de ejecución en comparación con aproximaciones iterativas.

Bayes Naive es un modelo de probabilidad condicional bayesiana, en donde dada una instancia del problema a clasificar, representada por un vector $x = (x_{1,\dots,} x_n)$, asigna a esta instancia probabilidades como se muestra en la ecuación 3.14[29].

$$p(C_k|x_{1,\dots,}x_n) \tag{3.14}$$

Para cada uno de los K posibles resultados o clases C_k, se tiene la probabilidad conjunta mostrada en la ecuación 3.15.

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{3.15}$$

4) Máquinas de soporte vectorial (SVM)

Las máquinas de soporte Vectorial (SVM), es un algoritmo de aprendizaje supervisado que asigna etiquetas a los objetos de cada clase. SVM maximiza una función matemática basada en los datos mediante el uso de un hiperplano y márgenes máximos definidos por los vectores de soporte (los puntos de datos más cercanos al hiperplano), que son líneas que separan los datos en dos conjuntos: "positivo" o " negativo".

Para realizar esta tarea, SVM utiliza un kernel para proyectar los datos en un espacio de mayor dimensión, obteniendo así un clasificador eficiente. Los núcleos más comunes son la función de base lineal, gaussiana y radial (RBF). Dependiendo del kernel empleado, SVM se puede clasificar en SVM lineal y no lineal. El primero es cuando los datos solo tienen dos clases y se pueden dividir linealmente. Por otro lado, si los datos no se pueden separar linealmente, los datos empleados se transforman en un espacio de características donde pueden separarse linealmente[30].

5) Clasificador K-Vecinos más Cercanos (KNN)

El clasificador kNN es usado en tareas de clasificación donde los datos no poseen etiqueta, por lo cual, el método asigna la clase a cada observación teniendo como base la similaridad entre datos. Las cualidades de las observaciones se recopilan tanto para el entrenamiento como para el conjunto de datos de prueba con el fin de comparar en una gráfica de dos dimensiones las características de los datos. Para esta tarea podemos tener más de dos conjuntos de datos e incorporar más de dos variables de interés [31].

De forma predeterminada, la función kernel utiliza la distancia euclidiana que se puede calcular con la ecuación 3.16.

$$\begin{aligned}&D(p,q) \\ &= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}\end{aligned} \tag{3.16}$$

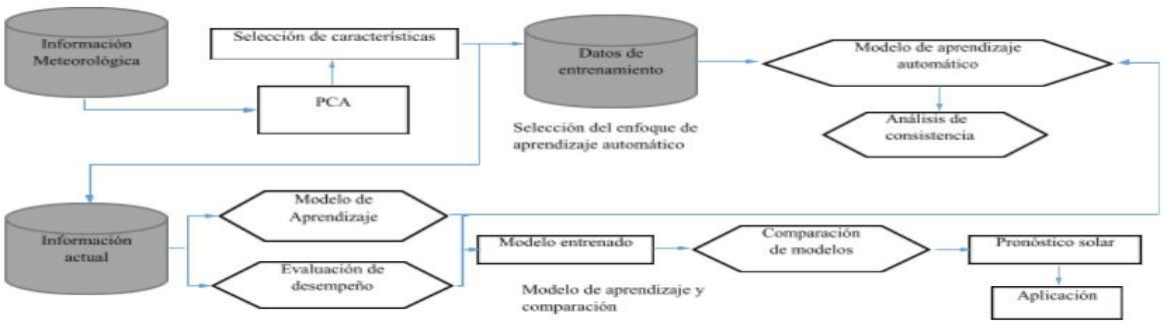

Fig 1. Modelo metodológico del marco general para la extracción de información y estimación de irradiación solar

Donde p y q son sujetos que comparar con n características. También existen otros métodos para calcular la distancia, como la distancia de Manhattan.

6) Métodos de conjuntos (Ensembled Methods)

Los métodos de conjunto son algoritmos de aprendizaje que construyen un conjunto de clasificadores y luego clasifican nuevos puntos de datos tomando un voto (ponderado) de sus predicciones. El método de conjunto original es el promedio bayesiano, pero los algoritmos más recientes incluyen la codificación de salida de corrección de errores, el ensacado y el impulso. Este documento revisa estos métodos y explica por qué los conjuntos a menudo pueden funcionar mejor que cualquier clasificador individual[32].

IV. METODOLOGÍA

La Metodología propuesta en este artículo se basa en los estudios previos presentados en la revisión del estado del arte (sección II), y en los modelos matemáticos (sección III).

La figura 1 presenta el modelo metodológico diseñado en el marco general para la extracción de información y estimación de irradiación solar; en la parte superior izquierda del modelo se tiene la entrada de los datos provenientes del observatorio meteorológico, en este caso usamos los datos provenientes de MERRA-2 (Modern-Era Retrospective analysis for Research and applications version 2)[33], estos datos son tratados mediante PCA para generar una nueva base de datos. En la selección de características se toman las 3 primeras componentes principales como se explica en la "*sección III, A, análisis por componente principales*", estos datos son usados en el entrenamiento del estimador mediante técnicas de aprendizaje supervisado, lo cual, es presentado en la parte inferior izquierda del modelo de la figura 1, los modelos generados son validados usando los datos de prueba como se muestra en la parte superior de la figura 1, este proceso se repite para cada una de las bases de datos provenientes de diferentes ubicaciones geográficas. Finalmente, se presentan los resultados de la comparación de los diferentes modelos usados en el estudio los cuales serán usados posteriormente para inferir correctamente la irradiación solar.

Para la estimación de la irradiación solar se hace uso de las técnicas de procesamiento de datos presentadas en la "*sección III, B, técnicas de procesamiento de datos*" y de validación cruzada para encontrar las tazas de fallo y acierto de cada uno de los algoritmos de aprendizaje supervisado utilizados.

A. Base de datos de irradiación solar

La base de datos de irradiación solar utilizada en el estudio fue tomada del observatorio de MERRA-2. La base de datos ofrece series temporales de temperatura (K), humedad relativa (%), presión (hPa), velocidad de viento (m/s) y dirección del viento (grados), lluvia (Kg/m^2), nieve (Kg/m^2), Profundidad de la nieve (m) e irradiación solar (W/m^2). Las bases de datos fueron generadas en una serie temporal con pasos de una hora entre muestra. Un resumen de los datos es presentado en la **Tabla I** en donde se indican cada una de las ubicaciones tomadas para el análisis del proyecto. La duración de las series temporales es de dos años, con mediciones cada hora, para el caso discreto se llevan estos valores de irradiación a valores diarios de irradiación.

TABLA I
MUESTRA DE DATOS AMBIENTALES USADOS

Latitud	Longitud	País [a]
10.3950	-75.3140	Colombia
48.2423	11.4003	Alemania
29.3861	108.549	China
-30.6000	24.4198	Suráfrica
-16.2252	143.706	Australia

B. Paneles solares

Los paneles solares fueron considerados en este artículo para la toma de decisiones del rango de irradiación en el entrenamiento de los sistemas de inferencia. Por lo cual, se revisó información por parte de diferentes fabricantes con el objetivo de hacer la selección de los rango de irradiación adecuados al momento de conformar los grupos para las salidas de los algoritmos.

Los grupos elegidos se muestran en la Tabla II, donde se indica la información por parte de los fabricantes así mismo la potencia generada por los dispositivos.

C. Validación Cruzada

La validación cruzada consiste en hacer K iteraciones en donde los datos de muestra se dividen en k subconjuntos. El sistema es entrenado con k-1 subconjuntos y el restante es utilizado para realizar las pruebas. Este proceso se realiza k veces, que es finalmente la cantidad de posibles combinaciones que se tienen de los datos. Este método es preciso y permite evaluar a partir de k combinaciones los datos de entrenamiento y de prueba, es un método robusto en comparación al método de partición, el cual en esencia es el mismo método pero realizando una sola combinación [24]. En el experimento desarrollado se utilizaron el 70% de los datos para el entrenamiento de los modelos computacionales, el 30% restante de los datos se usó para la validación de los resultados

V. Resultados Y Discusión de Resultados

Siguiendo el modelo presentado en la metodología, se analiza los metadatos generados mediante el PCA como se recomienda en [5]–[8], para esto, se usa las bases de datos presentadas en la sección IV, A, “Base de datos de irradiación solar” de MERRA-2 [33] en las ubicaciones presentadas en la **Tabla I.** El tratamiento de los datos mediante el uso de PCA genera una nueva base de datos que asocia el comportamiento de la irradiación y las tres componentes principales, con los rangos de potencia presentados en la **Tabla II,** un análisis similar se propone en [11], [17], finalmente la tabla III presenta los porcentajes de error del PCA con respecto a los datos originales al pasar de nueve variables independientes a tres, este error puede traducirse también como la perdida de información de los datos al transformar las variables de entrada.

TABLA III
Error Porcentual del PCA por país

País	Error % Caso Continuo	Error % Caso Discreto
Sudáfrica	5,4	7.8
USA	4,8	6.3
Alemania	3,6	5.9
Austria	4,5	7.1
Colombia	1,7	4.7
China	2,4	5.8

Los algoritmos fueron desarrollados e implementados en Matlab versión 2020b, los resultados de aplicar PCA a los datos de condiciones ambientales se observan en las figuras 2 a 5, en las cuales se analiza el comportamiento de los datos dado el rango de irradiación (cada color indica un rango de potencia generada dada la irradiación de entrada al panel solar), se intenta dar una interpretación de la dependencia que existe entre los rangos de radiación y la ubicación geográfica. Esto influye en los resultados finales del PCA.

Para el país de Colombia, se puede evidenciar en las figuras. 2 y 3 el comportamiento de los datos extraídos usando el PCA para dos y tres componentes principales, el color azul

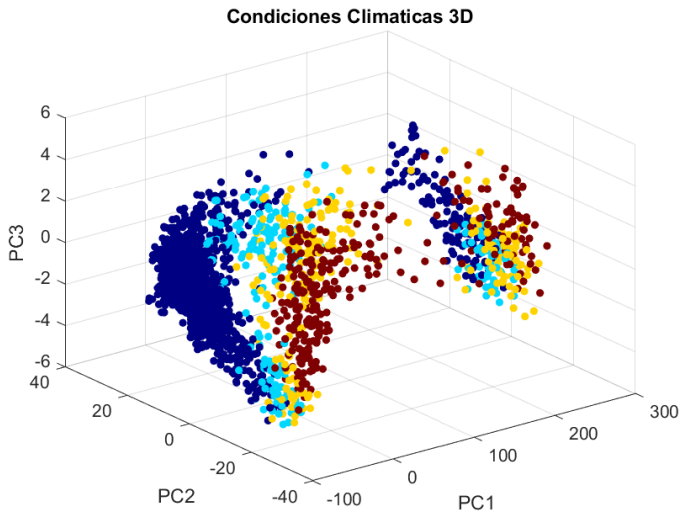

Fig. 2. Condiciones climáticas al aplicar PCA para las primeras 3 componentes principales tomando los datos en estado continuo para Colombia 2021 del 1/01 al 03/30.

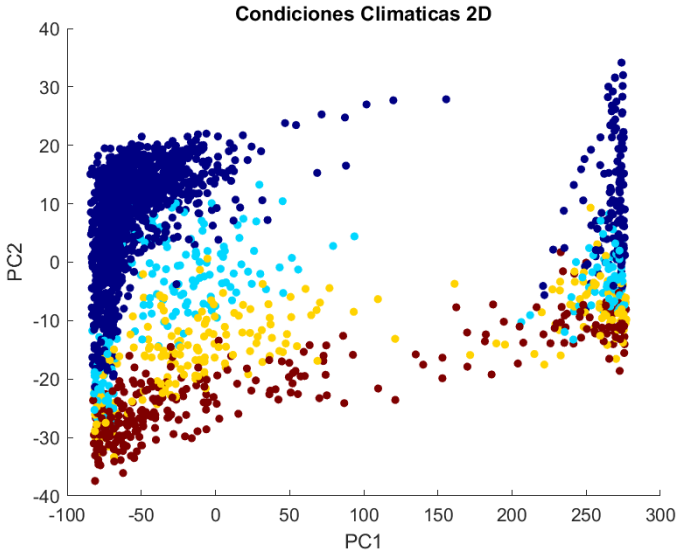

Fig. 3. Condiciones climáticas al aplicar PCA para las primeras 2 componentes principales tomando los datos en estado continuo para Colombia 2021 del 1/01 al 03/30

representa el rango de irradiación más bajo, mientras el color rojo representa el rango de irradiación más alto. Al observar los resultados arrojados por el PCA para esta base de datos se puede

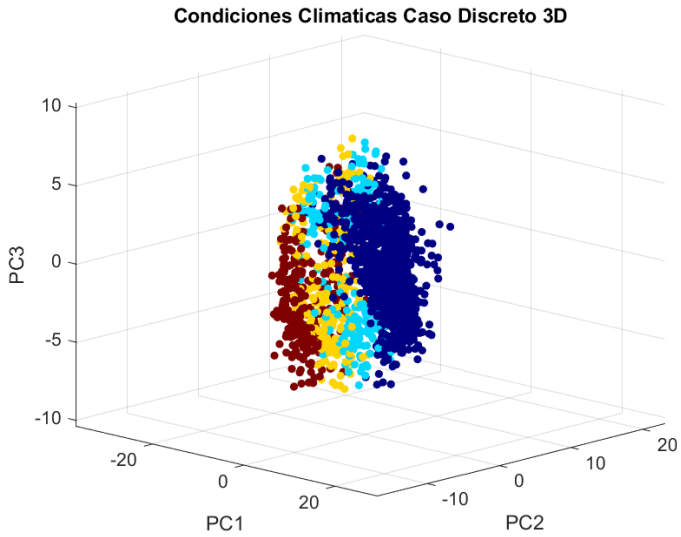

Fig. 4. Condiciones climáticas al aplicar PCA para las primeras 3 componentes principales tomando los datos en estado discreto para Colombia 2021 del 1/01 al 03/30

argumentar que los datos se comportan de forma cuasi-estacionaria durante el rango de tiempo estudiado, además de permitir la clasificación por rangos de radiación solar con un error medio cuadrático de 1.7%. El caso presentado en la base

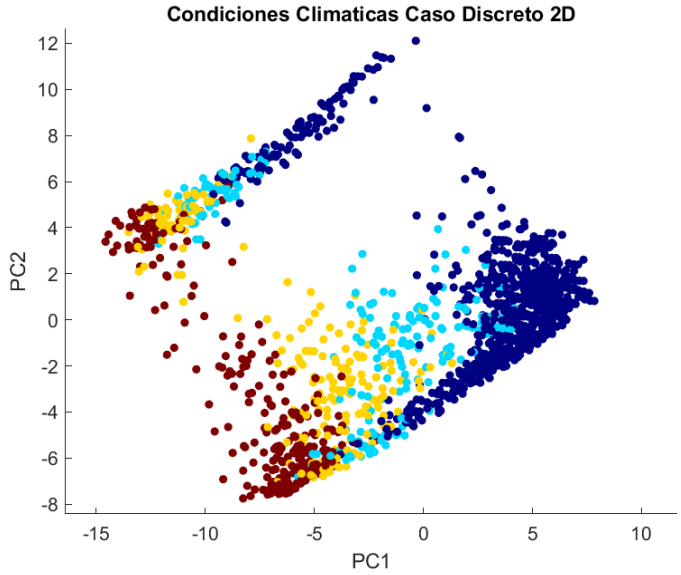

Fig. 5. Condiciones climáticas al aplicar PCA para las primeras 2 componentes principales tomando los datos en estado discreto para Colombia 2021 del 1/01 al 03/30

de datos Colombia muestra uno de los mejores resultados tanto si evaluamos de forma cualitativa como de forma cuantitativa los resultados del PCA, esto se observa en la **Tabla III**.

Por otra parte, en las figuras 4 y 5, se presentan los datos extraídos por el PCA sobre la base de datos discretos, que si bien, muestra un cambio en su comportamiento, sigue permitiendo la clasificación de los datos en grupos de irradiación con un error del 4.1% como se puede observar en la **Tabla III,** verificando los resultados en los estudios presentados en [13]–[15], [19]**.**

Para la segunda fase del estudio, se toman los valores resultantes de aplicar el PCA y la base de datos original como datos de entrada sobre los modelos de aprendizaje supervisado elegidos, siendo la propuesta de este estudio [13],[14]. De esta manera se obtienen los resultados que se presentan en la **Tabla IV**, en donde se relacionan los porcentajes de precisión o acierto que son una medida inversa al error de clasificación; los valores consignados en la **Tabla IV** son el resultado de la validación con los datos de prueba correspondientes al 30% del total de los datos como se mencionó anteriormente en la metodología, sin embargo, la validación con los datos de entrenamiento no se tuvo en cuenta para la construcción de la tabla que comparte similitudes con los resultados generados en [18], esto debido a que los resultados fueron mayormente positivos y

adicionalmente, debido a que el estudio está dirigido a establecer un marco referencial para nuevas investigaciones y aplicaciones, el caso de validación de entrenamiento no toma relevancia para el objeto de estudio como se explica en [17].

En la **Tabla IV** se presentan los valores de precisión en el acierto para cada uno de los modelos matemáticos mencionados en la sección II.

En las figuras 6 a 8 se presentan los resultados obtenidos de aplicar los modelos de entrenamiento supervisado sobre la base de datos generada mediante PCA y la base de datos original. Se puede observar los errores de acierto (marcas X) para cada caso, el mejor caso se presenta en los datos originales, lo cual lleva a concluir que la información contenida en dos o tres componentes principales no es suficiente para tener un sistema robusto, esto se observa en los datos en Colombia donde el error es del 11.1%, esto sucede cuando tenemos el modelo Naive Bayes entrenado con las tres componentes principales (caso discreto) en comparación con los datos originales en donde el error es solo del 8.3%, se concluye para esta etapa lo siguiente: El PCA permite una interpretación y la posibilidad de ver el comportamiento de los datos, pero la perdida de información es relevante en la tarea de entrenamiento de los clasificadores, por lo cual, al no tener una mejora representativa en la tarea de clasificación e inferencia, no se aconseja usar PCA como filtro para las bases de datos.

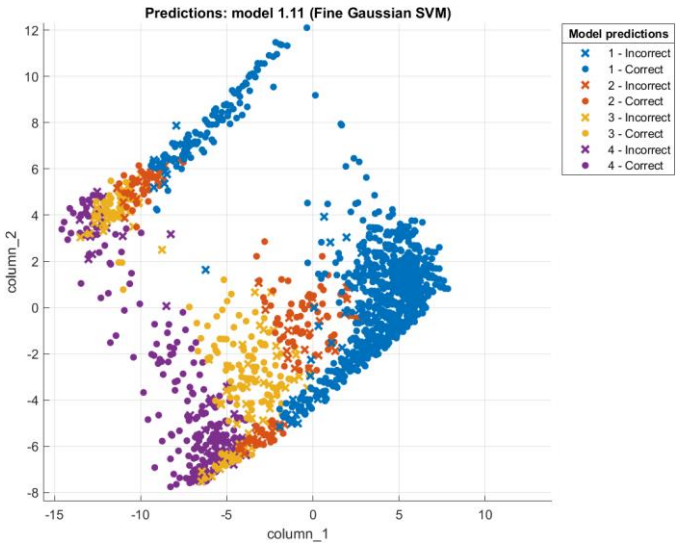

Fig. 7. Gráfico de inferencia usando SVM para el caso PCA discreto en Colombia.

Finalmente en la Tabla IV podemos diferenciar y evaluar algunos de los mejores y peores resultados del estudio, por ejemplo, en la figura 8, encontramos que la validación nos entrega un porcentaje de acierto de 99,8% al usar la base de datos original, seguido por un 87,1% resultado de la base de datos generada por el PCA presentado en la figura 6, y finalizando con el caso discreto en el cual se tuvo un resultado de 88,5% que podemos observar en la figura 7, estos resultados son similares a lo expuesto en [20], [21], en donde se observa el porcentaje de acierto en la estimación de variables.

TABLA IV
RESULTADOS DE VALIDACIÓN DE INFERENCIA DE ESTADOS

País	Modelo	PCA Discreto	PCA Continuo	Original
		Precisión %		
Sudáfrica	LDA	70	68,3	Error
	QDA	69,1	68,3	Error
	KNN	80	77,5	94,1
	Esemble	78,3	78,6	99,8
Colombia	LDA	82,5	81,7	Error
	QDA	84,3	84,4	Error
	Naive Bayes	81,1	79,8	91,7
	SVM	88,5	87,1	98,6
	KNN	87,9	86	93,6
	Esemble	87,6	86,4	99,8
China	LDA	78,8	77,2	96,9
	QDA	79,7	77,3	Error
	Naive Bayes	78,6	76,6	92
	SVM	80,3	79,9	98,2
	KNN	80,7	81,8	92,8
	Esemble	79,8	78,8	99,7

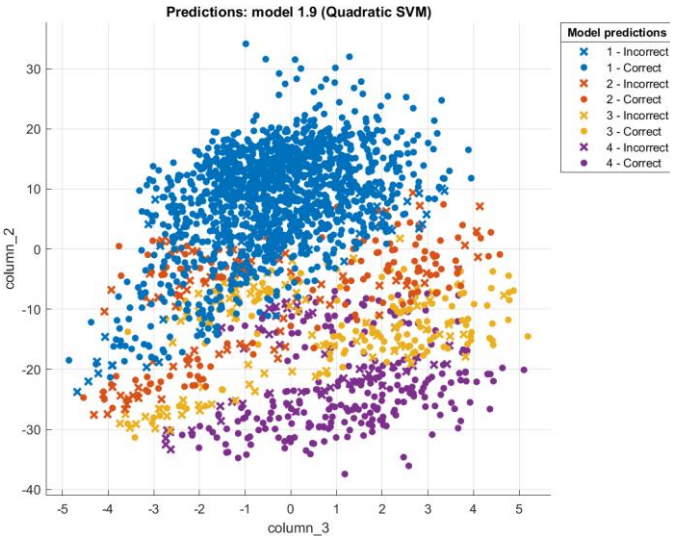

Fig. 8. Gráfico de predicción usando Quadractic SVM para el caso PCA continúo de los datos de Colombia.

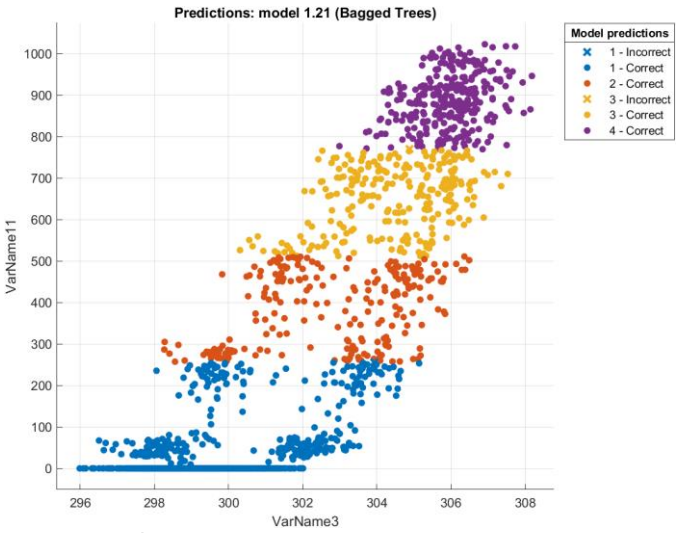

Fig. 6. Predicción usando Esembled Bagged Trees para el caso Original (Data) de los datos de Colombia.

VI. Conclusiones

Los PCA son una estrategia adecuada para el estudio de datos climáticos debido a que permiten visualizar de forma gráfica el comportamiento de los datos bajo la variable de interés irradiación solar e interpretar como se asocia la información oculta de los datos de condiciones climáticas.

La efectividad del PCA en el trabajo es evidente, ya que mediante la aplicación del PCA se pudo obtener los resultados presentados en las figuras y tablas que han servido de insumo para evaluar la posibilidad de generar un marco de referencia para el tratamiento de datos climáticos que sea independiente de una zona geográfica especifica.

La aplicación de PCA como primera fase de análisis y extracción de información muestra ser una estrategia adecuada para lograr la interpretación de los resultados, sin embargo, al momento de entrenar los sistemas de inferencia conformados por los clasificadores elegidos para el estudio, los resultados obtenidos tienen una tasa de error mayor que al usar la base de datos original como se expuso en los resultados presentados. Resulta ser particularmente útil aplicar PCA para definir contextos, encontrar errores en la adquisición de los datos climáticos, datos faltantes y peculiaridades sujetas a las diferentes zonas geográficas tales como: las estaciones del año o el tiempo solar, las cuales se presentaron como las principales dificultades en estudios previos como los presentados en [11], [17].

La aplicación de los diferentes modelos de clasificación han aportado una visión amplia sobre las posibilidad de crear un sistema robusto de inferencia de condiciones climáticas, esta hipótesis fue valorada a través del estudio y observada en detalle a través del comportamiento de los clasificadores para cada uno de los casos, en donde el caso Colombia muestra un claro ejemplo de la posibilidad de generar un sistema robusto, sin embargo, otros casos como el de Sudáfrica, en el cual, los resultados no son igual de favorables, nos marca el camino a seguir experimentando con nuevas técnicas con respecto a la extracción de información de bases climáticas para la inferencia de los rango de irradiación.

AGRADECIMIENTOS

Agradecimientos especiales a la Fundación Universitaria Tecnológico Comfenalco, en especial al grupo de investigación GISNET, por el apoyo y recursos brindados para el desarrollo de este trabajo de investigación.

Referencias

[1] M. Viviana and O. L. Castillo, “Colombian energy planning - Neither for energy, nor for Colombia,” *Energy Policy*, vol. 129, pp. 1132–1142, 2019, doi: https://doi.org/10.1016/j.enpol.2019.03.025.

[2] D. Silva Herran and T. Nakata, “Design of decentralized energy systems for rural electrification in developing countries considering regional disparity,” *Appl. Energy*, vol. 91, no. 1, pp. 130–145, 2012, doi: https://doi.org/10.1016/j.apenergy.2011.09.022.

[3] BP p.l.c., “Statistical Review of World Energy,” London, 2018. [Online]. Available: https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

[4] C. Washburn and M. Pablo-Romero, “Measures to promote renewable energies for electricity generation in Latin American countries,” *Energy Policy*, pp. 212–222, 2019, doi: 10.1016/j.enpol.2018.12.059.

[5] S. E. Hosseini and M. A. Wahid, “Hydrogen production from renewable and sustainable energy resources: Promising green energy carrier for clean development,” *Renew. Sustain. Energy Rev.*, vol. 57, pp. 850–866, 2016, doi: https://doi.org/10.1016/j.rser.2015.12.112.

[6] N. Abdelhafidi, N. E. I. Bachari, and Z. Abdelhafidi, “Estimation of solar radiation using stepwise multiple linear regression with principal component analysis in Algeria,” *Meteorol. Atmos. Phys.*, vol. 133, no. 2, pp. 205–216, 2021, doi: 10.1007/s00703-020-00739-0.

[7] K. Bouchouicha, N. Bailek, M. E.-S. Mahmoud, J. A. Alonso, A. Slimani, and A. Djaafari, “Estimation of Monthly Average Daily Global Solar Radiation Using Meteorological-Based Models in Adrar, Algeria,” *Appl. Sol. Energy*, vol. 54, no. 6, pp. 448–455, 2018, doi: 10.3103/S0003701X1806004X.

[8] X. Zhang and Z. Wei, “A Hybrid Model Based on Principal Component Analysis, Wavelet Transform, and Extreme Learning Machine Optimized by Bat Algorithm for Daily Solar Radiation Forecasting,” *Sustainability* , vol. 11, no. 15. 2019, doi: 10.3390/su11154138.

[9] J. Xue, T. L. Yip, B. Wu, C. Wu, and P. H. A. J. M. van Gelder, “A novel fuzzy Bayesian network-based MADM model for offshore wind turbine selection in busy waterways: An application to a case in China,” *Renew. Energy*, vol. 172, pp. 897–917, 2021, doi: https://doi.org/10.1016/j.renene.2021.03.084.

[10] K. Bouchouicha, M. A. Hassan, N. Bailek, and N. Aoun, “Estimating the global solar irradiation and optimizing the error estimates under Algerian desert climate,” *Renew. Energy*, vol. 139, pp. 844–858, 2019, doi: https://doi.org/10.1016/j.renene.2019.02.071.

[11] H. B. Tolabi, S. B. M. Ayob, M. H. Moradi, and M. Shakarmi, “New technique for estimating the monthly average daily global solar radiation using bees algorithm and empirical equations,” *Environ. Prog. Sustain. Energy*, vol. 33, no. 3, pp. 1042–1050, Oct. 2014, doi: https://doi.org/10.1002/ep.11858.

[12] U. Waqas, M. F. Ahmed, F. G. Awan, and Z. Hussain, “A Dimensionality Reduction-Based Approach to Select a Suitable Interpolator for the Mapping of Solar Irradiation Across Pakistan,” *MAPAN*, 2021, doi: 10.1007/s12647-021-00435-3.

[13] F. Li *et al.*, “Novel models to estimate hourly diffuse radiation fraction for global radiation based on weather type classification,” *Renew. Energy*, vol. 157, pp. 1222–1232, 2020, doi: https://doi.org/10.1016/j.renene.2020.05.080.

[14] Z. Song *et al.*, "General models for estimating daily and monthly mean daily diffuse solar radiation in China's subtropical monsoon climatic zone," *Renew. Energy*, vol. 145, pp. 318–332, 2020, doi: https://doi.org/10.1016/j.renene.2019.06.019.

[15] J. Li, Z. Wang, X. Cheng, J. Shuai, C. Shuai, and J. Liu, "Has solar PV achieved the national poverty alleviation goals? Empirical evidence from the performances of 52 villages in rural China," *Energy*, vol. 201, p. 117631, 2020, doi: https://doi.org/10.1016/j.energy.2020.117631.

[16] K. Ansari, S. K. Panda, and P. Jamjareegulgarn, "Singular spectrum analysis of GPS derived ionospheric TEC variations over Nepal during the low solar activity period," *Acta Astronaut.*, vol. 169, pp. 216–223, 2020, doi: https://doi.org/10.1016/j.actaastro.2020.01.014.

[17] U. Munawar and Z. Wang, "A Framework of Using Machine Learning Approaches for Short-Term Solar Power Forecasting," *J. Electr. Eng. Technol.*, vol. 15, no. 2, pp. 561–569, 2020, doi: 10.1007/s42835-020-00346-4.

[18] H. Bouzgou and C. A. Gueymard, "Fast short-term global solar irradiance forecasting with wrapper mutual information," *Renew. Energy*, vol. 133, pp. 1055–1065, 2019, doi: https://doi.org/10.1016/j.renene.2018.10.096.

[19] P. Chung *et al.*, "An intelligent control strategy for energy storage systems in solar power generation based on long-short-term power prediction," in *2020 8th International Electrical Engineering Congress (iEECON)*, 2020, pp. 1–4, doi: 10.1109/iEECON48109.2020.229485.

[20] A. Bakdi, W. Bounoua, S. Mekhilef, and L. M. Halabi, "Nonparametric Kullback-divergence-PCA for intelligent mismatch detection and power quality monitoring in grid-connected rooftop PV," *Energy*, vol. 189, p. 116366, 2019, doi: https://doi.org/10.1016/j.energy.2019.116366.

[21] S. Bandong, E. Leksono, A. Purwarianti, and E. Joelianto, "Performance Ratio Estimation and Prediction of Solar Power Plants Using Machine Learning to Improve Energy Reliability," in *2019 6th International Conference on Instrumentation, Control, and Automation (ICA)*, 2019, pp. 36–41, doi: 10.1109/ICA.2019.8916687.

[22] J. Zhang, Y. Chi, and L. Xiao, "Solar Power Generation Forecast Based on LSTM," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 2018, pp. 869–872, doi: 10.1109/ICSESS.2018.8663788.

[23] I. Jolliffe, "Principal Component Analysis," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096.

[24] C. M. Bishop, *Pattern recognition and machine learning*. New York : Springer, [2006] ©2006.

[25] C. Bishop, "Machine learning and the learning machine with Dr. Christopher Bishop," *Microsoft blog editor*, 2018. https://www.microsoft.com/en-us/research/blog/machine-learning-and-the-learning-machine-with-dr-christopher-bishop/.

[26] S. Balakrishnama and A. Ganapathiraju, "Institute For Signal And Information Processing Linear Discriminant Analysis-A Brief Tutorial."

[27] A. J. Izenman, "Linear Discriminant Analysis," in *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, New York, NY: Springer New York, 2008, pp. 237–280.

[28] Y. Qin, "A review of quadratic discriminant analysis for high-dimensional data," *WIREs Comput. Stat.* , vol. 10, no. 4, p. e1434, Jul. 2018, doi: https://doi.org/10.1002/wics.1434.

[29] H. Zhang and J. Su, "Naive Bayesian Classifiers for Ranking," in *Machine Learning: ECML 2004*, 2004, pp. 501–512.

[30] S. Suthaharan, "Support Vector Machine," in *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, Boston, MA: Springer US, 2016, pp. 207–235.

[31] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Ann. Transl. Med.*, vol. 4, no. 11, p. 218, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[32] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, 2000, pp. 1–15.

[33] "MERRA - www.soda-pro.com." http://www.soda-pro.com/web-services/meteo-data/merra (accessed Jun. 02, 2021).