Artículos

Enfoque multicriterio para la selección óptima de variables explicativas para modelos de pronóstico de la energía eléctrica de plantas solares fotovoltaicas

Multicriteria approach for the optimal selection of explanatory variables for forecast models of electrical energy from photovoltaic solar plants

Yajure-Ramírez, César A.

César A. Yajure-Ramírez
cyajure@gmail.com
Universidad Central de Venezuela, Venezuela

Revista Tecnológica ESPOL - RTE
Escuela Superior Politécnica del Litoral, Ecuador
ISSN: 0257-1749
ISSN-e: 1390-3659
Periodicidad: Semestral
vol. 35, núm. 3, 2023
rte@espol.edu.ec

Recepción: 09 Julio 2023
Aprobación: 19 Septiembre 2023

URL: http://portal.amelica.org/ameli/journal/844/8444930005/

DOI: https://doi.org/10.37815/rte.v35n3.1045

Resumen: Cuando se aborda un problema de pronóstico a través de modelos de regresión, se espera contar con el número óptimo de variables explicativas, y de no ser así, aplicar alguna técnica para reducir la dimensionalidad del problema. Actualmente, existe una variedad de métodos para seleccionar las características o variables explicativas, que a su vez caen dentro de distintas categorías, haciendo complejo sólo seleccionar el método idóneo para una aplicación específica. Entonces, el objetivo de esta investigación es presentar una metodología multicriterio para la selección óptima de las variables explicativas de un modelo de regresión, utilizando los métodos de selección de características como los criterios de decisión, y las variables explicativas como las alternativas. La metodología se ilustra a través del conjunto de datos de una planta solar fotovoltaica del Instituto Nacional de Estándar y Tecnología (NIST por sus siglas en inglés), de los Estados Unidos, tomando como variable objetivo a la energía eléctrica AC generada por la planta, y como variables explicativas a la irradiancia solar, la temperatura de los paneles solares, la temperatura ambiente, y la velocidad del viento. Se consideran métodos del tipo "filtro", del tipo "envoltura", y del tipo "incrustado". Utilizando la técnica multicriterio TOPSIS, se logró seleccionar la mejor variable para representar a la irradiancia solar con una ponderación de 1,00, a la temperatura de los paneles solares con 0,182, a la temperatura ambiente con 0,204, y a la velocidad del viento con 0,129.

Palabras clave: Correlación, componentes principales, LASSO, pesos, selección de características, TOPSIS.

Abstract: When a forecast problem is approached through regression models, it is expected to have the optimal number of explanatory variables and, if not, to be able to apply some technique to reduce the problem's dimensionality. Currently, there is a variety of methods to select the features or explanatory variables, which in turn fall into different categories, making it complex to select only the ideal method for a specific application. Therefore, this research aims to present a multicriteria methodology for the optimal selection of the explanatory variables of a regression model, using the feature selection methods as the decision criteria and the explanatory

variables as the alternatives. The methodology is illustrated through the data set of a photovoltaic solar plant from the National Institute of Standards and Technology (NIST) of the United States, taking the AC electricity generated by the plant as the objective variable and the temperature of the solar panels, the ambient temperature, and the wind speed as explanatory variables to solar irradiance. Methods of the "filter" type, the "wrapper" type, and the "embedded" type are considered. Using the TOPSIS multicriteria technique, it was possible to select the best variable to represent solar irradiance with a weighting of 1.00, the temperature of the solar panels of 0.182, the ambient temperature of 0.204, and the wind speed of 0.129.

Keywords: Correlation, principal components, LASSO, weights, feature selection, TOPSIS.

Introducción

Los métodos de selección de variables explicativas (*feature selection*) se han utilizado usualmente para la reducción de dimensionalidad, ya sea por una alta dimensionalidad del conjunto de datos y/o un bajo número de registros, o para cumplir con los requisitos específicos de cada algoritmo de aprendizaje automático, por ejemplo, el de no colinealidad de las variables explicativas. Asimismo, de acuerdo con Jović (2015), un alto número de características (variables explicativas) en contraste a un bajo número de registros podría llevar a un sobre ajuste del modelo que se esté creando, por lo que, se justifica obtener un subconjunto de características más pequeño, a partir del conjunto original. Por otra parte, en ciertas aplicaciones podrían aparecer particularidades que pudieran ser redundantes, y el analista debe decidir a priori cuáles de ellas utilizar y cuáles descartar.

Ahora bien, para llevar a cabo la selección de características, existe una variedad de métodos, que según Explorium (2023) los más populares son los del tipo "filtro" y los del tipo "envoltura" (*wrapper methods*). Los de tipo filtro utilizan métricas para desechar características irrelevantes, haciendo uso de pruebas estadísticas univariadas independientemente del modelo, por tanto, obtienen resultados más rápidos. Mientras que los métodos de envoltura miden la utilidad de la característica y hacen su selección con base a su nivel de importancia, al evaluar el subconjunto de características de acuerdo con los resultados de un predictor. Por lo anterior, pueden alcanzar un mejor desempeño en el predictor mencionado, pero toman un tiempo más largo en el proceso. Por otra parte, Li et al. (2021) agregan los métodos incrustados (*embedded methods*) en los que se integra el proceso de selección de variables explicativas con la etapa de entrenamiento del modelo.

En todo caso, cada método tiene sus ventajas y desventajas, por eso, sería de utilidad compararlos, por ejemplo, tipo filtro, envoltura, o incrustado, y seleccionar el mejor subconjunto de características de acuerdo con distintos criterios de selección. En ese sentido, el objetivo de esta investigación es presentar una metodología multicriterio para escoger las mejores variables explicativas a utilizar en la creación de modelos de regresión, empleando los métodos de selección de características como los criterios de decisión. La metodología se ilustra con un caso de estudio cuyos datos corresponden a las mediciones de las variables de una planta solar fotovoltaica del NIST, de los Estados Unidos, tomando como variable objetivo a la energía eléctrica AC generada por la planta.

Para alcanzar el objetivo de este estudio, se hizo una revisión de las investigaciones previas asociadas al tema tratado, luego del análisis, se pudo observar que ninguna de ellas tiene el enfoque de toma de decisiones multicriterio para la selección de variables explicativas. Esta revisión se presenta a continuación. En su investigación, Jomthanachai et al. (2022) plantean una metodología para la selección de variables explicativas

utilizando los métodos de filtro: análisis de correlación y análisis de componentes principales (PCA por sus siglas en inglés), además de los métodos incrustados: LASSO y Regresión de red elástica (*Elastic-net Regression*). Seguidamente, utilizan una serie de modelos de regresión derivados de algoritmos de aprendizaje automático para entrenar y validar el conjunto de datos. Los hallazgos indican que el conjunto de variables, obtenido con PCA y con regresión de red elástica, ofrecen los mejores resultados basado en el criterio de medición del error.

Por otra parte, Li et al. (2018) utilizan el método de mínima redundancia y máxima relevancia y el método de bosques aleatorios para la selección de variables explicativas en modelos para la predicción de la concentración de clorofila-a en la floración de algas. Combinan estos métodos con los algoritmos de aprendizaje automático máquinas de soporte vectorial y bosques aleatorios. El modelo, obtenido con la combinación bosques aleatorios/bosques aleatorios, alcanzó el mejor desempeño con el número más bajo de variables explicativas.

Así también, Otchere et al. (2022), usan ocho técnicas de selección de variables explicativas junto con el modelo regresor de gradiente reforzado (*Gradient Boosting Regressor*) para la caracterización de un yacimiento petrolífero marino. Los resultados indican que las técnicas bosque aleatorio, SelectKBest y Lasso obtuvieron el mejor desempeño para la permeabilidad, porosidad y predicciones de saturación de agua, respectivamente.

En su trabajo R et al. (2020), utilizan las técnicas de selección de variables explicativas F-test y umbral de varianza para obtener el conjunto de variables a usar para la predicción del cáncer de pecho. Para conseguir los modelos de clasificación utilizan los algoritmos de aprendizaje automático: Naive Bayes, máquina de vectores de soporte (SVM), árbol de decisiones, perceptrón multicapa (MLP), regresión logística y vecinos más cercanos (KNN), y tres técnicas de ensamblaje: embolsado, refuerzo y apilamiento. Además, ilustran la metodología a través de tres conjuntos de datos disponibles online, y concluyen que la técnica F-test junto con la técnica de ensamblaje de apilamiento es la que presenta mejor desempeño.

En la investigación de Frederick et al. (2019), se analiza el desempeño de cuatro procedimientos de selección de variables en la construcción de un modelo de regresión lineal múltiple. Estas técnicas son el método de búsqueda directa, el de selección hacia adelante, el método de eliminación hacia atrás y el de regresión por pasos. La variable objetivo es el producto interno bruto de Nigeria, y se tienen siete factores asociados al sistema económico del país como potenciales variables explicativas. Los modelos se evalúan utilizando el R., el VIF, y el error cuadrático medio. Luego del estudio, obtuvieron que el método de eliminación hacia atrás es el de mejor desempeño con una media de 1,67.

En otra investigación, los autores (Gebreyesus et al., 2023) introducen el método SHAP (Shapley Additive exPlanation) para identificar las variables explicativas relevantes para modelos de predicción de la demanda de energía de centros de datos, y comparan sus resultados con los métodos de selección tradicionales basados en el nivel de importancia. Los métodos fueron probados y validados utilizando un conjunto de datos reales de un Centro de datos HPC, un clúster CRESCO6 que consta de 20.832 núcleos. Los modelos fueron evaluados con las métricas MAE, RMSE, y MAPE. Al final, los resultados obtenidos demuestran que los modelos predictivos entrenados, utilizando las características seleccionadas con el método asistido por SHAP, funcionaron bien con un error menor y un tiempo de ejecución razonable en comparación con otros métodos.

Finalmente, Mesafint Belete & D.H (2020) utilizan los métodos de envolventes: selección hacia atrás de características, selección hacia delante de características, y selección recursiva de características, para escoger los atributos a utilizar en modelos de predicción del resultado de estado individual de la prueba del conjunto de datos de la Encuesta demográfica y de salud de Etiopía para el VIH/SIDA. Utilizan siete algoritmos de aprendizaje automático para clasificación, cuyos modelos se evalúan a través de las métricas exactitud, precisión, recall, y f1-score. Los resultados indican que los clasificadores bosque aleatorio, K-NN, y gradiente reforzado alcanzan los niveles de exactitud más altos después que se aplican los métodos envolventes.

Hasta este punto se ha realizado la revisión de las investigaciones que sustentan este trabajo, el resto del artículo se distribuye como se indica a continuación. En la sección dos se presenta la metodología utilizada, así como el conjunto de datos empleados para ilustrarla. Seguido, en la sección tres, se presentan y discuten los resultados obtenidos. Posteriormente, en la sección cuatro, se presentan las conclusiones que se derivan de la investigación. Finalmente, se esboza un listado con las referencias bibliográficas utilizadas.

Materiales y Métodos

La metodología involucra las etapas de un proyecto de ciencia de datos, tal como describen Cielen et al. (2016), y combinarlas con las etapas de un proceso de toma de decisiones multicriterio. En ese sentido, la primera etapa consistió en fijar el o los objetivos de la investigación, la segunda etapa implicó la obtención de los datos a analizar, y la tercera etapa residió en hacer el preprocesamiento de los datos. Seguidamente, se hizo la selección óptima de las características, para lo cual se eligieron los métodos a utilizar, pues estos definieron los criterios de decisión de la técnica multicriterio, cuyas alternativas serían las características o variables explicativas del conjunto de datos. En la Figura 1 se presenta de manera esquemática la metodología a aplicada.

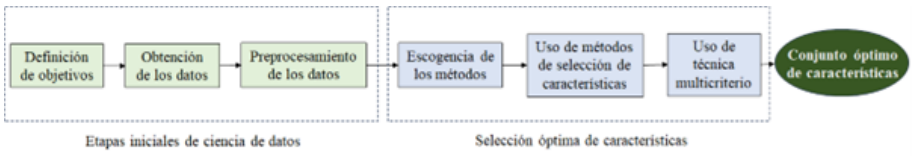

FIGURA 1
Metodología para selección óptima de características

Como se mencionó previamente, esta metodología se ilustra con el uso de los datos de las mediciones de las variables de una planta solar fotovoltaica. Entonces, el objetivo asociado a la primera etapa correspondió a desarrollar el pronóstico de la energía eléctrica AC generada por la planta, para lo que se desarrolló un modelo de regresión. En la segunda etapa, se obtuvieron los datos a utilizar, los cuales vinieron de fuentes internas o externas. Una vez que los datos estuvieron disponibles, lo siguiente consistió en hacer su preprocesamiento, lo que incluyó la detección e imputación de datos faltantes, la detección y corrección de datos duplicados, la transformación de variables, la combinación de variables para crear otras que son de interés, entre otras acciones que se encuentran en la investigación realizada por McKinney (2018).

Con los datos ya procesados, se procedió a escoger los métodos de selección de características, para lo que se debió considerar el tipo de datos (numéricos o categóricos), y el tipo de métodos (filtro, envoltura, o incrustado). Teniendo los métodos de selección de características (criterios de decisión), estos se aplicaron a las variables explicativas (alternativas a seleccionar y/o jerarquizar) para así obtener para cada una de ellas su ponderación por método. Con la ponderación de cada característica por método de selección, se procedió a aplicar la técnica de toma de decisiones multicriterio después de lo cual se tuvieron jerarquizados las características con una ponderación general.

Los métodos de selección de variables explicativas (criterios de decisión) tomados en cuenta fueron: análisis de correlación, y análisis de componentes principales, que se consideran tipo "filtro", eliminación recursiva de características, que se conoce del tipo "envoltura", y la técnica LASSO que se clasifica como del tipo "incrustado". Las variables explicativas consideradas (alternativas) fueron: irradiancia solar, temperatura de los paneles solares, temperatura ambiente, y velocidad del viento.

Toma de decisiones multicriterio

La toma de decisiones multicriterio (MCDM por sus siglas en inglés) está relacionada con abordar un problema de decisión en el que se tiene más de un criterio de decisión a considerar para la selección de la mejor opción, dentro de un conjunto de alternativas. De acuerdo con Eltarabishi et al. (2020), la MCDM se divide en toma de decisiones multiobjetivo (MODM por sus siglas en inglés), y toma de decisiones multiatributo (MADM por sus siglas en inglés). La MODM se caracteriza por tener un objetivo explícito y un espacio de decisión continuo (infinitas alternativas y atributos), mientras que la MADM se caracteriza por tener un objetivo implícito y un espacio de decisión discreto, con alternativas y atributos discretos.

Por otra parte, según Triantaphyllou et al. (1998), un problema de decisión multi atributo se podría representar a través de su matriz de decisión. Esta es una matriz (M×N) en la que el elemento a_{ij} indica el desempeño de la alternativa A. cuando es evaluada en términos del criterio de decisión C., (para i = 1, 2, 3, ..., M, y j = 1, 2, 3, ..., N). Cada uno de los criterios tiene un peso de importancia relativa w., los cuales, por lo general, son definidos por el "tomador de decisión". Entonces, dado un conjunto de alternativas y de criterios de decisión, se busca determinar la alternativa óptima con el grado más alto de "deseabilidad" con respecto a los criterios de decisión.

Entre los tipos de problemas de decisión se tienen problemas de selección, clasificación, jerarquización, y descripción. A esta investigación le interesa el problema de jerarquización, en el que se ordenan las alternativas desde la mejor hasta la peor, de acuerdo con un puntaje o por medio de comparaciones pareadas (Ishizaka & Nemery, 2013).

Para tratar con los problemas de decisión, se puede utilizar alguna de las distintas técnicas multicriterio disponibles para ello. En particular, para los problemas de jerarquización, los métodos disponibles son variados, pero en este estudio se utiliza la técnica para ordenamiento de las preferencias por similitud con las soluciones ideales (TOPSIS por sus siglas en inglés), para la jerarquización de las variables explicativas de los modelos de regresión.

De acuerdo con Velasquez & Hester (2013), la técnica TOPSIS posee un procedimiento sencillo, es fácil de usar y programar, y el número de pasos sigue siendo el mismo independientemente del número de criterios. Esta técnica se basa en seleccionar la mejor alternativa midiendo la distancia geométrica más corta a la solución positiva ideal, y la distancia geométrica más larga a la solución negativa ideal (Sahoo et al., 2022). Consta de una serie de pasos, el primero, común a todas las técnicas multicriterio de toma de decisiones, consiste en obtener la matriz de decisión. Posteriormente, se obtiene la matriz de decisión normalizada, luego la matriz de decisión normalizada ponderada, la solución ideal positiva y la solución ideal negativa, la distancia de cada alternativa con esas soluciones ideales y, finalmente, la cercanía relativa de cada alternativa con la solución ideal (Papathanasiou & Ploskas, 2018).

Obtención de los datos

Los datos se obtuvieron de la página web del NIST (National Institute of Standards and Technology, 2023). La información incluyó valores de variables climáticas y eléctricas, correspondientes a las mediciones minutales realizadas entre los años 2015 y 2018, provenientes de las estaciones de medición de una planta solar fotovoltaica, localizada en Maryland, Estados Unidos. Está compuesta de 1.152 paneles solares de silicio monocristalino marca Sharp, con 235 Wp por panel (Datasheet, 2023). Asimismo, cuenta con un único inversor de 260 kW nominales de potencia AC marca PVPowered (SolarDesignTool, 2023).

Las variables climáticas incluyeron mediciones de: irradiancia solar en vatios por metro cuadrado ("SEWSPOAIrrad_Wm2_Avg"), temperatura ambiente en grados Celsius ("SEWSAmbientTemp_C_Avg"), temperatura promedio en los paneles solares en grados

Celsius ("SEWSModuleTemp_C_Avg"), velocidad promedio del viento en metros por segundo ("WindSpeedAve_ms"), entre otras. En cuanto a la temperatura de los paneles solares, la variable que se mencionó previamente integra esa medición en una sola variable, pero en el conjunto de datos hay nueve variables adicionales de medición de temperatura de los paneles solares, cuyos sensores (RTD) están ubicados en distintos puntos de la planta. Asimismo, para la temperatura ambiente, la irradiancia solar, y la velocidad del viento hay una variable adicional, "AmbTemp_C_Avg", "RefCell1_Wm2_Avg", "WindSpeed_ms_Max", respectivamente.

En cuanto a las variables eléctricas se tuvieron mediciones de potencia activa AC en kilovatios ("'PwrMtrP_kW_Avg"), potencia reactiva en kilovoltio amperios reactivos ("PwrMtrP_kVAR_Avg"), potencia aparente en kilovoltio amperios aparentes ("PwrMtrP_kVA_Avg"), frecuencia eléctrica en Hertz ("PwrMtrFreq_Avg"), factor de potencia ("PwrMtrPF_Avg"), entre otras. El conjunto de datos constó de 2.103.810 registros (filas) correspondientes a las mediciones minutales de un total de 99 variables, separados en archivos con datos diarios, es decir, 1.461 archivos propios de cada uno de los días entre los años 2015 y el 2018.

Preprocesamiento de los datos

El primer paso consistió en combinar los 1.461 archivos de datos diarios, para crear cuatro archivos de datos, de acuerdo con cada uno de los cuatro años del período de estudio, para luego unirlos y crear un solo archivo, y así alcanzar la totalidad de registros mencionados previamente. Luego, se hizo un análisis de datos faltantes, resultando que de las noventa y nueve columnas (variables), sólo la fecha no presentó datos faltantes, siendo 9.587 el valor mínimo y 128.369 el valor máximo de datos faltantes por columna. Las variables con valores máximos de datos faltantes estaban asociadas al funcionamiento específico del inversor, las cuales no son de utilidad en esta investigación, por lo que fueron eliminadas. A continuación, de las variables restantes se eliminaron los registros con al menos un dato faltante, para quedar 1.997.418 registros sin datos faltantes. No se detectaron datos duplicados.

Posteriormente, utilizando la columna de potencia eléctrica AC, se creó la columna de la energía eléctrica AC ("energyAC_kWh"). Finalmente, los datos minutales se agruparon para obtener un set de datos con resolución horaria, el *quese* utilizó para ilustrar la metodología planteada.

Resultados y Discusión

En esta sección se presentan y discuten los resultados obtenidos luego de aplicar los métodos de selección de características, y la técnica de toma de decisión multicriterio. La variable objetivo considerada es la energía eléctrica AC generada por la planta, mientras que las variables explicativas (alternativas) son irradiancia solar, temperatura de los paneles solares, temperatura ambiente, y velocidad del viento.

Métodos de selección de características

En esta sección se presentan y aplican los métodos de selección de las variables explicativas, dos del tipo "filtro", dos del tipo "envoltura", y un método del tipo "incrustado".

Análisis de correlación

Como primer método se tuvo el análisis de correlación, para lo que se utilizó el método de Pearson, pero también los métodos de Spearman y Kendall, ya que, como bien lo mencionan Navlani et al. (2021), el primero es un método paramétrico, mientras que los otros dos métodos no imponen ninguna suposición con respecto a la distribución de los datos. Es importante recordar que el coeficiente de correlación varía entre "-1" y "+1", siendo negativo cuando la relación entre las variables respectivas es inversa, y positivo cuando esta relación es directa. Para interpretar los valores se toma en cuenta lo planteado por Ratner (2017, p. 26), quien postula que "valores entre 0 y 0,3 (0 y -0,3) indican una relación positiva (negativa) débil. Los valores entre 0,3 y 0,7 (−0,3 y −0,7) señalan una relación positiva (negativa) moderada. Los valores entre 0,7 y 1,0 (−0,7 y −1,0) evidencian una fuerte relación positiva (negativa)". En ese sentido, en la Tabla 1 se presenta el análisis de correlación de todas las variables explicativas con respecto a la variable objetivo.

TABLA 1
Coeficientes de correlación de variables explicativas con la variable objetivo

Variable	Pearson	Variable	Spearman	Variable	Kendall
RefCell1_Wm2_Avg	0,940	RefCell1_Wm2_Avg	0,925	RefCell1_Wm2_Avg	0,860
SEWSPOAIrrad_Wm2_Avg	0,922	SEWSPOAIrrad_Wm2_Avg	0,911	SEWSPOAIrrad_Wm2_Avg	0,829
RTD_C_Avg_5	0,673	RTD_C_Avg_5	0,675	RTD_C_Avg_5	0,501
RTD_C_Avg_9	0,667	RTD_C_Avg_3	0,670	RTD_C_Avg_3	0,497
RTD_C_Avg_4	0,662	RTD_C_Avg_9	0,669	RTD_C_Avg_9	0,496
RTD_C_Avg_3	0,655	RTD_C_Avg_4	0,665	RTD_C_Avg_4	0,493
RTD_C_Avg_7	0,655	RTD_C_Avg_7	0,659	RTD_C_Avg_7	0,487
RTD_C_Avg_6	0,616	RTD_C_Avg_1	0,631	RTD_C_Avg_1	0,465
SEWSModuleTemp_C_Avg	0,615	RTD_C_Avg_6	0,629	RTD_C_Avg_8	0,464
RTD_C_Avg_8	0,555	RTD_C_Avg_8	0,628	RTD_C_Avg_6	0,463
RTD_C_Avg_2	0,489	RTD_C_Avg_2	0,624	RTD_C_Avg_2	0,460
SEWSAmbientTemp_C_Avg	0,325	SEWSModuleTemp_C_Avg	0,620	SEWSModuleTemp_C_Avg	0,455
AmbTemp_C_Avg	0,309	SEWSAmbientTemp_C_Avg	0,369	SEWSAmbientTemp_C_Avg	0,257
WindSpeedAve_ms	0,307	AmbTemp_C_Avg	0,362	AmbTemp_C_Avg	0,252
WindSpeed_ms_Max	0,278	WindSpeed_ms_Max	0,358	WindSpeed_ms_Max	0,251
RTD_C_Avg_1	0,160	WindSpeedAve_ms	0,356	WindSpeedAve_ms	0,249

En la Tabla 1 también se puede observar que, para los métodos Spearman y Kendall, la jerarquización de las variables explicativas fue similar, sólo hay una diferencia en las posiciones de los sensores RTD de temperatura 6 y 8. En cuanto a los resultados con el método de Pearson, se nota que la jerarquización fue diferente a la de los otros dos métodos. En lo que sí coincidieron los tres métodos fue en los dos primeros lugares ocupados por las variables de la irradiancia solar, con una fuerte correlación directa con la energía eléctrica AC. Asimismo, con los tres métodos se obtuvo que la velocidad del viento, tanto la máxima como la promedio, ocupan los últimos lugares en la jerarquización, con una correlación débil y directa con la variable objetivo. Por último, se pudo decir que todas las variables explicativas tienen una relación directa con la variable objetivo. Es importante destacar que Lee (2019) utiliza esta técnica para seleccionar las mejores variables explicativas en un modelo de regresión lineal múltiple.

Análisis de componentes principales PCA

Esta técnica entra en el grupo de los algoritmos de aprendizaje automático no supervisados que, de acuerdo con Raschka & Mirjalili (2017, p. 142), "es una técnica de transformación lineal no supervisada que se usa

ampliamente en diferentes campos, más prominentemente para la extracción de características y reducción de la dimensionalidad". Por otra parte, Navlani et al. (2021, p. 319) plantean que

> El concepto principal de PCA es el descubrimiento de relaciones y correlaciones invisibles entre atributos en el conjunto de datos original. Los atributos altamente correlacionados son tan similares como para considerarse redundantes. Por lo tanto, PCA elimina tales atributos redundantes.

Cuando se utiliza para reducir la dimensionalidad de los datos, sin afectar la información significativa del conjunto de datos, PCA encuentra las direcciones de máxima varianza y proyecta los datos en un nuevo subespacio con igual o menos dimensiones que la original, que son precisamente las componentes principales. La ecuación de PCA para un conjunto de datos de . dimensiones, es la que se presenta en la ecuación 1. Se deduce que cada componente principal es una combinación lineal de las variables originales.

$$PC_j = w_{1j} \cdot X_1 + w_{2j} \cdot X_2 + \cdots + w_{pj} \cdot X_p \quad (1)$$

Siendo:

j-ésima componente principal

Las variables explicativas del conjunto de datos original

Los pesos de las variables explicativas en la j-ésima componente principal

Los pesos pueden interpretarse como el nivel de importancia que tiene cada variable en cada componente principal, y la suma de los cuadrados de los pesos, para una componente principal particular es igual a la unidad. Lo mismo aplica para la suma de los cuadrados de los pesos para una variable particular a lo largo de todas las componentes, y es así como según lo planteado por Ratner (2017), la suma de los cuadrados de los pesos a lo largo de todas las componentes principales para una variable explicativa particular, indica cuanta varianza aporta esta variable. Se cumple también que la primera componente principal es la que posee más información significativa, la segunda componente es la segunda con más información, y así sucesivamente.

En ese sentido, se aplicó el algoritmo PCA a los datos, obteniéndose que ya en las primeras seis componentes principales se explica el 100% de la varianza de los datos, tal como se observa en la Figura 2. Se puede ver que el modelo arrojó tantas componentes principales como variables en el conjunto de datos original. La primera componente explicó el 76% de la varianza de los datos, la segunda componente el 20%, y las siguientes cuatro, el 1% cada una.

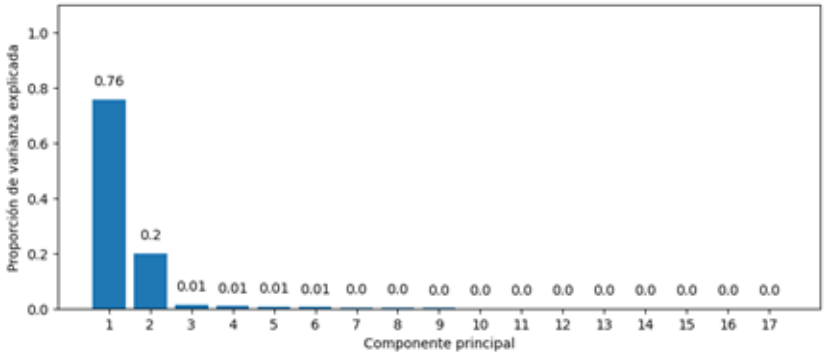

FIGURA 2
Proporción de varianza explicada por componente principal

Entonces, se elevaron al cuadrado los pesos de las variables explicativas, se multiplicaron por la proporción de varianza de la componente principal respectiva, y se sumaron esos valores para cada variable explicativa, pero sólo hasta la sexta componente principal. Los resultados se presentan en la Tabla 2, en la que las variables ya están jerarquizadas desde la de mayor peso hasta la de menor peso.

TABLA 2
Pesos de importancia según PCA

Variables	Pesos
RefCell1_Wm2_Avg	0,1538
SEWSAmbientTemp_C_Avg	0,1235
SEWSPOAIrrad_Wm2_Avg	0,1126
SEWSModuleTemp_C_Avg	0,1028
AmbTemp_C_Avg	0,0720
RTD_C_Avg_9	0,0689
RTD_C_Avg_4	0,0673
RTD_C_Avg_7	0,0663
RTD_C_Avg_5	0,0286
RTD_C_Avg_2	0,0181
RTD_C_Avg_3	0,0146
WindSpeedAve_ms	0,0091
RTD_C_Avg_8	0,0072
RTD_C_Avg_6	0,0063
WindSpeed_ms_Max	0,0019
RTD_C_Avg_1	0,0000

De la Tabla 2 se puede ver que las variables asociadas a la irradiancia solar, junto con la de temperatura ambiente, y la de la temperatura de los paneles solares, fueron las de mayor peso dentro del conjunto de datos.

Técnicas de eliminación recursiva de características

La técnica RFE (*Recursive Feature Elimination*) es del tipo envoltura e itera sobre el conjunto de datos hasta encontrar el subconjunto de variables explicativas que tienen el mejor desempeño de acuerdo con un modelo de regresión determinado. En cada iteración se va eliminando una o varias características simultáneamente. Es decir, generan muchos modelos con diferentes subconjuntos de variables explicativas, y se seleccionan aquellas que generan el modelo con el mejor desempeño de acuerdo con alguna métrica determinada. Usualmente, se utilizan modelos de regresión de árboles de decisión, los que generan niveles de importancia para las variables explicativas.

El lenguaje de programación Python tiene varias librerías que incluyen métodos para eliminación recursiva de características. Una de estas librerías es la de selección de características (*feature_selection*), la cual tiene la técnica denominada RFE. Esta fue la primera utilizada (RFE1), considerando un modelo de regresión de bosque aleatorio, para así utilizar los niveles de importancia de las variables explicativas. Los resultados obtenidos se muestran en la Tabla 3, en la cual se puede notar que las variables asociadas a la irradiancia solar fueron las que alcanzan el mayor valor de los pesos, mientras que las variables de los sensores RTD 2, 4, 6, y 8, fueron las de menor peso de importancia, según esta técnica. La variable de la velocidad del viento promedio tuvo mayor peso que la de velocidad máxima.

TABLA 3
Pesos de importancia RFE1

Variables	Pesos
RefCell1_Wm2_Avg	0,8970
SEWSPOAIrrad_Wm2_Avg	0,0180
RTD_C_Avg_7	0,0153
AmbTemp_C_Avg	0,0150
RTD_C_Avg_5	0,0081
RTD_C_Avg_9	0,0064
RTD_C_Avg_1	0,0054
SEWSAmbientTemp_C_Avg	0,0050
WindSpeedAve_ms	0,0046
RTD_C_Avg_3	0,0042
SEWSModuleTemp_C_Avg	0,0040
WindSpeed_ms_Max	0,0038
RTD_C_Avg_6	0,0036
RTD_C_Avg_8	0,0035
RTD_C_Avg_4	0,0034
RTD_C_Avg_2	0,0025

Otra librería de Python que contiene técnicas para la selección de variables explicativas es máquina de características (*feature-engine*), la que incluye a la técnica *RecursiveFeatureElimination* (RFE2). Esta se utiliza considerando el modelo de regresión *Gradient Boosting Regressor*, y los resultados se muestran en la Tabla 4, en la cual se puede ver que para este criterio las variables con mayor peso fueron las de la irradiancia solar, y las de menor peso correspondieron a los sensores RTD 2, 3, 4, 6, 8. La variable de la velocidad promedio del viento tuvo mayor peso que la de velocidad máxima.

TABLA 4
Pesos de importancia RFE2

Variables	Pesos
RefCell1_Wm2_Avg	0,6421
SEWSPOAIrrad_Wm2_Avg	0,3404
RTD_C_Avg_7	0,0110
AmbTemp_C_Avg	0,0020
WindSpeedAve_ms	0,0014
SEWSModuleTemp_C_Avg	0,0005
RTD_C_Avg_9	0,0005
RTD_C_Avg_5	0,0005
SEWSAmbientTemp_C_Avg	0,0004
RTD_C_Avg_1	0,0003
WindSpeed_ms_Max	0,0003
RTD_C_Avg_6	0,0003
RTD_C_Avg_8	0,0002
RTD_C_Avg_2	0,0000
RTD_C_Avg_3	0,0000
RTD_C_Avg_4	0,0000

Técnica de regresión LASSO

Es una técnica de regularización que se utiliza para prevenir sobreajuste en los modelos. Según Hackeling (2014, p. 40) "la regularización agrega información al problema, con frecuencia en la forma de una penalidad

a la complejidad, o al problema". LASSO (*Least Absolute Shrinkage and Selection Operator*) penaliza los coeficientes del modelo de regresión lineal múltiple agregando la norma L_1, lo que hará que muchos de los coeficientes se hagan igual a cero, y sí hay variables explicativas correlacionadas entre sí, sus coeficientes son los que se anulan, dejando sólo una de ellas con coeficiente no nulo. La función de costo de esta técnica se presenta en la ecuación 2, en la que el segundo sumando representa a la norma L_1.

$$\frac{1}{2 \cdot N_{entr}} \cdot \sum_{i=1}^{N_{entr}} \left(y_{real}^{(i)} - y_{pred}^{(i)} \right)^2 + \alpha \cdot \sum_{j=1}^{n} |a_j| \qquad (2)$$

Siendo:

N_{entr}: total de registros del set de entrenamiento del modelo

$y^{(i)}_{real}$: i-ésimo valor real de la variable objetivo

$y^{(i)}_{pred}$: i-ésimo valor predicho de la variable objetivo

a_j: coeficiente de la j-ésima variable explicativa

α: Es un hiperparámetro que define la intensidad de la penalización

Esta técnica se aplicó al conjunto de datos, tomando en cuenta un valor óptimo para el hiperparámetro igual a 0,1. Los resultados se presentan en la Tabla 5, aquí se puede observar que sólo los coeficientes de las variables explicativas asociadas a la irradiancia solar fueron diferentes de cero, resaltando una de ellas sobre la otra.

TABLA 5
Pesos de importancia LASSO

Variables	Pesos
RefCell1_Wm2_Avg	0,9883
SEWSPOAIrrad_Wm2_Avg	0,0513
SEWSAmbientTemp_C_Avg	0,0000
SEWSModuleTemp_C_Avg	0,0000
AmbTemp_C_Avg	0,0000
RTD_C_Avg_1	0,0000
RTD_C_Avg_2	0,0000
RTD_C_Avg_3	0,0000
RTD_C_Avg_4	0,0000
RTD_C_Avg_5	0,0000
RTD_C_Avg_6	0,0000
RTD_C_Avg_7	0,0000
RTD_C_Avg_8	0,0000
RTD_C_Avg_9	0,0000
WindSpeed_ms_Max	0,0000
WindSpeedAve_ms	0,0000

Selección multicriterio de las variables explicativas

Como se indicó previamente, se considera un problema de jerarquización en el que las variables explicativas son las alternativas y los métodos de selección de características corresponden a los criterios de decisión. Por otra parte, la técnica multicriterio a utilizar fue TOPSIS, y se consideró que los criterios de decisión tuvieron el mismo peso de importancia relativa. El primer paso, consistió en encontrar la matriz de decisión, la que se presenta en la Tabla 6. Para el caso del análisis de correlación se utilizó el resultado del método de Spearman, puesto que fue del tipo no paramétrico.

TABLA 6
Matriz de decisión

Alternativas	Correlación	PCA	RFE1	RFE2	LASSO
RefCell1_Wm2_Avg	0,9250	0,1538	0,8970	0,6421	0,9883
SEWSPOAIrrad_Wm2_Avg	0,9109	0,1126	0,0180	0,3404	0,0513
RTD_C_Avg_5	0,6752	0,0286	0,0081	0,0005	0,0000
RTD_C_Avg_3	0,6697	0,0146	0,0042	0,0000	0,0000
RTD_C_Avg_9	0,6690	0,0689	0,0064	0,0005	0,0000
RTD_C_Avg_4	0,6653	0,0673	0,0034	0,0000	0,0000
RTD_C_Avg_7	0,6590	0,0663	0,0153	0,0110	0,0000
RTD_C_Avg_1	0,6311	0,0000	0,0054	0,0003	0,0000
RTD_C_Avg_6	0,6292	0,0063	0,0036	0,0003	0,0000
RTD_C_Avg_8	0,6279	0,0072	0,0035	0,0002	0,0000
RTD_C_Avg_2	0,6240	0,0181	0,0025	0,0000	0,0000
SEWSModuleTemp_C_Avg	0,6199	0,1028	0,0040	0,0005	0,0000
SEWSAmbientTemp_C_Avg	0,3687	0,1235	0,0050	0,0004	0,0000
AmbTemp_C_Avg	0,3623	0,0720	0,0150	0,0020	0,0000
WindSpeed_ms_Max	0,3583	0,0019	0,0038	0,0003	0,0000
WindSpeedAve_ms	0,3556	0,0091	0,0046	0,0014	0,0000

El siguiente paso consiste en obtener la matriz de decisión normalizada, utilizando la ecuación 3. Los resultados se presentan en la Tabla 7.

$$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} a_{ij}^2}} \tag{3}$$

TABLA 7
Matriz de decisión normalizada

Alternativas	Correlación	PCA	RFE1	RFE2	LASSO
RefCell1_Wm2_Avg	0,366	0,535	0,999	0,883	0,999
SEWSPOAIrrad_Wm2_Avg	0,360	0,392	0,020	0,468	0,052
RTD_C_Avg_5	0,267	0,100	0,009	0,001	0,000
RTD_C_Avg_3	0,265	0,051	0,005	0,000	0,000
RTD_C_Avg_9	0,264	0,240	0,007	0,001	0,000
RTD_C_Avg_4	0,263	0,234	0,004	0,000	0,000
RTD_C_Avg_7	0,261	0,231	0,017	0,015	0,000
RTD_C_Avg_1	0,250	0,000	0,006	0,000	0,000
RTD_C_Avg_6	0,249	0,022	0,004	0,000	0,000
RTD_C_Avg_8	0,248	0,025	0,004	0,000	0,000
RTD_C_Avg_2	0,247	0,063	0,003	0,000	0,000
SEWSModuleTemp_C_Avg	0,245	0,358	0,004	0,001	0,000
SEWSAmbientTemp_C_Avg	0,146	0,430	0,006	0,001	0,000
AmbTemp_C_Avg	0,143	0,251	0,017	0,003	0,000
WindSpeed_ms_Max	0,142	0,007	0,004	0,000	0,000
WindSpeedAve_ms	0,141	0,032	0,005	0,002	0,000

Posteriormente, se obtuvo la matriz de decisión normalizada ponderada, al tomar en cuenta los pesos de importancia relativa de los criterios de decisión (métodos de selección de características). Ahora bien, para esta investigación se consideró que todos los métodos son igual de importantes, por lo que la matriz buscada fue igual a la presentada en la Tabla 7.

Seguidamente, y a partir de la Tabla 7, se alcanzó la solución ideal positiva, tomando los valores máximos para cada uno de los criterios de decisión. De igual forma, la solución ideal negativa se consiguió al tomar los valores mínimos. Los resultados se presentan en las ecuaciones 4 y 5.

$$A^* = \{0{,}366, 0{,}535, 0{,}999, 0{,}883, 0{,}999\} \quad (4)$$

$$A^- = \{0{,}141, 0{,}000, 0{,}003, 0{,}000, 0{,}000\} \quad (5)$$

A continuación, se calculó la distancia de cada alternativa con la solución ideal positiva D.*, y la distancia con la solución ideal negativa D.-. A partir de estos dos valores, se calculó la cercanía relativa de cada alternativa con la solución ideal C.*, utilizando la ecuación 6. Los resultados obtenidos se presentan en la Tabla 8, con las variables explicativas (alternativas) ya jerarquizadas de acuerdo con C.*.

$$C_i^* = \frac{D_i^-}{D_i^- + D_i^*} \quad (6)$$

TABLA 8
Variables explicativas jerarquizadas

Alternativas	Di*	Di-	Ci*
RefCell1_Wm2_Avg	0,000	1,763	1,000
SEWSPOAIrrad_Wm2_Avg	1,431	0,651	0,313
SEWSAmbientTemp_C_Avg	1,680	0,430	0,204
SEWSModuleTemp_C_Avg	1,677	0,373	0,182
RTD_C_Avg_9	1,691	0,270	0,138
RTD_C_Avg_4	1,694	0,264	0,135
RTD_C_Avg_7	1,679	0,261	0,134
AmbTemp_C_Avg	1,694	0,251	0,129
RTD_C_Avg_5	1,720	0,161	0,086
RTD_C_Avg_3	1,735	0,134	0,072
RTD_C_Avg_2	1,734	0,123	0,066
RTD_C_Avg_8	1,744	0,111	0,060
RTD_C_Avg_6	1,745	0,110	0,059
RTD_C_Avg_1	1,750	0,109	0,059
WindSpeedAve_ms	1,751	0,032	0,018
WindSpeed_ms_Max	1,760	0,007	0,004

De la columna C.* de la Tabla 8 se observa que la variable "RefCell1_Wm2_Avg" tiene mayor ponderación que "SEWSPOAIrrad_Wm2_Avg", por lo que, la primera fue la mejor para representar a la irradiancia solar en un eventual modelo de regresión. Asimismo, la variable óptima para representar a la velocidad del viento fue "WindSpeedAve_ms" por tener mayor peso que "WindSpeed_ms_Max". En cuanto a la temperatura ambiente, la mejor variable fue "SEWSAmbientTemp_C_Avg" al tener mayor peso que "AmbTemp_C_Avg". Finalmente, para representar a la temperatura de los paneles solares fue "SEWSModuleTemp_C_Avg", pues su ponderación final fue superior a la de todos los sensores de temperatura.

Conclusiones

Se desarrolló una metodología multicriterio para la selección óptima de las variables explicativas para modelos de regresión. Para el problema de toma de decisión multicriterio, los criterios de decisión fueron los métodos de selección de características, y las alternativas fueron las variables explicativas de un modelo de regresión. Se ilustró a través del caso de una planta solar fotovoltaica, y se seleccionó las mejores variables para representar

a la irradiancia solar, a la temperatura ambiente, a la temperatura de los paneles solares, y a la velocidad del viento.

Del análisis de correlación se obtuvo que todas las variables explicativas tuvieron coeficientes de correlación positivos con la energía eléctrica AC generada por la planta. De esas variables, las correspondientes a la irradiancia solar fueron las de mayor valor absoluto, luego las asociadas a la temperatura de los paneles solares, la temperatura ambiente y, por último, la velocidad del viento. En cuanto al análisis PCA, se obtuvo que una de las variables de la irradiancia solar fue la de mayor peso con 0,1538, seguida de una de la temperatura ambiente con 0,1235, y la variable que integra los valores de los sensores de la temperatura de los paneles solares ocupó el cuarto puesto con un peso de 0,1028.

De los dos métodos de selección de características de tipo envoltura se puede decir que arrojaron directamente los pesos de importancia de las variables explicativas, siendo los resultados obtenidos similares, y ocupando los primeros lugares las variables asociadas a la irradiancia solar, quedando cerca una de las variables asociadas a la temperatura ambiente. Por otra parte, de la aplicación de la técnica LASSO se obtuvo que sólo las variables asociadas a la irradiancia solar lograron coeficientes diferentes de cero.

Por lo antes expuesto, se recomienda continuar la investigación, aplicando las técnicas de decisión multicriterio no sólo para seleccionar las mejores variables explicativas, sino también para seleccionar el mejor modelo de regresión derivado de la aplicación de algoritmos de aprendizaje automático.

Referencias

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science*. Shelter Island, NY: Manning Publications Co.

Datasheet. (04 de Mayo de 2023). Datasheet. Obtenido de https://www.datasheets.com/en/part-details/nu-u235f2-sharp-46351940#datasheet

Eltarabishi, F., Omar, O., Alsyouf, I., & Bettayeb, M. (2020). Multi-Criteria Decision Making Methods And Their Applications– A Literature Review. *Proceedings of the International Conference on Industrial Engineering and Operations Management* (págs. 2654-2663). Dubai, UAE: IEOM Society International.

Explorium. (24 de June de 2023). Explorium. Obtenido de Access the right data to extend your go-to-market needs: https://www.explorium.ai.

Frederick, O., Maxwell, O., Ifunanya L, O., Udochukwu V, E., Kelechi C, O., Ngonadi O, L., & Kayode Idris, H. (2019). Comparison of Some Variable Selection Techniques in Regression Analysis. *American Journal of Biomedical Science & Research*, 281-293. DOI: 10.34297/AJBSR.2019.06.001044.

Gebreyesus, Y., Dalton, D., Nixon, S., De Chiara, D., & Chinnici, M. (2023). Machine Learning for Data Center Optimizations: Feature Selection Using Shapley Additive explanation (SHAP). *Future Internet MDPI*, https://doi.org/10.3390/fi15030088.

Hackeling, G. (2014). *Mastering Machine Learning with scikit-learn. Birmingham*, UK: Packt Publishing Ltd.

Ishizaka, A., & Nemery, P. (2013). Multi-Criteria Decision Analysis - Methods and Software. *West Sussex*, United Kingdom: John Wiley & Sons, Ltd.

Jomthanachai, S., Wong, W., & Khaw, K. (2022). An application of machine learning regression to feature selection: a study of logistics performance and economic attribute. *Neural Computing and Applications*, 15781–15805. https://doi.org/10.1007/s00521-022-07266-6.

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (págs. 1200-1205). Opatija, Croatia: IEEE Xplore. doi: 10.1109/MIPRO.2015.7160458.

Lee, W. M. (2019). *Python Machine Learning*. Indianapolis: John Wiley & Sons, Inc.

Li, X., Sha, J., & Zhong-Liang, W. (2018). Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. *Environmental Science and Pollution Research*, 19488–19498. https://doi.org/10.1007/s11356-018-2147-3.

Li, Y., Li, G., & Guo, L. (2021). Feature Selection for Regression Based on Gamma Test Nested Monte Carlo Tree Search. *Entropy MDPI*, https://doi.org/10.3390/e23101331.

McKinney, W. (2018). *Python for Data Analysis*. Sebastopol, CA: O'Reilly Media, Inc.

Mesafint Belete, D., & D.H, M. (2020). Wrapper Based Feature Selection Techniques On EDHS-HIV/AIDS Dataset. *European Journal of Molecular & Clinical Medicine*, 2642-2657.

National Institute of Standards and Technology. (04 de Mayo de 2023). National Institute of Standards and Technology. Obtenido de NIST: https://catalog.data.gov/dataset/nist-campus-photovoltaic-pv-arrays-and-weather-station-data-sets-05b4d

Navlani, A., Fandango, A., & Idris, I. (2021). *Python Data Analysis*. Birmingham, UK: Packt Publishing Ltd.

Otchere, D., Arbi Ganat , T., Ojero, J., Tackie-Otoo, B., & Taki, M. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, https://doi.org/10.1016/j.petrol.2021.109244.

Papathanasiou, J., & Ploskas, N. (2018). *Multiple Criteria Decision Aid - Methods, Examples and Python Implementations*. Cham, Switzerland: Springer Nature Switzerland AG.

R, D., Paul, I., Akula, S., Sivakumar, M., & Nair, J. (2020). F-test feature selection in Stacking ensemble model for breast cancer prediction. *Procedia Computer Science*, 1561-1570. http://dx.doi.org/10.1016/j.procs.2020.04.167.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning - Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow*. Birmingham: Packt Publishing Ltd.

Ratner, B. (2017). *Statistical and Machine-Learning Data Mining - Techniques for Better Predictive Modeling and Analysis of Big Data*. Boca Raton, FL: CRC Press Taylor & Francis Group.

Sahoo, B., Behera, R., & Pattnaik, P. (2022). A Comparative Analysis of Multi-Criteria Decision Making Techniques for Ranking of Attributes for e-Governance in India. *International Journal of Advanced Computer Science and Applications*, 65-70. https://dx.doi.org/10.14569/IJACSA.2022.0130311.

SolarDesignTool. (04 de Mayo de 2023). SolarDesignTool. Obtenido de SolarDesignTool site: http://www.solardesigntool.com/components/inverter-grid-tiesolar/PVPowered/137/PVP260kW/specification-data-sheet.html

Triantaphyllou, E., Shu, B., Nieto Sanchez, S., & Ray, T. (1998). Multi-Criteria Decision Making: An Operations Research Approach. *Encyclopedia of Electrical and Electronics Engineering*, 175-186.

Velasquez, M., & Hester, P. (2013). An Analysis of Multi-Criteria Decision Making Methods. *International Journal of Operations Research*, 56-66. http://www.orstw.org.tw/ijor/vol10no2/ijor_vol10_no2_p56_p66.pdf.