MACHINE LEARNING PARA LA PREDICCIÓN DE SERIES TEMPORALES EN INDICADORES DE DESARROLLO MUNDIAL

MIGUEL ALBERTO PLAZAS WADYNSKI

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FISICOMECÁNICAS

ESCUELA DE INGENIERÍA DE SISTEMAS

BUCARAMANGA

2017

MACHINE LEARNING PARA LA PREDICCIÓN DE SERIES TEMPORALES EN INDICADORES DE DESARROLLO MUNDIAL

MIGUEL ALBERTO PLAZAS WADYNSKI

TRABAJO DE GRADO PARA OPTAR AL TÍTULO DE

INGENIERO DE SISTEMAS

DIRECTOR

PH.D RAÚL RAMOS POLLÁN

ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FISICOMECÁNICAS

ESCUELA DE INGENIERÍA DE SISTEMAS

BUCARAMANGA

2017

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

TITULO: MACHINE LEARNING PARA LA PREDICCIÓN DE SERIES TEMPORALES EN INDICADORES DE DESARROLLO MUNDIAL*

AUTOR: MIGUEL ALBERTO PLAZAS WADYNSKI**

PALABRAS CLAVE: Machine Learning, Indicadores de Desarrollo Mundial, Dataset, Estimadores, Árbol de Decisión, Bosque Aleatorio, Máquina de Soporte Vectorial, Clasificación, Regresión, Coeficiente de Determinación.

DESCRIPCIÓN:

Los Indicadores de Desarrollo Mundial, son una base de datos estadística gratuita del Banco Mundial que nos otorga una perspectiva respecto el desarrollo global y calidad de vida de las personas. El Banco Mundial agrupa los indicadores en Conjuntos por temáticas, a saber: Agricultura y desarrollo rural, Cambio climático, Economía y crecimiento, etc. Kaggle, que es una plataforma para la predicción y el análisis de modelos, publicó un Dataset conformado por Indicadores de Desarrollo del Banco Mundial desde 1960 hasta 2015. En este proyecto se aplica Machine Learning para determinar el grado de predictibilidad, las relaciones y dependencias entre los diversos conjuntos de indicadores para Colombia y Suramérica. Para llevar a cabo la tarea de analizar el Dataset y el comportamiento de los indicadores de desarrollo mundial se diseñó un Flujo de Trabajo, donde tenemos un Conjunto Base conformado por los indicadores de desarrollo utilizados para predecir el valor de cada indicador de un conjunto Target. Lo que se hace es Preprocesar los datos tomados del Dataset de Kaggle, e iterar cada conjunto, Base y Target, utilizando los indicadores de un Conjunto Base para predecir cada indicador de un Conjunto Target, por medio de tres estimadores (Árbol de Decisión, Bosque Aleatorio, Máquina de Soporte Vectorial) y determinando para cada conjunto la media del Coeficiente de Determinación (R^2), que es la métrica seleccionada para evaluar el rendimiento de cada uno de los estimadores.

* Trabajo de grado

** Facultad de Ingenierías Físico-Mecánica. Escuela de Ingeniería de Sistemas e Informática. Director: Ph.D Raúl Ramos Pollan.

ABSTRACT

TITLE: MACHINE FOR LEARNING FOR THE PREDICTION OF THE TEMPORARY SERIES IN GLOBAL DEVELOPMENT INDICATORS*

AUTHOR: MIGUEL ALBERTO PLAZAS WADYNSKI**

KEYWORDS: Machine Learning, World Development Indicators, Dataset, Estimators, Decision Tree, Random Forest, Support Vector Machine, Classification, Regression Coefficient of Determination.

DESCRIPTION:

The World Development Indicators, are a free statistical database of the World Bank that gives us a perspective on the global development and quality of life of the people. The World Bank groups the indicators in sets by topics, namely: Agriculture and Rural Development, Climate Change, Economy and Growth, etc. Kaggle, which is a platform for model prediction and analysis, published a Dataset made up of World Bank Development Indicators from 1960 to 2015. In this project, Machine Learning is applied to determine the degree of predictability, relationships and dependencies between the different sets of indicators for Colombia and South America. To carry out the task of analyzing the Dataset and the behavior of the global development indicators a Workflow was designed, where we have a Base Set made up of the development indicators used to predict the value of each indicator of a Target Set. What is done is to preprocess the data taken from the Kaggle Dataset, and iterate each set, Base and Target, using the indicators of a Base Set to predict each indicator of a Target Set, by means of three estimators (Decision Tree, Random Forest, Support Vector Machine) and determining for each set the mean of the Coefficient of Determination (R^2), which is the metric selected to evaluate the performance of each of the estimators.

* Bachelor Thesis

** Faculty of Physical-Mechanical Engieering. School of Engineering and Computer Science. Director: Ph.D Raúl Ramos Pollan.

INTRODUCCIÓN

En nuestros días es de vital importancia la información, el lograr analizar variables específicas de determinado conjunto de datos nos permite visualizar un universo infinito de posibilidades desde cualquier ámbito que se observe bien sea económico, social, o educativo.

El Machine Learning está siendo aplicado de manera innovadora ya que ha permitido acelerar el proceso de automatización de muchos procesos actuales, por lo cual es interesante utilizarlo para analizar el comportamiento de las variables en cualquier entorno, especialmente en el Desarrollo Mundial. Para lo cual, resulta interesante analizar el comportamiento de los Indicadores de Desarrollo Mundial.

Los Indicadores de Desarrollo Mundial (IDM) son una base de datos estadística gratuita del Banco Mundial que nos otorga una perspectiva respecto el desarrollo global y calidad de vida de las personas, los cuales son compilados a partir de fuentes internacionales reconocidas oficialmente. Los IDM presentan los datos de desarrollo mundial más actuales y precisos disponibles, e incluye estimaciones nacionales, regionales y mundiales. La periodicidad de los registros es anual y la frecuencia de actualización es trimestral. De igual forma contiene estadísticas de aproximadamente 250 economías que datan de 1960 hasta la actualidad. [1]

El Banco Mundial agrupa los indicadores por temas en los siguientes conjuntos: Agricultura y desarrollo rural; Eficacia de la ayuda, Cambio climático, Economía y crecimiento; Educación, Energía y minería; Medio ambiente, Deuda externa, Sector financiero, Género, Salud, Infraestructura, Protección social y trabajo; Pobreza, Sector privado, Sector público, Ciencia y tecnología, Desarrollo social, Comercio y Desarrollo urbano. [2]

1 OBJETIVOS

1.1 Objetivo General

Definir, implementar y evaluar modelos predictivos de series temporales de indicadores de desarrollo del Banco Mundial sobre Colombia.

1.2 Objetivos Específicos

- Caracterizar el registro histórico de indicadores de desarrollo del Banco Mundial (RHID) de forma general, y consecuentemente tomar como enfoque principal Colombia y secundariamente los países de la región sur americana.
- Construir un flujo de trabajo para pre-procesar los datos del RHID.
- Construir modelos predictivos y medir su capacidad de predicción.
- Determinar qué indicadores de desarrollo son más valiosos para el desarrollo de modelos predictivos basados en Machine Learning para Colombia.
- Evaluar la confiabilidad estadística de los métodos desarrollados.

2 JUSTIFICACIÓN

Los Indicadores del Banco Mundial se usan en estudios estadísticos, proyecciones económicas y de manera general su aplicación más notoria es su participación en el soporte del cumplimiento de la Agenda 2030 para el Desarrollo sostenible.

La Asamblea General de la ONU adoptó el mes de septiembre del año 2015 la Agenda 2030 para el Desarrollo Sostenible, un plan de acción a favor de las personas, el planeta y la prosperidad, que también tiene la intención de fortalecer la paz universal y el acceso a la justicia.

La Agenda plantea 17 objetivos ilustrados en la Figura 1. los cuales son de carácter integrado e indivisible que abarcan las esferas económica, social y ambiental. [3]

Figura 1: *Objetivos de desarrollo sostenible según la ONU*

Fuente: *Naciones Unidas, "Objetivos de Desarrollo Sostenible", 2015. [En línea]. (Recuperado 10 de junio de 2017). Disponible en http://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/*

Es en este punto donde estos indicadores juegan un papel vital siendo el pilar angular por el cual se analiza e informa el progreso del cumplimiento de los mencionados objetivos.

El aplicar Machine Learning a estos datos nos otorga la posibilidad de encontrar relaciones y dependencias entre los indicadores de los países, así como también poder predecir indicadores o tendencias futuras.

En 2010, Kaggle[1] fue fundada como una plataforma para la predicción de modelos y análisis de competiciones en la que las empresas y los investigadores publican sus datos y los estadísticos y los mineros de datos de todo el mundo compiten para producir los mejores modelos. Kaggle también organiza concursos de reclutamiento en los cuales los científicos de datos compiten por la oportunidad de entrevistarse en empresas líderes en ciencia de datos como Facebook, Winton Capital y Walmart.

En enero de 2016, Kaggle lanzó su producto *Dataset*[2], haciendo una selección de conjuntos de datos públicos disponibles en Kaggle. Cada conjunto de datos tiene habilitados Scripts, así como un foro dedicado, que permite la conversación y la colaboración entre los científicos de datos y el trabajo que están haciendo en cada conjunto de datos. [4]

De hecho, Kaggle, tiene una competición para el análisis de los IDM por medio de un *Dataset* recopilado directamente de la base de datos pública del Banco Mundial. De allí fue de donde extraje el *Dataset* con los Indicadores registrados hasta el año de 2015.[3]

[1] https://www.kaggle.com/
[2] Dataset significa Conjunto de Datos
[3] https://www.kaggle.com/worldbank/world-development-indicators

3 MARCO TEÓRICO

3.1 Aprendizaje Automático

El Aprendizaje Automático o Machine Learning es una subárea central de la inteligencia artificial cuyo objetivo es diseñar algoritmos de aprendizaje que realicen el aprendizaje automáticamente sin asistencia o intervención humana.

El aprendizaje que se hace se basa siembre en algún tipo de observaciones o datos con el objetivo de encontrar patrones ó hacer algo mejor basándose en registros.[5]

Algunos ejemplos de Problemas de Aprendizaje Automático son: Reconocimiento de Óptico de caracteres, Detección de rostros, filtrado de correo basura, detección de fraude, diagnósticos médicos, motores de búsqueda.

En Machine Learning, la mayoría de *Dataset*, pueden ser representados como tablas que contienen valores numéricos. Cada fila es denominada como una observación, una muestra, o un dato puntal. Cada columna es denominada como una característica o una variable.

Vamos a llamar N el número de filas (o el número de puntos), y D el número de columnas (o número de características). El número D también es llamado dimensionalidad de los datos. La razón es porque podemos ver esta tabla como un conjunto E de vectores en un espacio con D dimensiones (o espacio vectorial). Aquí, un vector x contiene D números $(x_1, \dots, x_D)$, también llamado componentes.

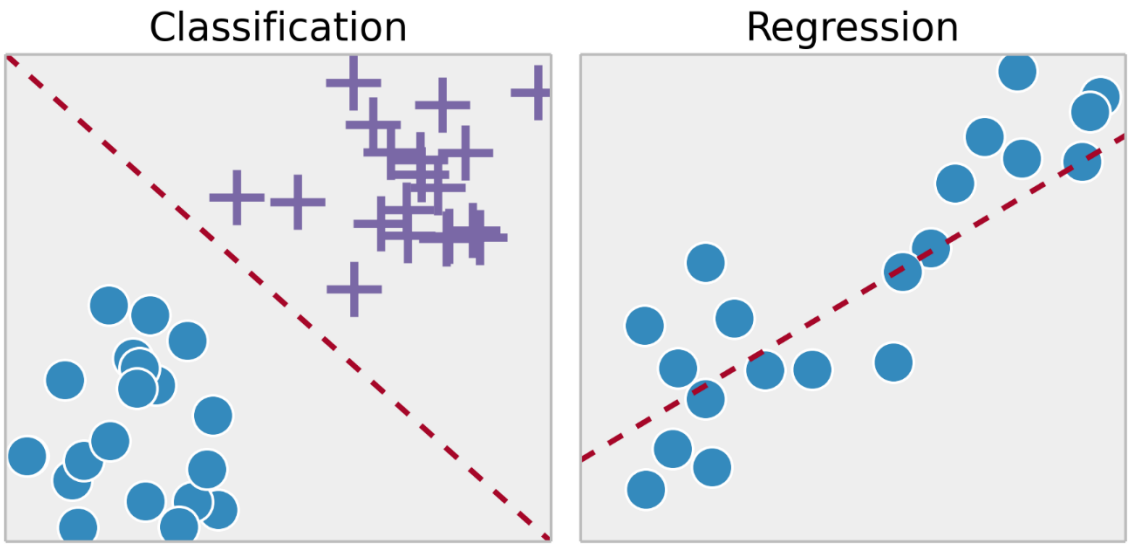
Generalmente se hace una distinción entre aprendizaje supervisado y aprendizaje no supervisado.

3.1.1 Aprendizaje Supervisado. Es cuando tenemos una etiqueta $\boldsymbol{y}$ asociada a cada punto de dato $\boldsymbol{x}$. El objetivo es aprender el mapeo de $\boldsymbol{x}$ a $\boldsymbol{y}$ de nuestros datos. Los datos nos dan este mapeo para un conjunto finito de puntos, pero lo que queremos es generalizar este mapeo. En otras palabras, queremos encontrar la etiqueta de cualquier punto $\boldsymbol{x}$ que no pertenece a nuestros datos.
Matemáticamente, el aprendizaje supervisado consiste en encontrar una función f que mapea un conjunto de puntos E a un conjunto de etiquetas F, conociendo un conjunto finito de asociaciones (x, y) lo cual es dado por nuestros datos. De forma general: Después de observar los pares (x_i, y_i), dado un nuevo x, podemos encontrar su correspondiente y aplicando la función f a x.

Una práctica común es dividir un conjunto de datos en dos subconjuntos: un conjunto de entrenamiento y un conjunto de prueba. Aprendemos la función f en el conjunto de entrenamiento, y la probamos en el conjunto de prueba. Esto es esencial para evaluar el poder predictivo de un modelo. Así mismo, generalmente se hace una distinción entre clasificación y regresión, dos casos particulares del aprendizaje supervisado se ilustran en la Figura 2.

- **Clasificación:** es cuando las etiquetas y pueden tomar sólo un número finito de valores (categorías). Cuando solo hay dos categorías, el problema se denomina clasificación binaria, y cuando hay más de dos categorías, el problema se denomina clasificación multiclase. Algunos ejemplos:
    - Reconocimiento de dígitos: x es una imagen con dígitos escritos a mano, y es un dígito entre 0 y 9.
    - Filtrado de Spam: x es un e-mail, y y es 0 o 1 si ese e-mail es Spam o no lo es.
- **Regresión:** es cuando las etiquetas de y pueden tomar cualquier valor real (continuo). Algunos ejemplos incluyen:
    - Predicción del Stock de un Local.
    - Predicción de ventas.
    - Detección de la edad de una persona en base a una fotografía.

Figura 2: *Diferencias entre clasificación y regresión.*



Fuente: *C. Rossant, "IPython Interactive Computing and Visualization Cookbook," 2014, pp. 269.*

3.1.2 Aprendizaje No Supervisado. Es cuando no tenemos ninguna etiqueta. Lo que queremos hacer es descubrir alguna estructura oculta en los datos. Es más difícil de comprender que el aprendizaje supervisado, el sentido de que no hay una pregunta y una respuesta precisas en general. A continuación, estos son algunos términos importantes relacionados con el aprendizaje no-supervisado:

- **Clustering:** Agrupamiento de puntos similares dentro de clústeres.
- **Estimación de densidad:** Estimación de una densidad de probabilidad que puede explicar la distribución de los puntos de los datos.
- **Reducción de dimensión:** Obtener una representación simple de los puntos de los datos que tienen una alta dimensionalidad mediante la proyección de estos en un espacio de menor dimensión. Esta técnica se utiliza especialmente para la visualización de datos. [6]

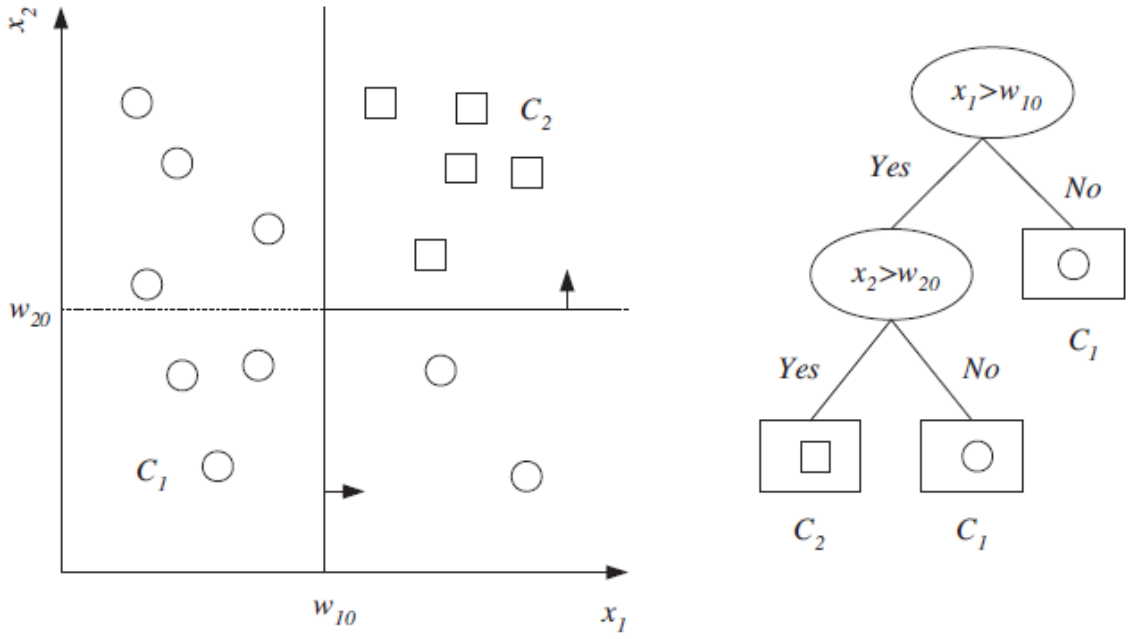
3.2 Métodos de Machine Learning

En Machine Learning tenemos diversos métodos para resolver los problemas de regresión o de clasificación. Como veremos más adelante el *Dataset* de los indicadores de desarrollo están compuestos por valores reales continuos, por lo cual la tarea a la cual nos enfrentamos es de regresión. Así pues, se decidió utilizar tres métodos de aprendizaje supervisado para esta tarea de regresión: *Decision Tree Regressor* (DTR), *Support Vector Regressor* (SVR) y *Random Forest Regressor* (RFR). Se seleccionaron estos tres métodos en particular debido a que su costo computacional es apropiado para el flujo de trabajo planteado más adelante.

3.2.1 Decision Tree. Un *Decision Tree* o *Árbol de Decisión* es una estructura de datos jerárquica que implementa la estrategia de dividir y conquistar. Es un método no paramétrico eficiente, que puede usarse tanto para la clasificación como para la regresión. [7]

De forma más detallada, es un modelo jerárquico para el aprendizaje por lo cual una región local es identificada en una secuencia de divisiones recursivas en un número menor de pasos. Un árbol de decisiones está compuesto por nodos de decisión internos y hojas terminales, en la Figura 3. se detalla un ejemplo. Cada nodo de decisión m implementa una función de prueba $f_m(x)$ con resultados discretos etiquetando las ramas. Dada una entrada, en cada nodo, se aplica una prueba y se toma una de las ramas dependiendo del resultado. Este proceso comienza en la raíz y se repite recursivamente hasta que se golpea un nodo hoja, momento en el cual el valor escrito en la hoja constituye la salida.

Figura 3: *Ejemplo de un Dataset y su correspondiente Árbol de Decisión*



Fuente: *E. Alpaydın, "Introduction to Machine Learning Second Edition," 2 edition., London, England: MIT Press, 2010, pp. 186.*

Un árbol de decisión es también un modelo no paramétrico en el sentido de que no asumimos ninguna forma paramétrica para las densidades de clase y la estructura de árbol no se fija a priori pero el árbol crece, las ramas y las hojas se agregan, durante el aprendizaje dependiendo de la complejidad del problema inherente a los datos.

Cada $f_m(x)$ define un discriminante en el espacio de entrada d-dimensional dividiéndolo en regiones más pequeñas que se subdividen más a medida que tomamos un camino desde la raíz hacia abajo. $f_m(\cdot)$ es una función simple y cuando se escribe como un árbol, una función compleja se descompone en una serie de decisiones simples. Diferentes métodos de árboles de decisión asumen diferentes modelos para $f_m(\cdot)$, y la clase de modelo define la forma del discriminante y la forma de las regiones. Cada nodo hoja tiene una etiqueta de salida, que en el caso de la clasificación es el código de clase y en la regresión es un valor numérico.

Un nodo hoja define una región localizada en el espacio de entrada donde las instancias que caen en esta región tienen las mismas etiquetas (en la clasificación) o salidas numéricas muy similares (en regresión). Los límites de las regiones están definidos por los discriminantes que están codificados en los nodos internos en el camino desde la raíz hasta el nodo de la hoja.

La colocación jerárquica de las decisiones permite una rápida localización de la región que cubre una entrada. Por ejemplo, si las decisiones son binarias, entonces en el mejor de los casos, cada decisión elimina la mitad de los casos. Si hay b regiones, entonces en el mejor de los casos, la región correcta se puede encontrar en $\log_2 b$ decisiones.

3.2.1.1 Decision Tree Clasiffier. En un *Decision Tree Clasifier* o *Árbol de Decision de Clasificación* tenemos que la buena calidad de una división se cuantifica por una medida de impurezas. Una división es pura si después de la división, para todas las ramas, todas las instancias que eligen una rama pertenecen a la misma clase. Sea para el nodo $\boldsymbol{m}$, $\boldsymbol{N_m}$ es el número de instancias de entrenamiento que alcanzan el nodo $\boldsymbol{m}$. Para el nodo raíz, es $\boldsymbol{N}$. $\boldsymbol{N_m^i}$ de $\boldsymbol{N_m}$ pertenece a la clase $\boldsymbol{C_i}$ con $\sum_i \boldsymbol{N_m^i} = \boldsymbol{N_m}$. Dado que una instancia alcanza un nodo $\boldsymbol{m}$, el estimado para la probabilidad de la clase $\boldsymbol{C_i}$ es:

$$\hat{P}(C_i|x, m) \equiv p_m^i = \frac{N_m^i}{N_m}$$

3.2.1.2 Decision Tree Regresor. Un *Decision Tree Regresor* o *Árbol de Decisión Regresor* se construye casi de la misma manera que uno de clasificación, excepto que la medida de impurezas apropiada para la clasificación se sustituye por una medida apropiada para la regresión. Digamos que para el nodo $\boldsymbol{m}$, $\boldsymbol{X_m}$ es el subconjunto de $\boldsymbol{X}$ que llega al nodo $\boldsymbol{m}$; es decir, es el conjunto de todo $\boldsymbol{x} \in \boldsymbol{X}$ que satisface todas las condiciones en los nodos de decisión en el camino desde la raíz hasta el nodo $\boldsymbol{m}$.

Sea:

$$b_m(x) = \begin{cases} 1 \; si \; x \; \in \; X_m : x \; alcanza \; el \; nodo \; m \\ 0 \; de \; otra \; manera \end{cases}$$

En regresión, la buena calidad de una división se mide por el error cuadrático medio del valor estimado.

Recordemos que:

$$E_m = \frac{1}{N_m} \sum_t (r^t - g_m)^2 b_m(x^t)$$

Donde:

$$N_m = |X_m| = \sum_t b_m(x^t)$$

$$g_m = \frac{\sum_t b_m(x^t)\, r^t}{\sum_t b_m(x)}$$

$$r^t = g(x^t) + \in ,$$

En un nodo, se usa la media de las salidas requeridas de las instancias que alcanzan todo el nodo. [7]

3.2.2 Support Vector Machine. Un *Support Vector Machine* (SVM) ó *Máquina de Soporte Vectorial* es una representación de las muestras de entrenamiento como puntos en el espacio, mapeados de modo que las muestras de las diversas categorías son divididas por una clara brecha considerada como hiper-plano que es tan amplia como sea posible. Un hiper-plano es una línea que divide el espacio de una variable de entrada.
En SVM, se selecciona un hiper-plano que separe mejor los puntos en el espacio de la variable de entrada en base a su clase.

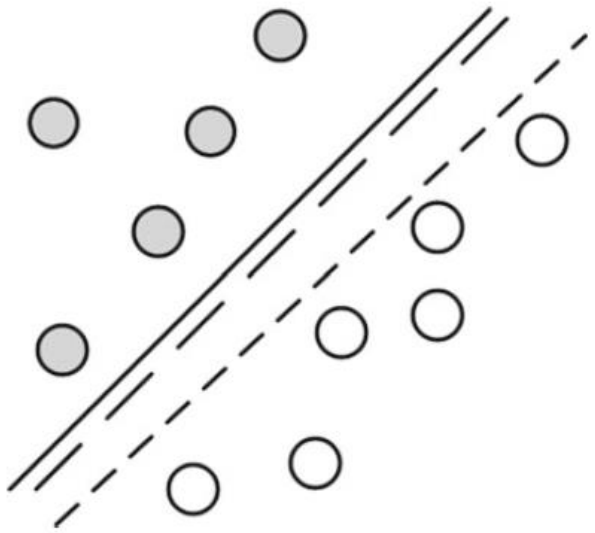
Luego se mapean nuevas muestras dentro del mismo espacio y se predice que pertenezcan a una categoría u otra basándose en qué lado de la brecha caen. [8]

3.2.2.1 Support Vector Clasiffier. Considere el problema de clasificación bidimensional mostrado en la Figura. 4. Esta figura nos muestra dos clases (círculos grises y blancos) que pueden ser separados por cualquiera de las líneas mostradas. Específicamente, cualquiera de tales líneas de separación puede escribirse como el lugar de los puntos ($\boldsymbol{x}$) en el plano bidimensional que satisfacen lo siguiente,

$$\beta_0 + \beta^T x = 0$$

Para clasificar una x arbitraria usando esta línea, sólo se calcula el signo de $\beta_0 + \beta^T x$ y se asigna una clase al signo positivo y la otra clase al signo negativo. Para especificar únicamente una línea de separación (o, hiperplano en un espacio de mayor dimensión) necesitamos criterios adicionales.

Figura 4: Clasificación Bidimensional

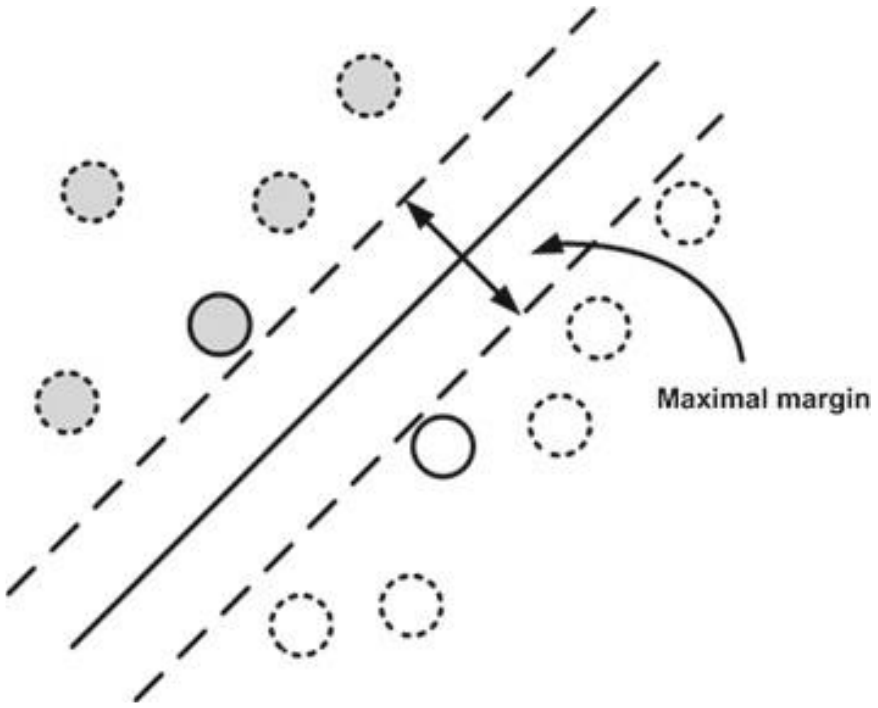


Fuente: *J. Unpingco, "Python for Probability, Statistics, and Machine Learning" Springer, 2016, pp. 250.*

Ahora bien, en la Figura 5. Se muestran los datos con dos líneas paralelas a la frontera que forman un margen alrededor de la línea central de separación. El algoritmo de *Maximal marigin* o Margen Máximo, encuentra la margen más amplia y la única línea separadora. Como consecuencia, el algoritmo descubre los elementos de los datos que tocan los márgenes. Estos son los elementos de apoyo.

Los otros elementos alejados de la frontera no son relevantes para la solución. Esto reduce la varianza del modelo porque la solución es insensible a la eliminación de elementos distintos a estos elementos de soporte (por lo general una pequeña minoría). [9]

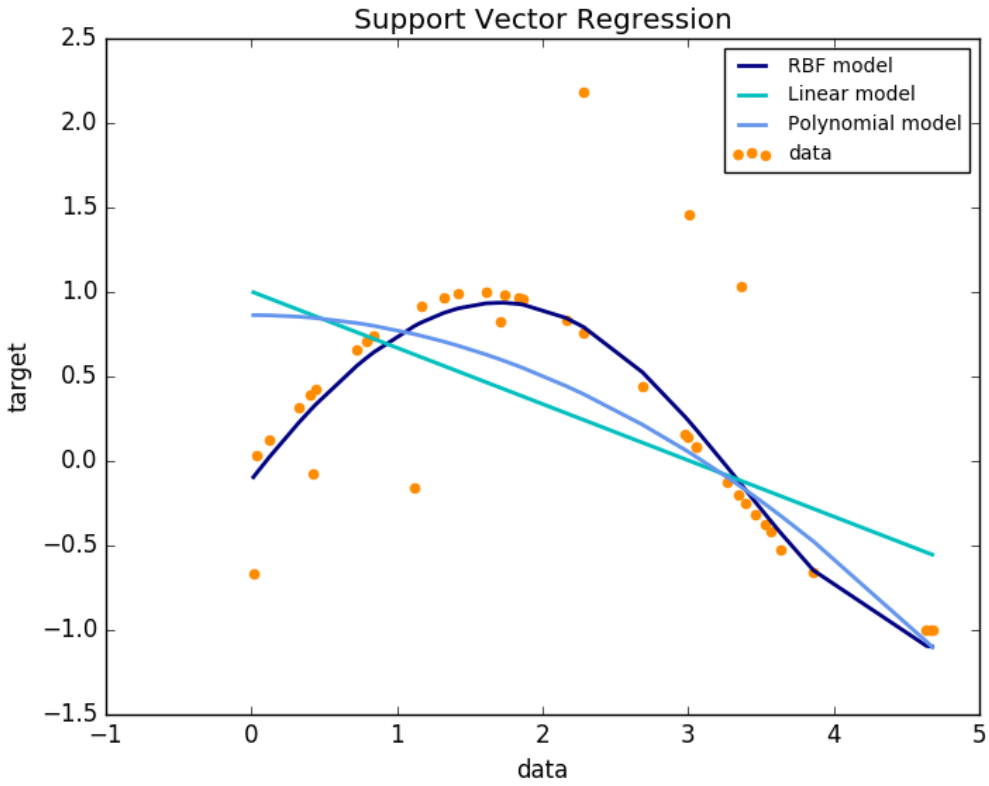
Figura 5: Algoritmo de margen máximo



Fuente: *J. Unpingco, "Python for Probability, Statistics, and Machine Learning" Springer, 2016, pp. 251.*

3.2.2.2 Support Vector Regressor. El método de Clasificación de Soporte Vectorial puede extenderse para resolver problemas de regresión. El modelo producido por la Regresión de Soporte Vectorial sólo depende de un subconjunto de los datos de entrenamiento, porque la función de costo para construir el modelo ignora cualquier información de entrenamiento cercana a la predicción del modelo. Al igual que con las clases de clasificación, el método de ajuste tomará como argumento vector de $\boldsymbol{x}$ e $\boldsymbol{y}$, sólo que en este caso $\boldsymbol{y}$ se espera que tenga valores de coma flotante en lugar de valores enteros. [10]

Figura 6: Vector de Soporte Regresor utilizando Kernels lineales y no-lineales



Fuente: *Scikit-learn developers, "Support Vector Regression (SVR) using linear and non-linear kernels." [En línea]. (Recuperado 13 de junio de 2017). Disponible en http://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html#sphx-glr-auto-examples-svm-plot-svm-regression-py.*

En la Figura 6. se ilustra un ejemplo de utilizar un Vector de Soporte Regresor utilizando Kernels lineales y no-lineales en una dimensión.

3.2.3 Random Forest. Los *Random Forest* (RF) ó *Bosques Aleatorios,* son conjuntos de árboles de decisión. Múltiples Árboles de decisión son entrenados y agregados para formar un modelo que es más robusto que cualquiera de los árboles individuales. [11]

Una forma de reducir la varianza de una estimación es promediar juntas muchas estimaciones.

Por ejemplo, es posible entrenar M árboles diferentes en diferentes subconjuntos de los datos, escogiendo aleatoriamente con reemplazo, y luego computar el conjunto, así:

$$f(x) = \sum_{m=1}^{M} \frac{1}{M} f_m(x)$$

Donde f_m es el m-ésimo árbol. Esta técnica es conocida como *bagging* (Breiman 1996), Que significa "Agregación de bootstrap".

Desafortunadamente, simplemente volver a ejecutar el mismo algoritmo de aprendizaje en diferentes subconjuntos de los datos puede resultar en Base altamente correlacionados, lo que limita la cantidad de reducción de la varianza que es posible. La técnica conocida como *Bosques Aleatorios* (Breiman 2001a) trata de des-correlacionar la base de los aprendices aprendiendo árboles basados en un subconjunto de variables de entrada elegido al azar, así como un subconjunto de casos de datos elegido aleatoriamente. Tales modelos tienen a menudo una muy buena precisión predictiva (Caruana y Niculescu-Mizil 2006), y se han utilizado ampliamente en muchas aplicaciones (por ejemplo, para el reconocimiento de la pose del cuerpo usando el popular sensor de Kinect de Microsoft (Shotton et al., 2011)). [12]

3.2.3.1 Random Forest Classifier. Para tareas de clasificación, se utilizan *Árboles de Decisión Clasificadores*, los cuales utilizan el criterio de división de Gini, que se calcula así:

$$Gini = n_I \sum_{k=1,\dots,K} p_{kI}(1 - p_{kI}) + n_D \sum_{k=1,\dots,K} p_{kD}(1 - p_{kD})$$

Donde,

$$p_{kI} = Proporción\ de\ la\ Clase\ k\ en\ el\ nodo\ Izquierdo$$

$$p_{kD} = Proporción\ de\ la\ Clase\ k\ en\ el\ nodo\ Derecho$$

3.2.3.2 Random Forest Regresor. Para tareas de regresión, se utilizan *Árboles de Decisión Regresores*, los cuales utilizan el criterio de división de Suma Residual de Cuadrados [13], que se calcula así:

$$SRC = \sum_{Izquierda}(y_i - {y_I}^*)^2 + \sum_{Derecha}(y_i - {y_D}^*)^2$$

Donde,

$${y_I}^* = Promedio\ del\ valor\ de\ y\ para\ el\ nodo\ Izquierdo$$

$${y_D}^* = Promedio\ del\ valor\ de\ y\ para\ el\ nodo\ Derecho$$

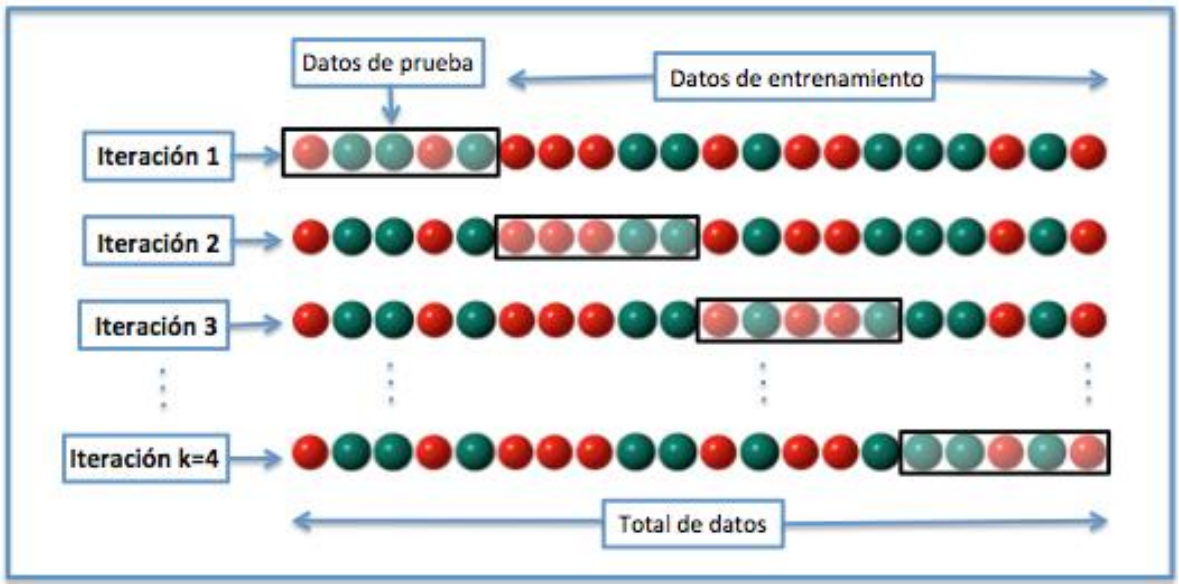
3.3 Selección de modelos y validación

Para tener una confianza en el resultado en el proceso de entrenamiento de los estimadores utilizamos el método denominado *CrossValidation* ó Validación Cruzada y así mismo con el objetivo de mejorar el rendimiento y optimizar los métodos de Machine Learning hemos probado un método de optimización denominado *GridSearchCV*.

3.3.1 CrossValidation. *CrossValidation* ó Validación Cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y asegurar que son independientes de la partición entre los datos de entrenamiento y de prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación en diferentes particiones.

3.3.1.1 K-Fold. En la validación cruzada de K iteraciones o K-fold cross-validation los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. En la Figura 7. es posible apreciar una ilustración de este iterador.

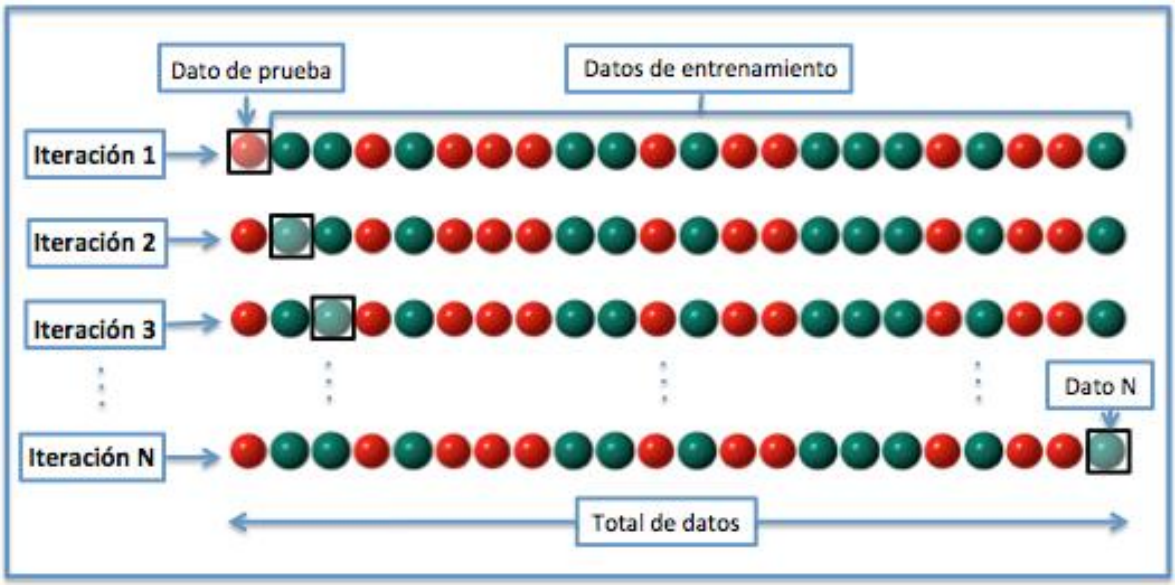
Figura 7:Validación cruzada de K iteraciones con K igual a 4



Fuente: *Wikipedia contributors, "Validación cruzada de K iteraciones con K=4" Wikipedia, La enciclopedia libre., 2017. [En Línea]. (Recuperado 14 de junio de 2017). Disponible en https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada#/media/File:K-fold_cross_validation.jpg*

3.3.1.2 Leave One Out. La *Validación Cruzada Dejando Uno Fuera* o *Leave-one-out cross-validation* (LOOCV) implica separar los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento. En la Figura 8. se aprecia una representación de este método. [14]

Figura 8: Validación cruzada dejando uno fuera



Fuente: *Wikipedia contributors, "Validación cruzada dejando uno fuera" Wikipedia, La enciclopedia libre., 2017. [En Línea]. (Recuperado 14 de junio de 2017). Disponible en https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada#/media/File:Leave-one-out.jpg*

3.3.1.3 ShuffleSplit. Existen utilidades para generar índices que se pueden utilizar para generar divisiones de datos según las diferentes estrategias de Validación Cruzada. El iterador *ShuffleSplit* genera un número de particiones independientes de conjuntos de datos de train/test. Las muestras se barajan primero y luego se dividen en train y test. [15]

3.3.2 GridSearchCV. En Machine Learning se tiene el problema de elegir un conjunto de hiperparámetros para un algoritmo con el objetivo de optimizar el rendimiento, por lo que la forma tradicional de realizar la optimización del hiperparámetro es un *GridSearchCV*, que simplemente es una búsqueda exhaustiva a través de un subconjunto especificado manualmente del espacio del hiperparámetro de un algoritmo de aprendizaje. [16]

3.4 Métrica de Rendimiento

La métrica de rendimiento seleccionada es el Coeficiente de Determinación (R^2), que es un número que indica la proporción de la varianza en la variable dependiente que es objetivo de la variable independiente. Ese coeficiente proporciona una medida de cuán bien se prevén las futuras muestras por el modelo. La mejor puntuación posible es 1.0, y puede ser negativo (porque el modelo puede ser arbitrariamente peor). Si $\widehat{y_i}$ es el valor predicho de la i-ésima muestra y el y_i es el correspondiente valor verdadero, luego la puntuación R^2 estimada sobre $n_{muestras}$ muestras se definida así:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{muestras}-1}(y_i - \widehat{y_i})^2}{\sum_{i=0}^{n_{muestras}-1}(y_i - \bar{y})^2}$$

Donde $\bar{y} = \frac{1}{n_{muestras}} \sum_{i=0}^{n_{muestras}-1} y_i$ [17]

4 ESTADO DEL ARTE

El Machine Learning aplicado a los Indicadores de desarrollo mundial ha tenido algunos estudios desarrollados como, por ejemplo: Estimar el volumen de tráfico ferroviario basándose en los indicadores de desarrollo mundial. La política de transporte europea que se define en *este artículo* apoya el desplazamiento de la carretera al ferrocarril y al transporte por barco. La hipótesis de este artículo es que los cambios en el entorno económico incluyen en el volumen del tráfico ferroviario. Por lo tanto, un modelo para la predicción del volumen de tráfico ferroviario aplicado en diferentes contextos económicos podría ser una herramienta valiosa para los planificadores del transporte. El modelo fue construido utilizando técnicas comunes de Machine Learning que se aprenden de la experiencia pasada.

En la preparación del modelo, los indicadores de desarrollo mundial definidos por el Banco Mundial fueron utilizados como parámetros de entrada. [18]

De forma similar se planteó un Modelo ecológico de la huella usando la técnica de Máquina de Soporte Vectorial. La huella ecológica per cápita (HE) es una de las medidas de sostenibilidad ambiental más reconocidas. Su objetivo es cuantificar los recursos biológicos de la Tierra necesarios para apoyar la actividad humana. En este trabajo, nos presentan cinco factores que influyen la HE per cápita. Estos factores son: Producto Interno Bruto (PIB) nacional, urbanización (independiente del desarrollo económico), distribución del ingreso (medido por el coeficiente de Gini), dependencia de las exportaciones (medida por el porcentaje de exportaciones al PIB total), e intensidad del servicio (medido por el porcentaje de servicio al PIB total). Se llevó a cabo un nuevo modelo de huella ecológica basado en una Maquina de Soporte Vectorial (SVM), que es un método de Machine Learning basado en el principio de minimización del riesgo estructural de la teoría del aprendizaje estadístico para calcular la HE per cápita de 24 naciones utilizando datos de 123

naciones. La precisión del cálculo se midió por error absoluto promedio y error relativo promedio. Los resultados fueron 0,004883 y 0,351078% respectivamente. Estos resultados demuestran que el modelo HE basado en SVM tiene un buen rendimiento de cálculo. [19]

Así mismo, resulta interesante encontrar que también se ha efectuado un Análisis de Componentes Principales sobre Indicadores de Desarrollo Humano de China. En este estudio se utilizó el análisis del componente principal ponderado para medir y analizar el progreso del desarrollo humano en las provincias chinas desde 1990. Las tendencias del desarrollo humano en el período de la transición del mercado en varias provincias de China se discutieron en términos de impacto en la salud pública, así como el desarrollo económico. La asociación del componente principal obtenido del estudio y el índice de desarrollo humano notificado por el Programa de las Naciones Unidas para el Desarrollo fue estimado por el coeficiente de correlación de rango de Spearman. [20]

También encontramos un trabajo en el que se plantearon los Indicadores de desarrollo para la búsqueda de un criterio de necesidades básicas. La medición de los esfuerzos de desarrollo en los países en desarrollo por lo general se ha centrado en el crecimiento del PNB per cápita y conceptos relacionados. Cada vez más, los economistas del desarrollo se han dado cuenta de que el crecimiento del producto o de los ingresos por sí mismos no son indicadores adecuados del desarrollo y que la reducción de la pobreza y la satisfacción de las necesidades humanas básicas son objetivos que deben aparecer en una medida de desarrollo. Ha habido un creciente interés en diseñar mejores medidas de desarrollo, incluyendo modificaciones del PNB, indicadores sociales y sistemas asociados de cuentas sociales, e índices compuestos de desarrollo. Una revisión de estos enfoques y conceptos apunta a la conclusión de que el uso de indicadores sociales y humanos es el suplemento más prometedor para el PNB, particularmente si el trabajo sobre indicadores sociales se hace en áreas centrales del enfoque de necesidades básicas. [21]

Resulta igualmente interesante que ya se ha efectuado un estudio de los Indicadores de desarrollo en el turismo y desarrollo sostenible. En este documento se esboza el contexto histórico en el que aparecen los indicadores y se vincula con la necesidad de mejorar los sistemas de información. Las conclusiones se hacen con particular atención a las cuestiones emergentes, la más reciente de las cuales es la del desarrollo sostenible. Debido a la poca fiabilidad de los datos ya la dificultad de definir los límites del turismo como actividad económica, la articulación de conjuntos de indicadores de desarrollo sostenible (IDS) parece ser aún más difícil para el turismo que para otros sectores industriales. Los intentos recientes y actuales muestran una gran variedad de métodos y resultados. Se refieren:

- Las diversas exigencias impuestas a los datos,
- La escala geográfica a la que se refieren los indicadores y
- El tipo de política que se invita a fomentar el desarrollo sostenible: políticas públicas, Autorregulación, etc.

Los resultados indican que un cierto conjunto de cuestiones planteadas por el desarrollo sostenible es privilegiado, mientras que otros son dejados de lado. Esto subraya la necesidad de revisar los aspectos más teóricos de los debates sobre el desarrollo sostenible dentro de los intentos prácticos de construir indicadores. [22]

5 METODOLOGÍA

5.1 Volumetría de los Datos

Los datos fueron extraídos de Kaggle y estos a su vez provienen del Banco Mundial. Se tiene exactamente **5'656.458** de datos registrados en 6 columnas que almacenan respectivamente: Nombre del Pais, Código del Pais, Nombre del Indicador, Código del Indicador, Año y Valor registrado. Un ejemplo en la Tabla 1.

Tabla 1: Muestra del Dataset extraído de Kaggle

Index	Country Name	Country Code	IndicatorName	IndicatorCode	Year	Value
5646130	Colombia	COL	Time required to register property (days)	IC.PRP.DURS	2015	16.0
5646131	Colombia	COL	Time required to start a business (days)	IC.REG.DURS	2015	11.0
5646132	Colombia	COL	Time to prepare and pay taxes (hours)	IC.TAX.DURS	2015	239.0
5646133	Colombia	COL	Time to resolve insolvency (years)	IC.ISV.DURS	2015	1.7
5646134	Colombia	COL	Total tax rate (% of commercial profits)	IC.TAX.TOTL.CP.ZS	2015	6.97

Un análisis inicial de la volumetría del Dataset se muestra en la Tabla 2.

Tabla 2: Volumetría del Dataset

	Datos	Países	Indicadores	Año Mínimo	Año Máximo
Total	5.656.458	247	1344	1960	2015
Suramérica	969.259	41	1339	1960	2015
Colombia	37.227	1	1299	1960	2015

Así mismo es importante tener en cuenta la cantidad de indicadores que cada conjunto tiene, lo cual se ilustra en la Tabla 3.

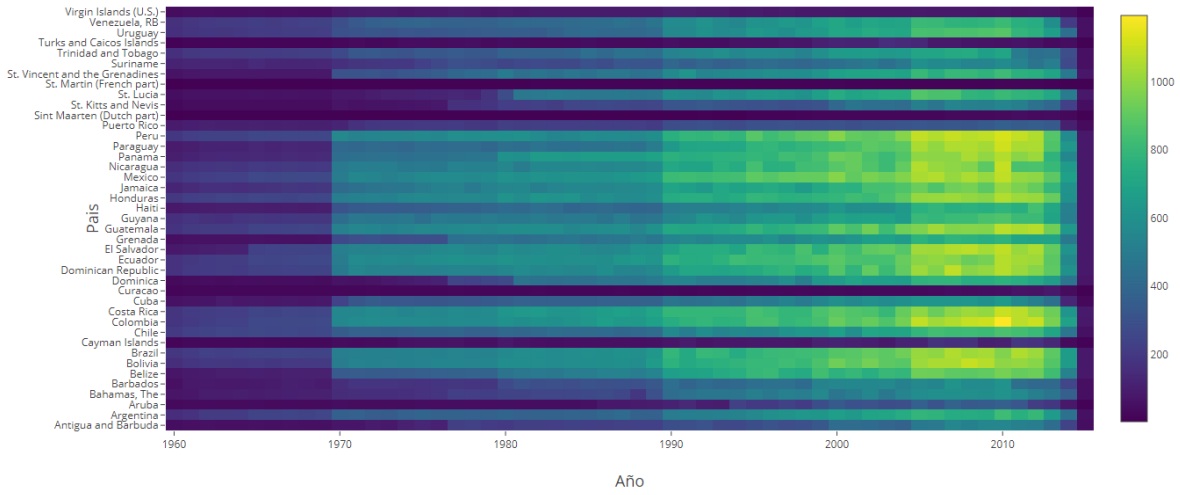
Tabla 3: Cantidad de indicadores registrados para cada conjunto

Etiqueta	Conjunto	Número de Indicadores
Agricultura	Agricultura y desarrollo rural	48
Ambiente	Medio ambiente	112
Ayuda	Eficacia de la ayuda	71
Ciencia	Ciencia y tecnología	13
Clima	Cambio climático	80
Comercio	Comercio	144
Deuda	Deuda externa	229
Economía	Economía y crecimiento	260
Educación	Educación	168
Energía	Energía y minería	53
Finanzas	Sector financiero	67
Género	Género	177
Infraestructura	Infraestructura	43
Pobreza	Pobreza	25
Privado	Sector privado	165
Público	Sector público	97
Salud	Salud	152
Social	Desarrollo social	35
Trabajo	Protección social y trabajo	133
Urbano	Desarrollo urbano	24
Total		**2096**

Observando se tiene un total de 2096 indicadores agrupados según el Banco Mundial, sin embargo, el *Dataset* contiene un máximo de 1344, este fenómeno sucede porque el Banco Mundial repite indicadores en los diversos conjuntos.

5.1.1 Suramérica. Para comprender mejor el número de indicadores registrados a lo largo de los años para cada país de Suramérica se diseñó un mapa de densidad que se puede observar en la Figura 9.

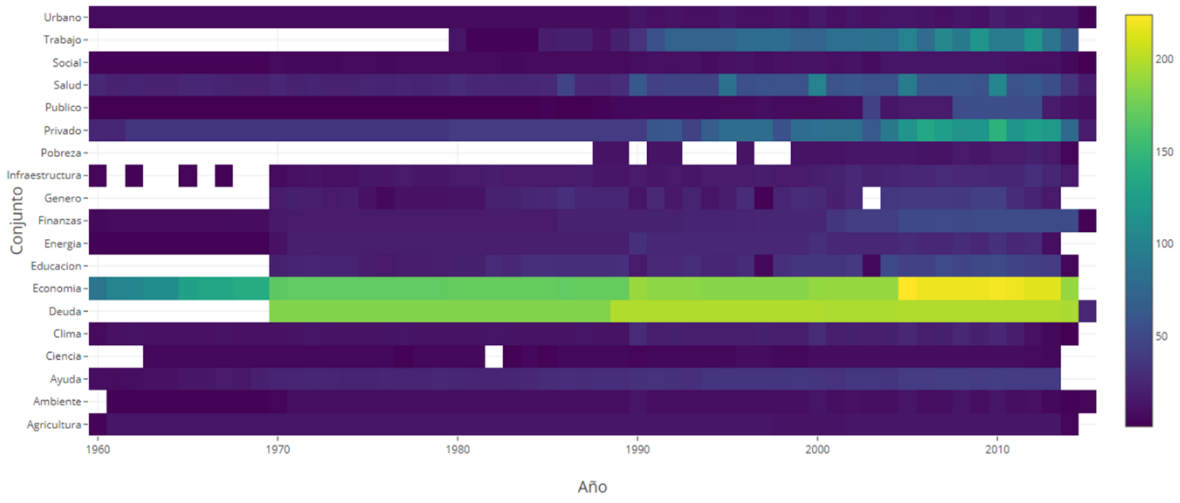
Figura 9: Indicadores registrados por país en Suramérica



En este mapa de densidad se observa que la mayor densidad de indicadores registrados para Suramérica se encuentra en la década más reciente, es decir, en el rango de años de 2005 a 2015. También es importante notar que en cada década hay un cambio significativo en el registro de indicadores, es decir, que a lo largo de los años ha ido aumentando de manera exponencial el registro de datos para este continente. Así como también que, en el último año registrado, es decir, 2015, casi no hay registros y esto sucede debido a que es el año en el que se publicó el *Dataset*, por lo cual es posible que ese mismo año no hayan estado listos todos los informes de las diversas fuentes del Banco Mundial, lo que explicaría el porqué hay tan pocos indicadores registrados ese año.

5.1.2 Colombia. Así mismo se decidió que para Colombia se analizaría también la cantidad de indicadores registrados, sólo que esta vez sería desde cada conjunto de indicadores a lo largo de los años. El mapa de densidad lo podemos observar en la Figura 10.

Figura 10: Indicadores registrados por conjunto en Colombia



Se observa que no hay indicadores registrados representados como un espacio en blanco en el mapa de densidad. Así mismo, es importante detallar que los conjuntos de indicadores con mayor densidad de registros a lo largo de los años para Colombia es el de Economía y el de Deuda.

De igual forma, resulta interesante observar que la mayoría de conjuntos de indicadores no poseen más de 50 registros a lo largo de los años, esto se puede interpretar como que en nuestro país el hábito de conservar registros de la mayoría de indicadores se ha mantenido en un plano inferior.

Como parte del trabajo, se ha tenido que definir métodos para tratamiento de los datos ante esa falta de información y serán expuestos a continuación por medio del flujo de trabajo.

5.2 Flujo de Trabajo

Para llevar a cabo la tarea de analizar el Dataset y el comportamiento de los Indicadores de Desarrollo Mundial se diseñó un Flujo de Trabajo, el cual se ilustra en la Figura 11. De igual manera, existe una convención de cada forma implementada en el flujo de trabajo, y esta se detalla en la Tabla 4.

Tabla 4: Convención para el flujo de trabajo

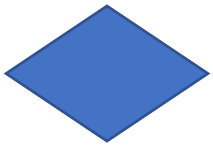
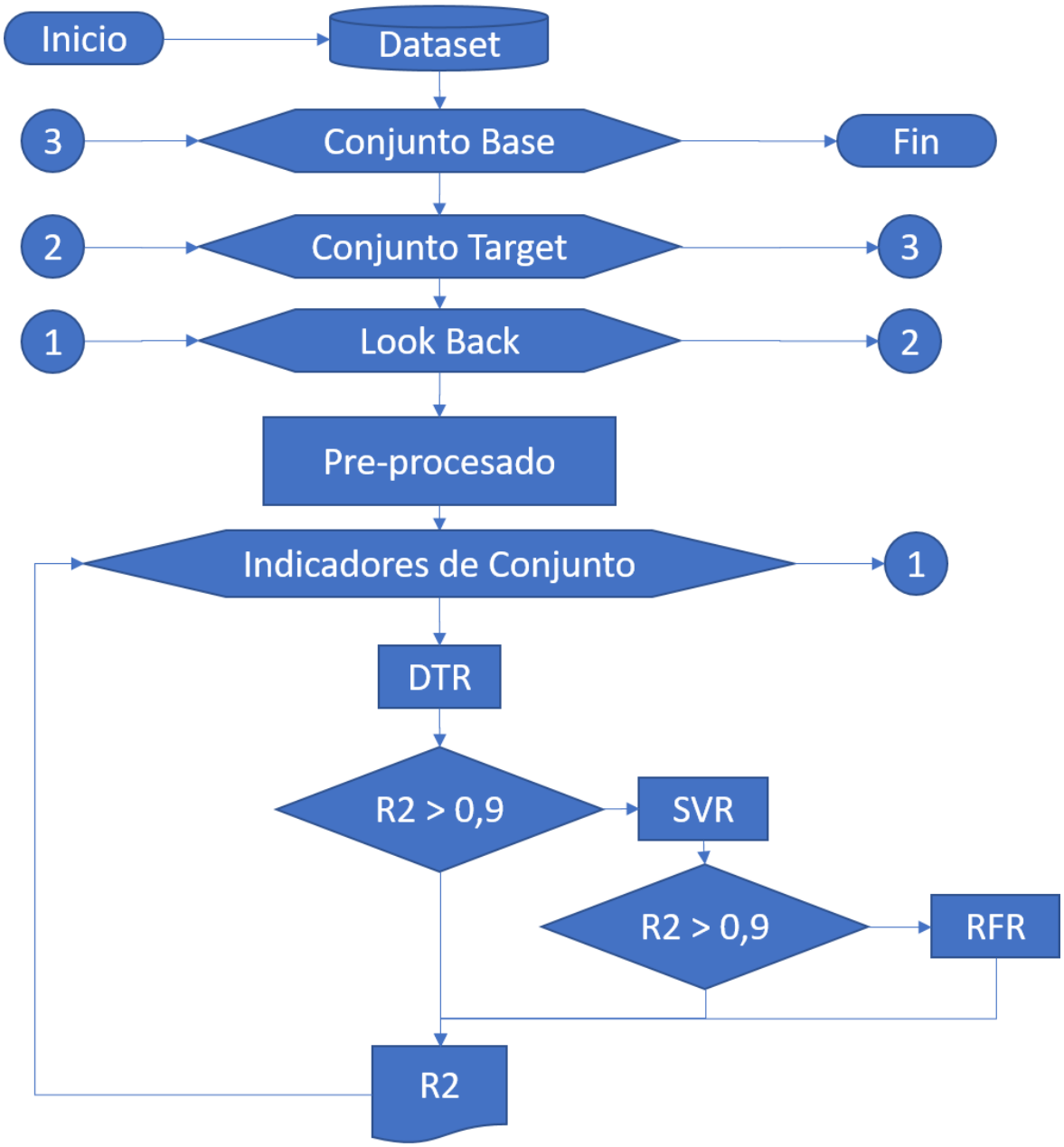
Símbolo	Nombre	Significado
	Terminador	Indica el inicio o la terminación.
	Decisión	Indica un punto en el flujo en que se produce una bifurcación del tipo “SI – NO”.
	Actividad	Representa una actividad llevada a cabo en el proceso.
	Base de Datos	Empleado para representar una base de datos.
	Iteración	Indica que una instrucción se debe ejecutar varias veces.
	Documento	Se utiliza para representar la salida de datos.
	Conector	Sirve para enlazar dos partes cualesquiera del diagrama.

Figura 11: Flujo de trabajo detallado



5.2.1 Conjunto Base, Target y Look Back. Tal como lo indica el Flujo de Trabajo, al iniciar se accede a la base de datos proveída por Kaggle, luego se tienen varios iteradores, descritos a continuación:

- **Conjunto Base:** son los indicadores de desarrollo utilizados para predecir cada indicador de un conjunto Target.
- **Conjunto Target:** son los indicadores de desarrollo que se van a predecir en base a un conjunto Base.
- **Look Back:** es el número de años tomados del conjunto Base para predecir cada indicador del conjunto Target. Por ejemplo, con un Look Back de 3 a partir del año 2012, se utilizarían los registros del año 2012, 2011 y 2010 del conjunto Base para predecir los registros del año 2012 de un indicador del conjunto Target.
- **Indicadores de Conjunto:** son los indicadores de un conjunto Target.

Lo que se hace es iterar cada conjunto, Base y Target, utilizando los indicadores de un conjunto Base para predecir cada indicador de un conjunto Target, y determinando para cada conjunto la media del Coeficiente de Determinación (R^2). Si el R^2 es negativo al momento de predecir un indicador, entonces al momento de calcular la media del conjunto se toma este valor negativo como 0.

Ahora bien, dependiendo del Look Back, el total de iteraciones aumentará, es decir, si se utiliza un Look Back de 1, se estará utilizando los registros del año anterior. Para calcular el total de iteraciones basta con resolver la siguiente ecuación:

$$Total\ Iteraciones = C_{Base} * C_{Target} * Look\ Back * k$$

Donde,

C_{Base} es la cantidad de Conjuntos Base

C_{Target} es la cantidad de Conjuntos Target

k es el número de indicadores de un Conjunto Target

Así mismo es importante tener en cuenta el fenómeno anteriormente mencionado de la repetición de algunos indicadores en varios conjuntos. Lo que se hace es quitar los indicadores del Conjunto Target que están repetidos en el Conjunto Base.

5.2.2 Preprocesado de Datos. Para predecir cada indicador en un Conjunto de indicadores se estructuran los datos como se muestra en la Tabla 5.

Tabla 5. Estructura aplicada a los datos

País	**Indicador 1,2,3…n**	**Indicador 1,2,3…n**	**Indicador 1,2,3…n**	**…**	**Indicador 1,2,3 … n**	**Y Indicador**
A	Año inicial	Año inicial-1	Año inicial-1	…	Año inicial - Look back	Año inicial
A	Año inicial-1	Año inicial-2	Año inicial-2	…	Año inicial -1 - Look back	Año inicial - 1
A	Año inicial-2	Año inicial-3	Año inicial-3	…	Año inicial -2 - Look back	Año inicial - 2
…	…	…	…	…	…	…
A	Año inicial– Rango Año	Año inicial– Rango Año-1	Año inicial– Rango Año-1	…	Año inicial - Rango Año - Look back	Año inicial - Rango Año
B	Año inicial	Año inicial-1	Año inicial-1	…	Año inicial - Look back	Año inicial
B	Año inicial-1	Año inicial-2	Año inicial-2	…	Año inicial -1 - Look back	Año inicial – 1
B	Año inicial-2	Año inicial-3	Año inicial-3	…	Año inicial-2 - Look back	Año inicial – 2
…	…	…	…	…	…	…
B	Año inicial – Rango Año	Año inicial - Rango Año-1	Año inicial - Rango Año - 1	…	Año inicial - Rango Año - Look back	Año inicial - Rango Año

Donde *Look Back* es el rango de años antes de cada año de un *Rango Año* determinado, el *Rango Año* es el rango de años antes del *Año inicial* y el *Y indicador* es el indicador que se quiere predecir.

A continuación, un ejemplo en la Tabla 6.

- Año inicial = 2012
- Rango Año = 5
- Look Back = 3

Tabla 6. Ejemplo estructura aplicada a los datos

	GDP[4]	GNP[5]	GNI[6]	GDP	GNP	GNI	GDP	GNP	GNI	Rural
Colombia	2012	2012	2012	2011	2011	2011	2010	2010	2010	2012
Colombia	2011	2011	2011	2010	2010	2010	2009	2009	2009	2011
Colombia	2010	2010	2010	2009	2009	2009	2008	2008	2008	2010
Colombia	2009	2009	2009	2008	2008	2008	2007	2007	2007	2009
Colombia	2008	2008	2008	2007	2007	2007	2006	2006	2006	2008
Chile	2012	2012	2012	2011	2011	2011	2010	2010	2010	2012
Chile	2011	2011	2011	2010	2010	2010	2009	2009	2009	2011
Chile	2010	2010	2010	2009	2009	2009	2008	2008	2008	2010
Chile	2009	2009	2009	2008	2008	2008	2007	2007	2007	2009
Chile	2008	2008	2008	2007	2007	2007	2006	2006	2006	2008

Es importante especificar que cada año representa el valor del indicador.

Una vez aplicada la estructura de la Tabla 5. se Preprocesan los datos así:

1. Eliminar los indicadores que tienen todos los valores como *NaN* [7], por lo que lo que se hace aquí es eliminar todos los indicadores que no tienen registro para todos los países y rango de años deseado.
2. Imputar los valores *NaN* con el promedio de cada indicador, excepto el indicador Y.
3. Eliminar las filas donde el valor del indicador Y es *NaN.*

[4] GPD: Siglas en inglés para Gross Domestic Product, es decir, Producto Interno Bruto.
[5] GNP: Siglas en inglés para Gross Nactional Product, es decir, Producto Nacional Bruto.
[6] GNI: Siglas en inglés para Gross National Income, es decir, Ingreso Nacional Bruto.
[7] NaN: Siglas en inglés para Not a Number la cual es una expresión para los datos que faltan.

4. Eliminar los indicadores altamente correlacionados (coeficiente de correlación mayor a 0.7) con el indicador Y, evitando redundancia debido a que en ocasiones se repiten indicadores en diferente escala.
5. Normalizar los datos con *StandardScaler* de *Scikit-learn* que es una función que estandariza las características de forma que quedan con media cero y varianza uno.
6. Dividir los datos normalizados en: 80% para el Train y 20% para Test, para ello se utiliza la función *Train Test Split* de *Scikit-learn.*

5.2.3 Estimadores. Después de Preprocesar los datos, se prueban los estimadores en esta secuencia: Primero se utiliza un DTR, si la puntuación de $\boldsymbol{R^2}$ es mayor que 0.9 (90%), entonces se continua al siguiente indicador, de lo contrario se prueba con un SVR, y si falla, finalmente se prueba con un RFR. Si ninguno de los estimadores logró un rendimiento superior a 0.9, entonces se almacena el mejor resultado entre todos.

Se seleccionó ese orden específico porque el costo computacional de un DTR es menor comparado con un SVR y un RFR.

6 CONFIGURACIÓN EXPERIMENTAL

Para el análisis de series temporales se deben utilizar los datos a lo largo del tiempo, por lo que se seleccionó un Look Back igual a 3 debido a que, si se selecciona un número mayor, entonces el costo computacional aumentaba drásticamente.

6.1 Suramérica

Se sabe que el Banco Mundial tiene registros desde 1960 hasta 2015, por lo que es posible seleccionar cualquier año que se desee entre ese rango. Para encontrar un buen rango se analizó el número de indicadores registrados para cada uno de los países de América del Sur a lo largo de los años. En base a la Figura 9. se seleccionó el rango de años igual a 5 años a partir del 2008 hasta el 2012 porque se observa una densidad muy alta de indicadores registrados en este rango.

6.2 Colombia

Para Colombia se analizó el número de indicadores registrados para cada conjunto de indicadores. En base a la Figura 10. se seleccionó el rango de años igual a 15 años a partir del año 2000 hasta 2015 porque se encontró que existen más datos registrados en este rango. De igual manera es importante tener en cuenta que se seleccionó un rango mayor de años que Suramérica porque sucedió que después de hacer el Preprocesado con un rango de 5 años la tabla estaba compuesta por menos de 2 filas de registros, y es necesario que los datos que le se le entreguen a los estimadores tengan al menos 2 filas de registros válidos para que estos puedan trabajar, debido a que, si no los tienen no funcionan. Así que para resolver este problema se eliminaron los indicadores que tienen menos de 5 registros en un rango de años igual a 15 años.

6.3 Software y Hardware

Se utilizó Anaconda Continuum Analytics[8] con el lenguaje de programación Python[9] en su versión 2.7 en un entorno de Jupyter[10]. En cuanto a Hardware inicialmente se efectuaron las pruebas con un ordenador de escritorio personal que cuenta con las siguientes características:

- CPU: Intel ® Core ® 2 Duo E7200 @ 2.53 GHz[11]
- RAM: 8 GB
- HD: 1 disk SATA 500 GB

Debido a que el procesador es de tan sólo un núcleo, el tiempo computacional era de varios días, manteniendo encendido el ordenador sin completar los cálculos, por lo que se decidió utilizar la infraestructura de la Supercomputadora GUANE[12], la cual posee las siguientes características:

- CPU: 2x Intel ® Xeon ® CPU E5645 @ 2.40GHz[13]
- RAM: 104 GB
- HD: 1 disk SAS 200GB

Es importante resaltar que la cantidad de iteraciones para el flujo de trabajo y el uso de los estimadores es necesario un procesador de alto rendimiento, como los que posee la Supercomputadora. Se utilizaron dos procesadores de GUANE, lo cual agilizó considerablemente el flujo de trabajo permitiendo llevar a cabo más combinaciones de parámetros.

[8] https://www.continuum.io/what-is-anaconda
[9] https://www.python.org/
[10] http://jupyter.org/
[11] https://ark.intel.com/es/products/35348/Intel-Core2-Duo-Processor-E7200-3M-Cache-2_53-GHz-1066-MHz-FSB
[12] http://www.sc3.uis.edu.co/servicios/hardware/
[13] https://ark.intel.com/es/products/48768/Intel-Xeon-Processor-E5645-12M-Cache-2_40-GHz-5_86-GTs-Intel-QPI

7 RESULTADOS

7.1 Suramérica

Utilizando en conjunto todos los países de la región suramericana se detallan en la Tabla 7. las mejores combinaciones.

Tabla 7: Mejores combinaciones para Suramérica

Base	Target	Look Back	# Indicadores	$\overline{R^2}$
Privado	Ciencia	1	11	0.902
Urbano	Ciencia	2	13	0.892
Comercio	Ciencia	2	11	0.880
Comercio	Ciencia	3	11	0.877
Infraestructura	Ciencia	2	13	0.875
Energía	Ciencia	1	13	0.863
Ciencia	Ciencia	3	13	0.856
Privado	Ciencia	2	11	0.856
Urbano	Ciencia	1	13	0.842
Comercio	Ciencia	1	11	0.841

Resulta interesante observar que al combinar Comercio y Ciencia con un Look Back de 2, es decir, utilizando en conjunto los registros de los últimos dos años, da un mejor resultado que si se utilizan sólo los registros del año anterior. El mismo fenómeno sucede con el conjunto de indicadores de Urbano. Sin embargo, en el caso de Privado no es así, pues se observa una mejora considerable al utilizar los registros del año inmediatamente anterior, en vez de utilizar los dos años anteriores.

De igual manera, resulta interesante observar que el Conjunto que mejor resultados otorga como Conjunto Target es el de Ciencia, es decir, Ciencia y Tecnología, con un rendimiento de más del 80% en todos los casos.

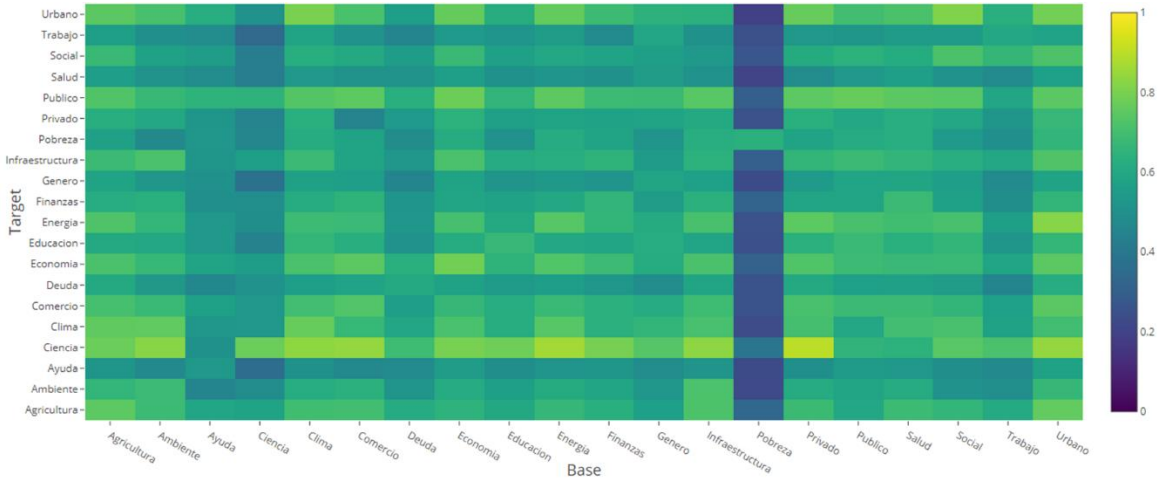
Así mismo, en la Tabla 8. se observan las peores combinaciones para Suramérica. Resulta curioso que sea el conjunto de indicadores de Pobreza el que peor rendimiento haya dado, otorgando una noción de que es el conjunto que peor sirve para predecir cualquier otro conjunto de indicadores.

Tabla 8: Peores combinaciones para Suramérica

Base	**Target**	**Look Back**	**# Indicadores**	$\overline{R^2}$
Pobreza	Urbano	3	21	0.172
Pobreza	Urbano	1	21	0.192
Pobreza	Ayuda	2	69	0.195
Pobreza	Salud	1	152	0.201
Pobreza	Ayuda	3	69	0.208
Pobreza	Urbano	2	21	0.210
Pobreza	Salud	2	152	0.211
Pobreza	Salud	3	152	0.216
Pobreza	Ambiente	1	111	0.224
Pobreza	Energía	3	53	0.226

En ningún caso, se supera un $\overline{R^2}$ mayor a 0.3 (30%), lo cual indica que el rendimiento es bastante bajo. A pesar de ello, es posible detallar una leve mejora de rendimiento al utilizar un Look Back de 1, es decir, el año inmediatamente anterior para los conjuntos de indicadores de Urbano. Con la finalidad de tener una perspectiva más global, en la Figura 12. se ilustra un Mapa de Calor, o Mapa de Densidad, de todas las combinaciones posibles de los Conjuntos de indicadores de desarrollo, donde cada intersección entre un Conjunto Base y un Conjunto Target representa el $\overline{R^2}$.

Figura 12: Mapa de densidad para Suramérica con un Look Back de 1



Resulta muy interesante denotar que el rendimiento de manera global es bastante bueno, el $\overline{R^2}$ se mantiene en la mayoría de combinaciones entre un 0.6 y 0.8, es decir, entre un 60% y un 80% de precisión.

Se observa también, que tal como se expuso en la Tabla 8. el rendimiento del Conjunto de Indicadores de Pobreza como Conjunto Base, resulta bastante bajo, es decir, que no es recomendable tomar este conjunto de indicadores como referencia para predecir los demás conjuntos de indicadores.

La Figura 13. y la Figura 14. son los mapas de densidad para Suramérica al aplicar un Look Back de 2 y 3 respectivamente. Es posible apreciar un leve incremento del rendimiento en general respecto de la Figura 12.

Figura 13: Mapa de densidad para Suramérica con un Look Back de 2

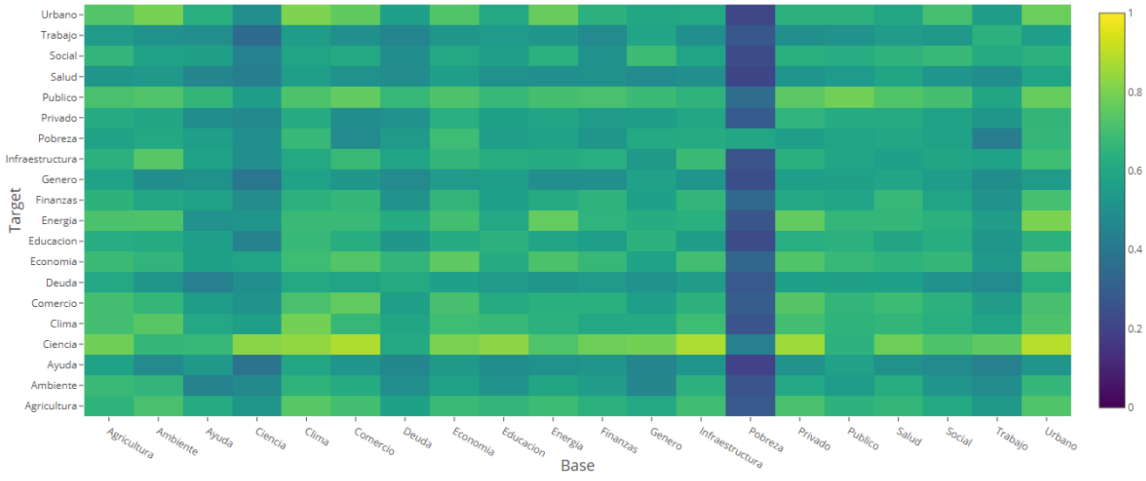
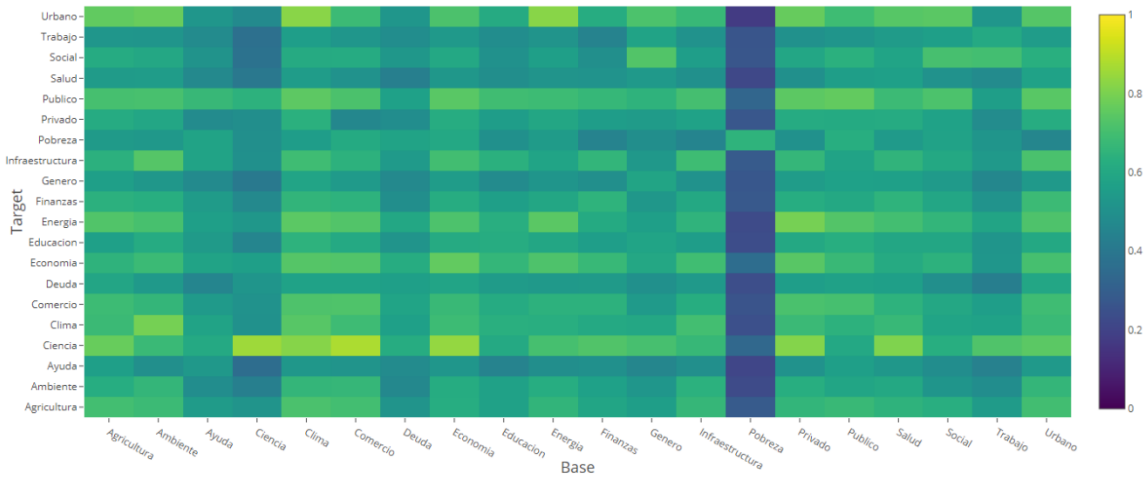


Figura 14: Mapa de densidad para Suramérica con un Look Back de 3



Teniendo en cuenta las observaciones de las Tablas 7 y 8 es posible apreciar que el aplicar un Look Back superior a 3 difícilmente mejorará el rendimiento general.

Con el objetivo de encontrar los conjuntos con mayor grado de predictibilidad, promediamos el rendimiento de cada conjunto como Base y Target.

Tabla 9: Resumen Conjuntos Base para Suramérica

Mejores	$\overline{R^2}$	**Peores**	$\overline{R^2}$
Desarrollo urbano	0.691	Pobreza	0.284
Agricultura	0.658	Ciencia y Tecnología	0.496
Cambio climático	0.656	Eficacia de la ayuda	0.531
Economía y crecimiento	0.656	Deuda externa	0.548
Sector Privado	0.648	Protección social y trabajo	0.554

En la Tabla 9. se observa en orden descendente los cinco mejores y peores conjuntos Base. Claramente el conjunto Base que mejor predice a los demás conjuntos es el de Desarrollo Urbano, y de forma similar le siguen Agricultura, Cambio Climático, Economía y crecimiento, y Sector privado.

Tabla 10: Resumen Conjuntos Target para Suramérica

Mejores	$\overline{R^2}$	**Peores**	$\overline{R^2}$
Ciencia y tecnología	0.747	Eficacia de la ayuda	0.486
Sector público	0.688	Salud	0.499
Desarrollo urbano	0.669	Protección social y trabajo	0.507
Economía y crecimiento	0.660	Género	0.520
Cambio climático	0.645	Deuda externa	0.536

En la Tabla 10. se observa en orden descendente los cinco mejores y peores conjuntos Target. Claramente el conjunto Target más fácil de predecir es el de Ciencia y tecnología, y de forma similar el Sector público, Desarrollo urbano, Economía y crecimiento, y Cambio climático.

Resulta interesante detallar que, para Suramérica, los conjuntos que son al mismo tiempo mejor Base y Target son: Desarrollo Urbano, Cambio Climático y Economía y crecimiento.

7.2 Colombia

Utilizando los registros de Colombia, en la Tabla 11. se detallan las mejores combinaciones.

Tabla 11: Mejores combinaciones para Colombia

Base	Target	Look Back	# Indicadores	$\overline{R^2}$
Social	Urbano	2	24	0.722
Social	Urbano	3	24	0.693
Urbano	Urbano	1	24	0.669
Energía	Urbano	2	20	0.662
Urbano	Salud	1	149	0.645
Clima	Urbano	2	16	0.643
Social	Salud	2	145	0.641
Salud	Urbano	2	21	0.639
Salud	Salud	1	152	0.638
Economía	Urbano	2	24	0.637

Resulta bastante interesante observar que el Conjunto de indicadores que mejor rendimiento da como Conjunto Target es el de Urbano, es decir, Desarrollo Urbano con un rendimiento $\overline{R^2}$ aproximado de 0.6 (60%).

De igual forma, es posible observar que el utilizar un Look Back de 2, es decir, los últimos dos años anteriores al año que se desea predecir, resulta en la mayoría de las combinaciones en un mejor rendimiento antes que si se utilizaran los registros de los últimos tres años, es decir, un Look Back de 3.

Así mismo, el conjunto de Indicadores de la Salud resulta tener un excelente rendimiento bien sea como Conjunto Base, o como Conjunto Target, en ambos casos su rendimiento es bastante alto.

En contraparte, en la Tabla 12. se observa observan las peores combinaciones para Colombia.

Tabla 12: Peores combinaciones para Colombia

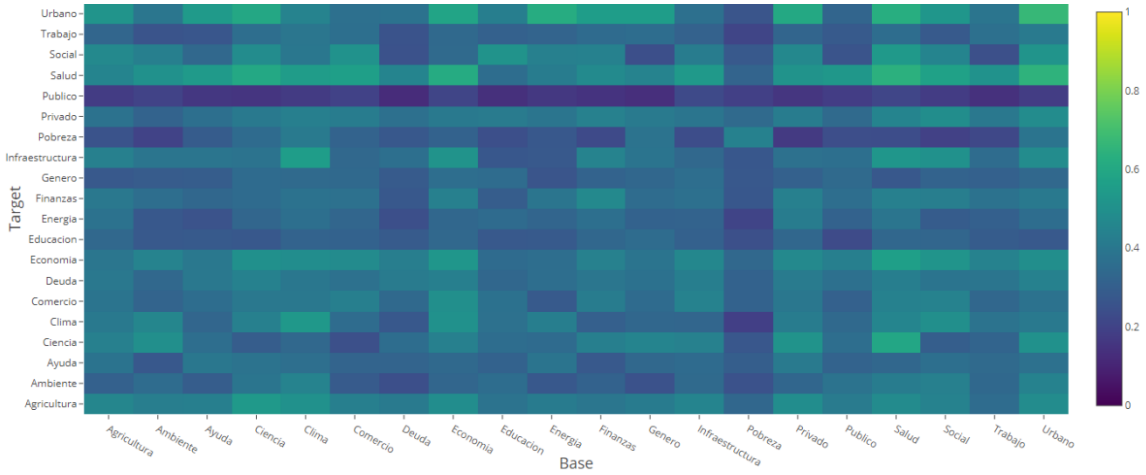
Base	Target	Look Back	# Indicadores	$\overline{R^2}$
Agricultura	Público	3	97	0.107
Finanzas	Público	2	95	0.112
Infraestructura	Público	2	97	0.113
Infraestructura	Público	3	97	0.117
Trabajo	Público	2	97	0.118
Ayuda	Público	3	97	0.121
Deuda	Público	1	97	0.123
Energía	Público	2	97	0.128
Publico	Público	3	97	0.130
Género	Público	1	97	0.131

Aquí el rendimiento más bajo resulta de utilizar los indicadores del Sector Público como Conjunto Target. El rendimiento $\overline{R^2}$ es realmente deplorable, ni siquiera alcanza el 0,15 (15%). Incluso aplicando varios Look Back de 2 y 3 el rendimiento no aumenta significativamente para ninguna combinación.

Resulta curioso observar que los peores resultados para Colombia se den al combinar los indicadores de Infraestructura y Sector público.

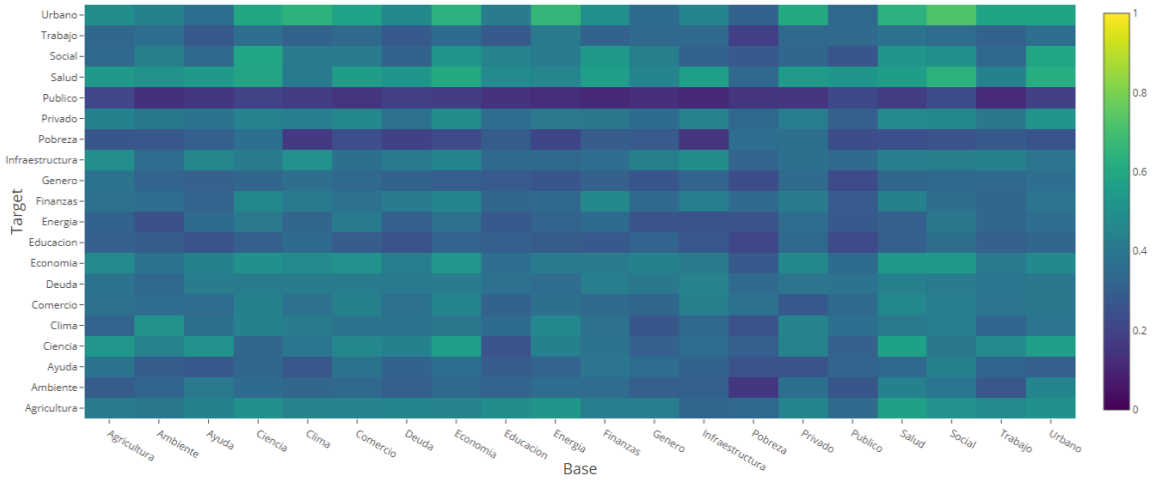
De forma similar, buscando tener una perspectiva más global, se ilustran las combinaciones de todos los conjuntos en un Mapa de Calor o Mapa de Densidad en la Figura 15. El rendimiento en general es claramente bastante bueno, el $\overline{R^2}$ se mantiene en la mayoría de combinaciones entre un 0.5 y 0.6, es decir, entre un 50% y un 60% de precisión.

Figura 15: Mapa de densidad para Colombia con un Look Back de 1



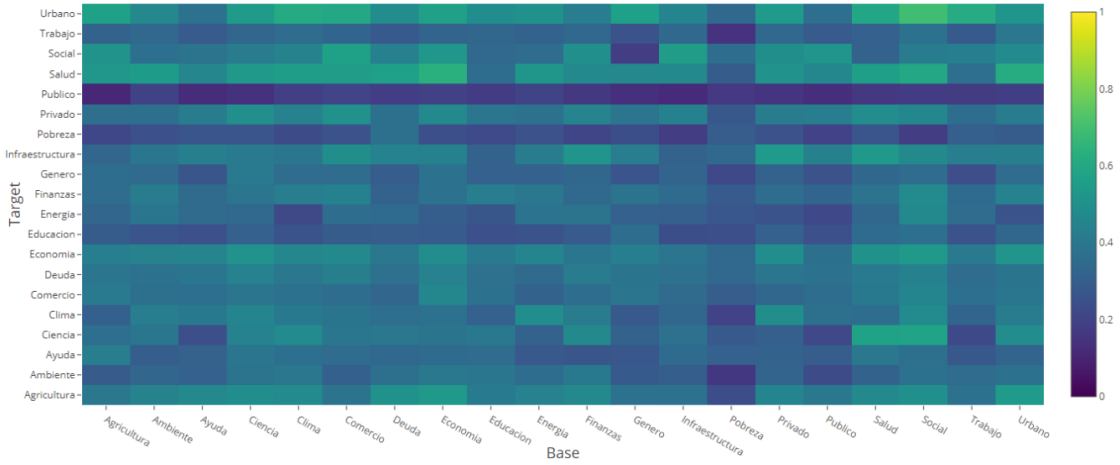
Como se mencionó anteriormente, es claramente visible que el Conjunto Target que peor resultado da es el Conjunto de indicadores del Sector Público.

Figura 16: Mapa de densidad para Colombia con un Look Back de 2



La Figura 16. y la Figura 17. son los mapas de densidad para Colombia al aplicar un Look Back de 2 y 3 respectivamente. Es posible apreciar un leve incremento del rendimiento en general respecto de la Figura 15.

Figura 17: Mapa de densidad para Colombia con un Look Back de 3



Teniendo en cuenta las observaciones de las Tablas 11 y 12 el aplicar un Look Back de más de 3 difícilmente mejorará el rendimiento general.

Como se mencionó anteriormente, promediamos el rendimiento de cada conjunto como Base y Target y analizamos los cinco mejores y peores conjuntos para Colombia.

Tabla 13: Resumen Conjuntos Base para Colombia

Mejores	$\overline{R^2}$	**Peores**	$\overline{R^2}$
Salud	0.434	Pobreza	0.279
Desarrollo urbano	0.431	Deuda externa	0.316
Economía y crecimiento	0.415	Sector público	0.333
Cambio climático	0.409	Educación	0.337
Ciencia y tecnología	0.399	Protección social y trabajo	0.338

En la Tabla 13. se observa claramente que el conjunto Base que mejor predice a los demás conjuntos es el de Salud, y de forma similar Desarrollo Urbano, Economía y crecimiento, Cambio climático y Ciencia y tecnología.

Tabla 14: Resumen Conjuntos Target para Colombia

Mejores	$\overline{R^2}$	**Peores**	$\overline{R^2}$
Salud	0.510	Sector público	0.170
Desarrollo urbano	0.486	Pobreza	0.281
Economía y crecimiento	0.443	Educación	0.298
Agricultura	0.434	Género	0.315
Sector privado	0.401	Energía y minería	0.321

En la Tabla 14. se observa claramente que el conjunto Target que es más fácil de predecir es el de la Salud y de forma similar el Desarrollo urbano, Economía y crecimiento, Agricultura y Sector privado.

Resulta interesante detallar que, para Colombia, los conjuntos que son al mismo tiempo mejor Base y Target son: Salud, Desarrollo Urbano y Economía y crecimiento.

7.3 Tiempo Computacional

Para computar todas las iteraciones posibles a la Supercomputadora le tomó:

- Para Suramérica 38815.64 segundos, aproximadamente 11 horas
- Para Colombia 28783.77 segundos, aproximadamente 8 horas

7.4 Rendimiento Original vs GridSearchCV

Buscando mejorar el rendimiento de los estimadores se decidió probar una búsqueda de parámetros exhaustiva por medio de un *GridSearchCV*.

Luego de aproximadamente un mes de estar efectuando cálculos se completó la búsqueda exhaustiva, y se decidió comparar el rendimiento de dos métodos de asignación paramétrica: Original y *GridSearchCV*. Original significa que los estimadores utilizan los parámetros que les asigna *Scikit-learn* por defecto y *GridSearchCV* significa que los estimadores probaron todos los parámetros posibles en un rango predefinido buscando la mejor asignación paramétrica.

Para efectuar dicha comparación se calcula el porcentaje de mejora y se promedia el rendimiento para cada caso.

En la Figura 18. es posible apreciar que si se utiliza un *GridSearchCV* el promedio de mejora del Coeficiente de Determinación $\overline{R^2}$ será máximo de un 16%, en lugar de si se utilizan los parámetros por defecto.

Figura 18: Mejora de rendimiento promedio del $\overline{R^2}$ al aplicar GridSearchCV vs Original

Ahora bien, en lo que respecta al tiempo utilizado, en la Figura 19. se observa que si se utiliza un *GridSearhCV* significa que el rendimiento promedio del tiempo se incrementa aproximadamente en más de un 3500%.

Figura 19: Mejora de rendimiento promedio del Tiempo al aplicar GridSearchCV vs Original

8 CONCLUSIONES

- Aplicar Machine Learning para el análisis de Indicadores de Desarrollo Mundial permitió encontrar correlaciones no triviales entre los diferentes conjuntos.
- Caracterizar el registro histórico de indicadores de desarrollo del Banco Mundial permitió analizar la volumetría y densidad de los datos, lo cual consecuentemente permitió desarrollar el flujo de trabajo.
- Establecer un flujo de trabajo adecuado permite realizar un correcto tratamiento de los datos, el propuesto cumplió satisfactoriamente su función.
- Los modelos predictivos construidos tuvieron un costo computacional apropiado para el flujo de trabajo teniendo en cuenta la cantidad de indicadores y la volumetría de los datos.
- Promediar el Coeficiente de Determinación para cada conjunto de indicadores permitió evaluar la confiabilidad estadística de los métodos desarrollados.
- En base la Tabla 9. los conjuntos de indicadores Base que mejor predicen a los demás conjuntos en Suramérica son: Desarrollo Urbano, Agricultura, Cambio climático, Economía y crecimiento y Sector Privado.
- En base la Tabla 10. los conjuntos de indicadores Target que son más fácilmente predecibles para Suramérica son: Ciencia y tecnología, Sector público, Desarrollo urbano, Economía y crecimiento y Cambio climático.
- En base la Tabla 13. los conjuntos de indicadores Base para Colombia son: Salud, Desarrollo Urbano, Economía y crecimiento, Cambio climático y Ciencia y tecnología
- En base la Tabla 14. los conjuntos de indicadores Target para Colombia son: Salud, Desarrollo urbano, Economía y crecimiento, Agricultura y Sector privado.

9. RECOMENDACIONES

- Realizar este estudio cada año logrando así validar el comportamiento de los diversos conjuntos de indicadores a lo largo de los años.
- Priorizar, en la medida de lo posible, el uso de los mejores conjuntos Base en la toma de decisiones que influyan en el desarrollo de un país.
- Desarrollar políticas de Inversión en los mejores conjuntos Base, ya que estos son el Pilar Angular desde donde se puede predecir muy bien cualquier otro conjunto de indicadores.
- Efectuar la combinatoria de subconjuntos de Indicadores, debido a que, si se tiene en cuenta el nombre de cada indicador, entonces es posible dividir cada conjunto en subconjuntos, aunque la cantidad de iteraciones aumentaría drásticamente, siendo necesario más tiempo y poder computacional.
- Realizar este mismo estudio y análisis para otros continentes.
- Analizar exógenamente con otros continentes, es decir, agregar países de otros continentes a los datos de Colombia y observar resultados.
- Paralelizar, siempre que sea posible, los núcleos de los procesadores en los estimadores asignando al parámetro *n_jobs* el valor de -1.

REFERENCIAS BIBLIOGRÁFICAS

[1] World Bank Group, “Worldwide Governance Indicators | Data.” [En línea]. (Recuperado 13 de Junio de-2017) Disponible en http://data.worldbank.org/data-catalog/worldwide-governance-indicators

[2] World Bank Group, “Indicators,” *World Bank*. [En línea]. (Recuperado 13 de Junio de 2017) Disponible en http://data.worldbank.org/indicator?tab=all

[3] A. General, L. A. General, and A. General, “Asamblea General,” vol. 13689, 2015.

[4] Wikipedia contributors, “Kaggle,” *Wikipedia, The Free Encyclopedia.* [En línea]. (Recuperado 14 de Junio de 2017) Disponible en https://en.wikipedia.org/w/index.php?title=Kaggle&oldid=790433588

[5] T. M. Learning, “What is Machine Learning ? Examples of Machine Learning Problems Goals of Machine Learning Research,” pp. 1–6, 2008.

[6] ROSSANT, Cyrille. “IPython Interactive Computing and Visualization Cookbook,” J. Dharmaraj, D. Nambiar, and K. Narayanan, Eds. Packt Publishing, 2014, pp. 268–270.

[7] ALPAYDIN, Ehem. “Introduction to Machine Learning Second Edition,” 2 edition., London, England: MIT Press, 2010, pp. 188–193.

[8] Wikipedia contributors, “Support vector machine,” *Wikipedia, The Free Encyclopedia.*, [En línea]. (Recuperado 13 de Junio de 2017) Disponible en https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=784160753.

[9] UNPINGCO, José. “Python for Probability , Statistics , and Machine Learning,” Springer, 2016, pp. 250–251.

[10] Scikit-learn developers, “Support Vector Regression,” *scikit-learn 0.18.1 documentation*, 2017. [En línea]. (Recuperado 14 de Junio de 2017) Disponible en http://scikit-learn.org/stable/modules/svm.html#regression

[11] ROSSANT, Cyrille. “IPython Interactive Computing and Visualization Cookbook,” 2014, pp. 268–270.

[12] MURPHY, Kevin. “A Probabilistic Perspective,” 2012, pp. 550–551.

[13] JAMES, Gareth, *et al*. “An Introduction to Statistical Learning,” Springer, 2013, p. 312.

[14] Wikipedia contributors, “Cross-validation (statistics),” *Wikipedia, The Free Encyclopedia.*, 2017. [En línea]. (Recuperado 14 de Junio de 2017) Disponible en https://en.wikipedia.org/w/index.php?title=Cross-validation_(statistics)&oldid=784123568

[15] Scikit-learn developers, “Random permutations cross-validation a.k.a. Shuffle & Split.” [En línea]. (Recuperado 14 de Junio de 2017) Disponible en http://scikit-learn.org/stable/modules/cross_validation.html#random-permutations-cross-validation-a-k-a-shuffle-split

[16] Scikit-learn developers, “Tuning the hyper-parameters of an estimator.” [En línea]. (Recuperado 14 de Junio de 2017) Disponible en http://scikit-learn.org/stable/modules/grid_search.html#grid-search

[17] Scikit-learn developers, “Model evaluation: quantifying the quality of predictions,” *scikit-learn 0.18.1 documentation*, 2017. [En línea]. (Recuperado 14 de Junio de 2017) Disponible en http://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

[18] LAZAREVIĆ, Luka; KOVAČEVIĆ, Miloš y POPOVIĆ, Zdenka. “Rail Traffic Volume Estimation,” vol. 13, pp. 133–141, 2015.

[19] MA, Haibo; CHANG, Wenjuan y CUI, Guangbai. “Ecological Footprint Model Using the Support Vector Machine Technique,” *PLoS One*, vol. 7, no. 1, pp. 1–5, 2012.

[20] LAI, Dejian, “Principal Component Analysis on Human Development Indicators of China,” *Soc. Indic. Res.*, vol. 61, no. 3, pp. 319–330, Mar. 2003.

[21] HICKS, Norman y STREETEN, Paul. “Indicators of development: The search for a basic needs yardstick,” *World Dev.*, vol. 7, no. 6, pp. 567–580, 1979.

[22] CERON, Jean Paul y DUBOIS Ghislain. “Tourism and Sustainable Development Indicators: The Gap between Theoretical Demands and Practical Achievements,” *Curr. Issues Tour.*, vol. 6, no. 1, pp. 54–75, 2003.

BIBLIOGRAFÍA

- ALPAYDIN, Ehem. “Introduction to Machine Learning Second Edition,” 2 edition., London, England: MIT Press, 2010, pp. 188–193.
- CERON, Jean Paul y DUBOIS Ghislain. “Tourism and Sustainable Development Indicators: The Gap between Theoretical Demands and Practical Achievements,” Curr. Issues Tour., vol. 6, no. 1, pp. 54–75, 2003.
- HICKS, Norman y STREETEN, Paul. “Indicators of development: The search for a basic needs yardstick,” World Dev., vol. 7, no. 6, pp. 567–580, 1979.
- JAMES, Gareth, et al. “An Introduction to Statistical Learning,” Springer, 2013, p. 312.
- LAI, Dejian, “Principal Component Analysis on Human Development Indicators of China,” Soc. Indic. Res., vol. 61, no. 3, pp. 319–330, Mar. 2003.
- LAZAREVIĆ, Luka; KOVAČEVIĆ, Miloš y POPOVIĆ, Zdenka. “Rail Traffic Volume Estimation,” vol. 13, pp. 133–141, 2015.
- MA, Haibo; CHANG, Wenjuan y CUI, Guangbai. “Ecological Footprint Model Using the Support Vector Machine Technique,” PLoS One, vol. 7, no. 1, pp. 1–5, 2012.
- MURPHY, Kevin. “A Probabilistic Perspective,” 2012, pp. 550–551.
- ROSSANT, Cyrille. “IPython Interactive Computing and Visualization Cookbook,” J. Dharmaraj, D. Nambiar, and K. Narayanan, Eds. Packt Publishing, 2014, pp. 268–270.
- UNPINGCO, José. “Python for Probability, Statistics, and Machine Learning,” Springer, 2016, pp. 250–251.