APLICACIÓN DE UN ALGORITMO DE OPTIMIZACION SECUENCIAL MINIMA (SMOreg)

Cristian Duney Bermudez Quintero, Jhon Heyder Murillo Mejia, Hernan Dario Jimenez Arboleda.
Universidad cooperativa de Colombia- Bogotá, Colombia

Email: cristian.bermudez@campusucc.edu.co, jhon.murillom@campusucc.edu.co. hernan.jimeneza@campusucc.edu.co
Grupo Ingeniería De Telecomunicaciones

Resumen: *En este trabajo se realiza la aplicación de un modelo predictivo basado en el algoritmo de regresión SMOreg, que consiste en la modelación del comportamiento de datos utilizando la regresión estadística. El modelo fue implementado sobre registros de nacimientos y defunciones en Colombia, lo cual nos permitió predecir comportamientos futuros del crecimiento de la población, direccionados a mejorar la toma de decisiones en políticas sanitarias y sociales en el país.*
Nuestra principal fuente de información fue el Departamento Administrativo Nacional de Estadística (DANE), de donde obtuvimos la información anual de los nacimientos y defunciones en Colombia organizada demográficamente. Posteriormente, se realizó la preparación de los datos clasificados por departamento para administrarlos en el modelo estadístico. Una vez obtenidos los datos, se emplearon modelos predictivos a partir de Tecnologías de Información (TI) como los Métodos Supervisados Automatizados. Adicionalmente, se utilizó la herramienta de minaría de datos Weka para aplicar el algoritmo SMOreg sobre los datos obtenidos.
El desarrollo del presente proyecto y sus resultados servirán como insumo para mejorar una de las problemáticas que afrontan las estadísticas vitales; el sub registro a partir de la implementación de herramientas TI. La aplicación de la inteligencia de negocios, tableros de control y cubos dinámicos permitirán generar alertas tempranas que permitan obtener datos confiables para la toma oportuna de decisiones.

Palabras claves:
Regresión, Modelos Estadísticos, Similitud, Aprendizaje Automatizado, Series de Tiempo.

Abstract: *In this work, the application of a predictive model based on the SMOreg regression algorithm, which consists of modeling the behavior of data using statistical regression, is performed. The model was implemented on records of births and deaths in Colombia, which allowed us to predict future behaviors of population growth, aimed at improving decision-making in health and social policies in the country.*
Our main source of information was the National Administrative Department of Statistics (DANE), from where we obtained the annual information on births and deaths in demographically organized Colombia. Subsequently, the preparation of the data classified by department was carried out to administer them in the statistical model. Once the data were obtained, predictive models were used from Information Technology (IT) as Automated Supervised Methods. Additionally, the Weka data mining tool was used to apply the SMOreg algorithm to the data obtained.
The development of this project and its results will serve as an input to improve one of the problems faced by vital statistics; the underreporting from the implementation of IT tools. The application of business intelligence, control panels and dynamic cubes will generate early warnings to obtain reliable data for timely decision making.

Keywords:
Regression, Statistical Models Similarity. Automated Learning, Time Series.

I. INTRODUCCIÓN

Por medio del uso de algoritmos de Machine Learning o aprendizaje automático, realizaremos el acondicionamiento y prueba de los datos recopilados de las series históricas proporcionadas por el DANE, con el propósito de aplicar un modelo predictivo que nos permita evaluar el comportamiento de los nacimientos y las defunciones en los diferentes departamentos del país.
Para el desarrollo de este proyecto se requirió de una exploración y análisis bibliográfico de modelos estadísticos principalmente del algoritmo de regresión SMOreg, así como el estudio y la capacitación en el manejo de la herramienta de uso libre para minería de datos llamado Weka, creado por la universidad de Waikato, Nueva Zelanda para el modelamiento de datos y análisis predictivo. Esta herramienta se integra al plugin forecasting para la gestión de datos e inteligencia de negocios. Adicionalmente, estas herramientas nos ayudan a desarrollar modelos de predicción y automatizar resultados. Las bases de datos con las cuales se trabajó fueron recopiladas de la página web del DANE datadas desde el año 2000 al año 2016, esto con el fin de poder realizar predicciones y generar resultados con el menor porcentaje de error.

II. TRABAJOS RELACIONADOS

Omar González Amor Santa Clara, Cuba (2015), Titulo: Predicción de parámetros de energía eólica utilizando modelos de regresión - más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta. El objetivo general de este trabajo es desarrollar un sistema capaz de predecir el comportamiento de los parámetros de la energía eólica contenidos en series de tiempo utilizando modelos de regresión [1]

Yvonne Gala García, Madrid, 25 de septiembre de 2013, Titulo: Algoritmos SVM para problemas sobre big data – estudiar el estado del arte de las máquinas de vectores soporte tanto en Clasificación como en regresión. Para ello se analiza la teoría básica de optimización y la teoría clásica de SVM, posteriormente se procede a profundizar en tres de los algoritmos más importantes, Sequential Minimal Algorithm, Dual Coordinate Descend Method, y Stochastic sub-gradient descent Cuyo objetivo es mejorar las predicciones dadas por el European Centre for Medium-Range Weather Forecasts (ECMWF).se justifica por el creciente interés tanto en la predicción de energías renovables, siendo las energías del futuro, la manera de trabajarlo es comparar tres enfoques algorítmicos diferentes.[2]

Carlos Blanco González, Fernando García Diez, Madrid España Título: Minería de Datos: Predicción de las condiciones meteorológicas - El objetivo es obtener un sistema de predicción meteorológica para la ciudad de Madrid, para lo cual se usan técnicas de minería de datos sobre un conjunto extenso de datos extraídos durante aproximadamente 8 años en el aeropuerto de Barajas. La minería de datos proporciona toda una serie de técnicas para la selección, procesado, clasificación, evaluación e interpretación de la información, para poder conseguir que, dado un gran volumen de datos se consiga extraer un conocimiento y así poder predecir futuros comportamientos o acciones. Mediante las diferentes etapas en que se divide el proceso de minería de datos (pre procesado, selección de características, extracción de conocimiento, interpretación y evaluación) conseguimos pasar de una secuencia de datos a un modelo de conocimiento. El cual es apoyado mediante la herramienta WEKA en la que se pueden desarrollar e implementar diferentes etapas de las ya mencionadas en la minería de datos. [3]

Pablo Casas Muñoz, Roberto García Sánchez, Madrid España Título: Predicción meteorológica usando WEKA – desarrolla un sistema real de predicción meteorológica para la ciudad de Madrid, utilizando técnicas de aprendizaje automático para obtener los modelos de caracterización y predicción, empleando como herramienta la plataforma Weka. El punto de partida es la información METAR (Meteorological Actual Report) disponible en Internet. El código METAR, que fue establecido por la World Meteorological Organization (WMO) y adoptado por todas las naciones del mundo, sirve para representar

condiciones meteorológicas en aeropuertos o estaciones meteorológicas, habitualmente con medidas horarias. El código METAR es preciso y fácil de leer y proporciona información muy valiosa que utilizan líneas aéreas y organizaciones gubernamentales de todo el mundo. Para esta práctica se proporciona un conjunto de datos meteorológicos sobre el aeropuerto de Madrid Barajas (código internacional, LEMD) desde el año 1996 al 2004. [4]

III. OBJETIVOS

OBJETIVO GENERAL

Aplicar un algoritmo de optimización secuencial mínima para regresión (SMOreg) que permita la caracterización de las estadísticas de nacimientos y defunciones, con el fin de predecir el comportamiento en los años futuros y de esta manera tener un plan de acción que pueda mejorar la toma de decisiones en políticas sanitarias y sociales en el país.

OBJETIVOS ESPECIFICOS

•Identificar las variables que alimentaran el modelo estadístico de regresión lineal SMOreg.

•Analizar el Algoritmo de Optimización Secuencial Mínima basado en regresión (SMOreg).

•Predecir el comportamiento de los nacimientos y las defunciones en Colombia con el fin de evaluar su comportamiento futuro para los próximos 5 años.

IV. DESARROLLO

FUENTES DE INFORMACION

Nuestras principales fuentes de información fueron las tablas suministradas por el profesor Cristhian Bermúdez, complementando con las descargadas por medio de la página del DANE en el Archivo Nacional de Datos (ANDA) para los nacimientos y defunciones de los años 2000 al 2016 los cuales se encuentran en su página oficial www.dane.gov.co

HERRAMIENTAS PARA EL PROCESAMIENTO DE LOS DATOS

Para la preparación de los datos se utilizó el paquete estadístico SAS esta herramienta se utilizó para construir las tablas por departamento, mes, año de los nacimientos y las defunciones que alimentaran el modelo estadístico
También fue indispensable el uso de la herramienta Excel en la cual se validaban los datos obtenidos en cada paquete y modelo estadístico obtenido en la herramienta SAS, así como las pruebas realizadas con cada una de las fórmulas de SMOreg con el fin de comparar los porcentajes de error obtenidos para cada una de ellas.
Como herramienta tecnológica para la construcción del modelo estadístico se utilizó el software para minería de datos WEKA.

WEKA

Software desarrollado en java por la universidad de Waikato en nueva Zelanda
Proceso de instalación del Weka
Realizamos la respectiva descarga de la página oficial:
https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Instalación Weka

Una vez ejecutado Weka evidenciamos en el siguiente menú

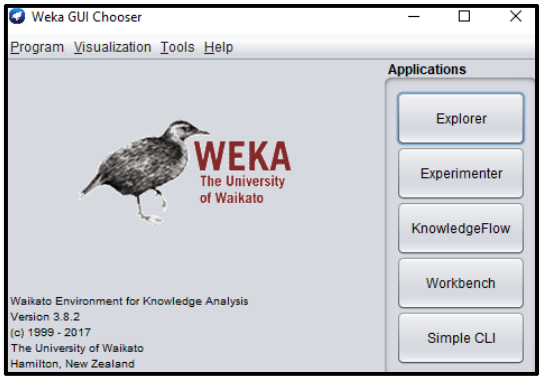


Figura 1. Ventana Principal Software Weka

HERRAMIENTAS PARA EL ANALISIS DE RESULTADOS

Para el análisis de los de los datos se utilizaron herramientas graficas como:

Matriz de relación: esta opción del software Weka permitió evaluar el comportamiento de cada una de las variables que alimentara el modelo.

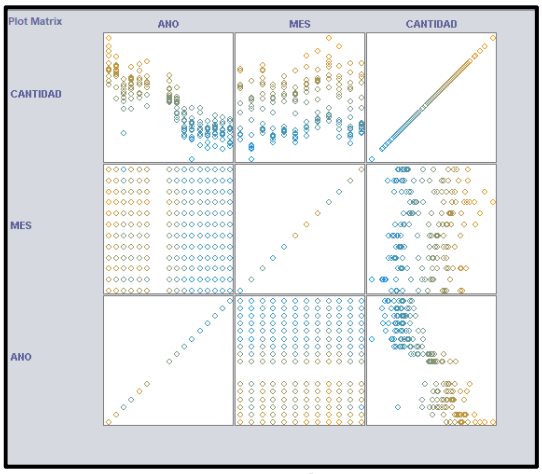


Figura 2. Matriz de relación - variables base de Nacimientos

Dentro del proceso de análisis y técnicas de minería de datos se llevó a cabo el análisis de los datos con el paquete de ofimática Excel, el cual nos permitió la incorporación de fórmulas y representación de datos.

Gráficas de dispersión del Plugin Forecasting: En las cuales se logró evidenciar el comportamiento predictivo del algoritmo.

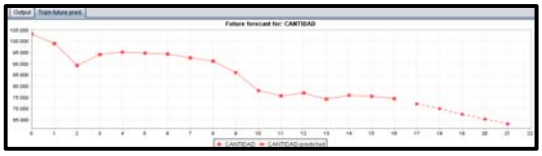


Figura 3. Muestra de Datos Plugin Forecasting

V. METODOLOGÍA

Gráficas de dispersión del Plugin Forecasting: En las cuales se logró evidenciar el comportamiento predictivo del algoritmo.

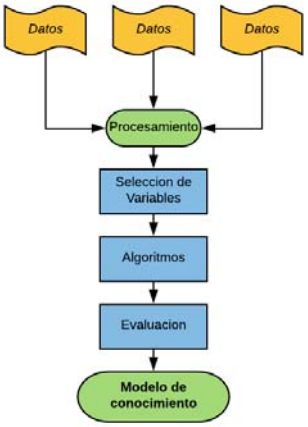


Figura 4. Metodología Minería de Datos

Se realiza el tratamiento de datos de acuerdo a los procesos comunes de minería de datos como lo son procesamiento, selección de variables, algoritmos, evaluación y modelos de conocimiento.

PREPARACION Y PROCESAMIENTO DE LOS DATOS

- **Preparación y alistamiento de los datos**

La fuente oficial de información son los datos de los nacimientos y las defunciones dispuestos por el DANE ubicados en el Archivo Nacional de Datos (ANDA) en su página oficial www.dane.gov.co.
Los datos que se obtuvieron corresponden a series históricas desde el año 2000-2016.

Comprende 3 libros los cuales se encuentra clasificados en Nacimiento, Defunciones Fetales y Defunciones no Fetales
Estos 3 libros contienen las siguientes variables:

- Año
- Mes
- Departamento
- Cantidad de sucesos presentados

- **Características de los archivos**

Tabla 1
Base de datos N1

AÑO	MES	DEPARTAMENTO	CANTIDAD
2000	Enero	Antioquia	8447
2000	Enero	Atlántico	3648
2000	Enero	Bogotá	10688
2000	Enero	Bolívar	2568
2000	Enero	Boyacá	1905
2000	Enero	Caldas	1381
2000	Enero	Caquetá	535
2000	Enero	Cauca	1453

4 atributos (variables)
6726 instancias (rango de datos)

Tabla 2.
Base de datos N2

AÑO	DEPARTAMENTO	CANTIDAD
2000	Antioquia	103153
2000	Atlántico	43926
2000	Bogotá	131170
2000	Bolívar	32474
2000	Boyacá	23070
2000	Total	752834

3 Atributos (variables)
576 instancias (rango de datos)

- **Adaptación de los datos**

Los datos de tipo carácter deben ser transformados en datos de tipo numérico para que puedan ser procesados por el algoritmo SMOreg. Por esta razón se realiza el remplazo de atributos como mes y departamento de la siguiente manera:

Tabla 3.
Conversión de mes a número

MES	MES FORMA NUMERICA
ENERO	1
FEBRERO	2
MARZO	3
ABRIL	4
MAYO	5
JUNIO	6
JULIO	7
AGOSTO	8
SEPTIEMBRE	9
OCTUBRE	10
NOVIEMBRE	11
DICIEMBRE	12

Tabla 4.
Conversión de los departamentos a cód. Numérico

DEPARTAMENTO	COD. DEPART.
Antioquia	5
Atlántico	8
Bogotá, D.C.	11
Bolívar	13
Boyacá	15
Caldas	17
Caquetá	18
Cauca	19
Cesar	20
Córdoba	23
Cundinamarca	25
Chocó	27
Huila	41
La Guajira	44
Magdalena	47
Meta	50
Nariño	52
Norte de Santander	54
Quindío	63
Risaralda	66
Santander	68
Sucre	70
Tolima	73
Valle del Cauca	76
Arauca	81
Casanare	85
Putumayo	86
San Andrés y Providencia	88
Amazonas	91
Guainía	94
Guaviare	95
Vaupés	97

Vichada	99

Los anteriores códigos son reemplazados de acuerdo al sistema de clasificación del DANE

- **Migración de los datos en Excel al formato WEKA**

Debido a que las bases de datos no cuentan con ninguna extensión de archivo comprendida para la interpretación del aplicativo, se requiere de un proceso de conversión y tratamiento de los datos. Los datos son convertidos a formato Weka por medio de aplicaciones de uso libre como la que podemos obtener del siguiente enlace https://sourceforge.net/projects/exceltoarffconv/.

Figura 5. Conversión de los archivos a formato WEKA

- **Selección de variables**

En este proceso se realizó el análisis de los atributos de las bases de datos, con el fin de determinar la relación o dependencia que existe entre cada uno de ellos, y así, tener evidencia estadística para el proceso de selección de las variables más relevantes. En este proceso de selección se tuvieron en cuenta los siguientes aspectos:

- Identificación de variables.
- Técnicas de análisis.

- **Técnicas de análisis de datos**

La técnica de regresión lineal permite identificar relaciones entre variables numéricas y construir modelos de regresión en donde se tiene una variable de salida y múltiples variables de entrada.

- Variable de salida (dependiente).
- Variables de entrada (independientes).

Dada una muestra de datos donde cada uno es representado por un vector X con N dimensiones determinadas por las variables que definen cada uno de los datos. En las técnicas de regresión lineal se busca estimar la función que mejor explique el comportamiento de los datos. Para lo anterior se debe utilizar el procedimiento de mínimos cuadrados, en el cual se estima el vector de coeficientes que minimiza el error. Este proceso también es llevado a cabo por el algoritmo de clasificación MVS el cual utiliza SMOreg para el entrenamiento de datos.

ANALISIS DEL MODELO ESTADISTICO SMOREG

Para la evaluación del algoritmo nos dirigimos a la opción Explorer en el aplicativo de WEKA, una vez allí se visualizan diferentes opciones como la posibilidad de transformar o normalizar los datos para su tratamiento tal y como se ilustra en la siguiente figura:

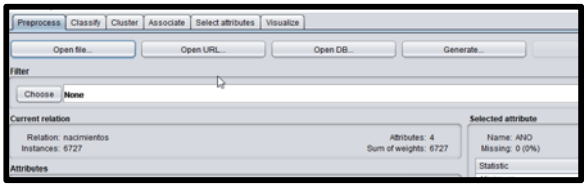


Figura 8. Ventana menú Explorer

En los menús u opciones superiores del programa podemos observar una serie de opciones a continuación descritas

- Open File: apertura de archivos .arf
- Open Url : Archivos de origen web
- Open DB: apertura de archivos de bases de datos como mysql
- Filter Chose : aplicación de métodos de transformación

A continuación, evidenciamos en la figura 9. Las propiedades de cada uno de los atributos

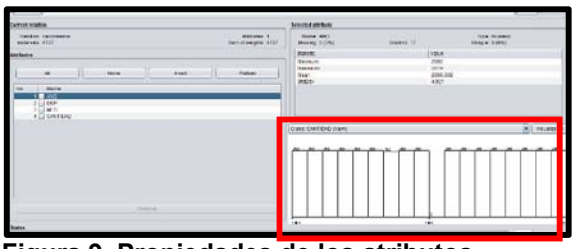


Figura 9. Propiedades de los atributos

WEKA interpreta los datos y realiza un resumen de sus propiedades donde nos indican el tipo de dato, el valor mínimo o máximo y la desviación estándar de los datos.

DESCRIPCION DE RESULTADOS ALGORITMO SMOreg

El significado de los datos arrojados por el algoritmo SMOreg es descrito a continuación:

Correlation Coefficient

Este valor nos indica la correlación que existe entre las variables o atributos de los datos ingresados tal y como se ha mencionado anteriormente y se encuentra comprendido entre 0 y 1. Adicionalmente, El supuesto de este coeficiente nos indica que entre más cercano sea este valor de correlación a 1 la dispersión de los datos será menor y por lo tanto también la efectividad del modelo en la predicción.

Mean Absolute Error

Error Absoluto Promedio, determina la distancia entre los datos predichos y los reales, calculando la diferencia entre los 2 valores. Con el fin de obtener un comportamiento global de los datos, realiza el promedio del error de cada evaluación individual, incluyendo el valor absoluto en la medida de la distancia para evitar compensaciones de error.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|$$

θ^= Valor predicho.
Θi = Valor real.
N = Cantidad de registros.

Root Mean Squared

Este valor de error es calculado para determinar la distancia entre el valor predicho y el valor real, calcula la diferencia entre los dos valores. Con el fin de obtener un comportamiento global de los datos, realiza el promedio de las diferencias elevadas al cuadro de la ecuación con el fin de evitar que el error disminuya. Adicionalmente se calcula la raíz cuadrada del resultado para regresar a las unidades de los datos originales:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2}$$

,

θ^= Valor predicho.
Θi = Valor real.
N = Cantidad de registros.

Relative Aboslute Error

Determina la diferencia promedio entre el valor predicho y el valor real en comparación a la diferencia entre el valor real y el promedio de datos.

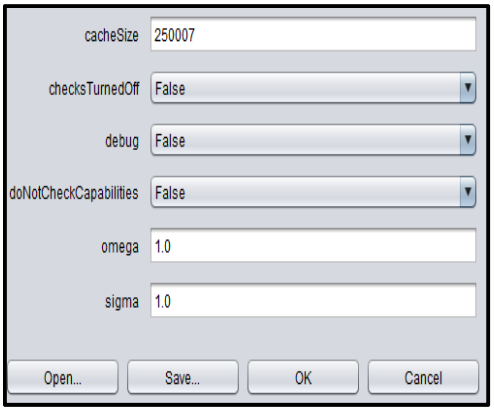
$$\mathrm{RAE} = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta} - \theta_i|}$$

,

El supuesto de estos errores nos indica que cuanto menor sean los valores de porcentaje de error en el modelo relacionado por el software weka mejor será el comportamiento en la efectividad y validación del modelo predictivo.

Por lo tanto, en cuanto menor sean los valores de las cantidades descritas anteriormente, mayor será la efectividad del modelo en la predicción de nuevos datos y como consecuencia son una medida de la validez del modelo arrojado por Weka.

KERNEL



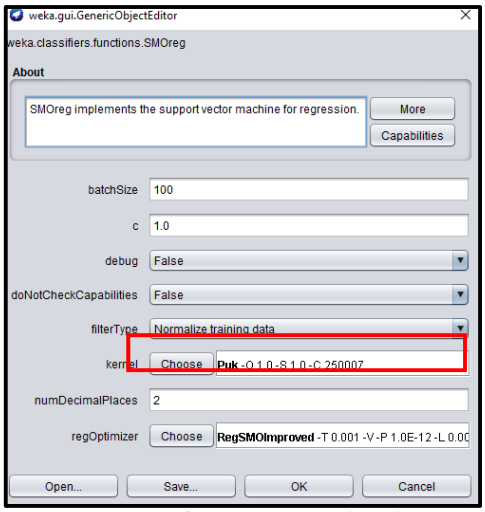


Figura 10. Parámetros de Configuración del Kernel Algoritmo SMOreg

Al seleccionar el algoritmo SMOreg para su evaluación, este nos proporcionara una serie de opciones que podemos modificar, como las que podemos ver en la Ilustración 18. Debemos de configurar y aplicar el kernel que mejor se ajusta al entrenamiento de los datos, el kernel seleccionado para la predicción y ajuste del modelo final fue el kernel PUK.
Teniendo en cuenta la naturaleza y comportamiento de los datos de entrada en un espacio dimensional se debe contemplar el uso de un tipo de kernel que facilite la clasificación de los datos por medio de funciones matemáticas que realizan la transformación del problema de clasificación no-lineal a un problema de clasificación lineal, transformando su espacio dimensional original a un espacio con mayor cantidad de dimensiones.

Datos no separables linealmente

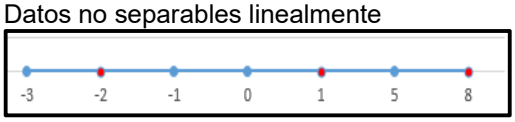


Figura 11. Ejemplo de Vectores

Pero proyectados en 2 dimensiones

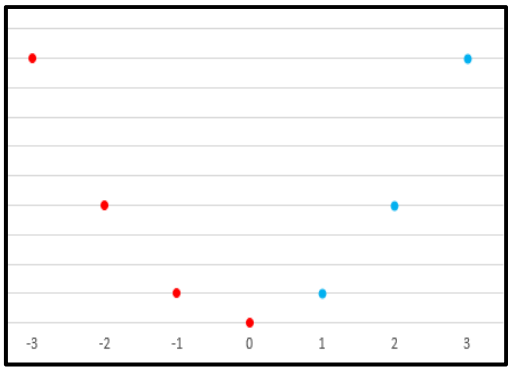


Figura 12. Vectores en Dimensiones

En las figuras 11 y 12 se evidencia la forma de clasificación y funcionalidad del uso del kernel sobre los datos.
Kernel PUK (Pearson Universal Kernel) es una función matemática que depende de 2 parámetros. A partir de la selección de los valores de estos parámetros su forma funcional oscila entre una loretziana y gaussiana para determinar el mejor método de clasificación de los datos.

$$f(x) = \frac{H}{[1 + (\frac{2(x - x_0)\sqrt{2^{(\frac{1}{w})} - 1}}{\sigma})^2]}$$

Donde **H** es el valor máximo de pico de la función y **W** es un parámetro cercanamente relacionado con el comportamiento de las colas de las función y sigma **(σ)** está relacionado con el ancho a la mitad de altura de la función.

NORMALIZACION

El filtro de normalización sobre el modelo consiste en la conversión de los valores numéricos cuyo rango de valores va desde un valor máximo a un valor mínimo, a un rango de valores comprendido entre 0 y 1, esto quiere decir que el valor más alto de un atributo o variable será convertido a su equivalente "**1**" y su valor más bajo será convertido a "**0**". A continuación, se relaciona la fórmula que satisface dicha conversión:

$$ValorNorm = \frac{Valor - Min}{Max - Min}$$

- **Valor=** Valor del atributo mes o año o cantidad.
- **Min=** Valor mínimo de los registros mes o año o cantidad.
- **Max**= Valor máximo de los registros mes o año o cantidad.

Finalmente después de aplicar el filtro de normalización dentro de los parámetros del algoritmo y la aplicación del kernel PUK, se elige el modelo representado en la figura 13 como el modelo más idóneo para generar las predicciones a futuro de las estadísticas de nacimientos y defunciones, dado que los resultados obtenidos representan el mayor porcentaje de efectividad (coeficiente de correlación cercano a (1) y menor error porcentual), basado en los supuestos definidos por el algoritmo SMOreg.

```
SMOreg

weights (not support vectors):
 -       0.9285 * (normalized) ANOS
 +       0      * (normalized) DEP
 +       0.9127



Number of kernel evaluations: 153 (89.071% cached)

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.9281
Mean absolute error                   2926.455
Root mean squared error               3714.849
Relative absolute error                 33.1661 %
Root relative squared error             38.6774 %
Total Number of Instances               17
```

Figura 13. Resultados SMOreg con datos normalizados De Antioquia

- Correlation coefficient: 0,9281.
- Relative absolute error: 33.1661%.
- Root relative squared error: 38.6774%.

VI. RESULTADOS

Los resultados de las predicciones de cada uno de los modelos estadísticos de los diferentes departamentos se relacionan a continuación:

ALGORITMO SMOreg UTILIZADO PARA LA PREDICCION Y COMPORTAMIENTO TOTAL NACIONAL

```
SMOreg

Support vectors:
+1.0 * k[0]
+0.05071662070956696 * k[1]
-1.0 * k[2]
-1.0 * k[3]
+1.0 * k[4]
+1.0 * k[5]
-0.38668650363380086 * k[6]
-1.0 * k[7]
+1.0 * k[8]
+1.0 * k[9]
-1.0 * k[10]
-1.0 * k[11]
+1.0 * k[12]
-1.0 * k[13]
+1.0 * k[14]
+0.33596988292423385 * k[15]
-1.0 * k[16]
 +       0.4515
```

```
-1.0 * k[11]
+1.0 * k[12]
-1.0 * k[13]
+1.0 * k[14]
+0.33596988292423385 * k[15]
-1.0 * k[16]
 +       0.4515



Number of kernel evaluations: 153 (75.041% cached)

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.8528
Mean absolute error                  10582.6118
Root mean squared error              16198.6611
Relative absolute error                 38.6915 %
Root relative squared error             53.9699 %
Total Number of Instances               17
```

Figura 14. Algoritmo SMOreg Total Nacional

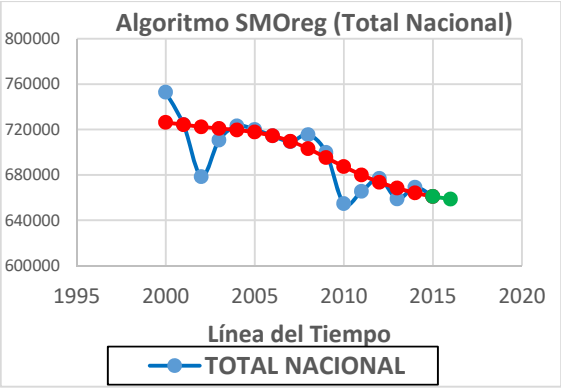


Figura 15. Resultados de Aplicar SMOreg en los Registros Total Nacional

Tabla 5
Predicciones 5 Años Total Nacional

Registro Total de la Predicción de Nacimientos en Colombia					
Años	2017	2018	2019	2020	2021
Total Nacimientos	664028	664096	666192	673350	674990

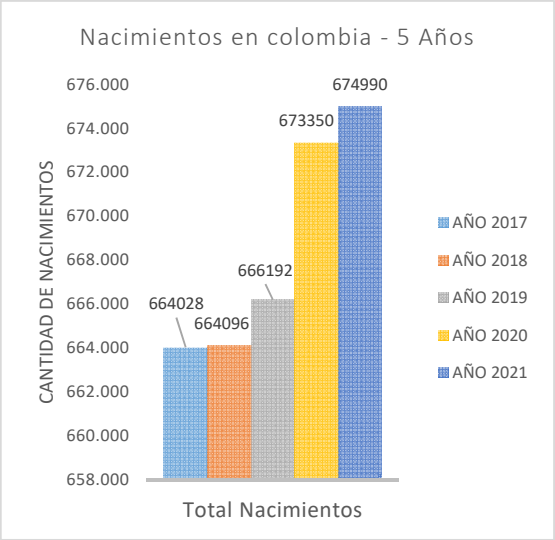


Figura 16. Nacimientos en los próximos años

MODELO UTILIZADO PREDICCION DEFUNCIONES –NO FETALES REGISTRO TOTAL

```
=== Classifier model (full training set) ===

SMOreg

Support vectors:
-1.0 * k[0]
+0.058668367865322035 * k[1]
+1.0 * k[2]
+1.0 * k[3]
-1.0 * k[4]
-1.0 * k[5]
+0.29450379765491674 * k[6]
-1.0 * k[7]
+1.0 * k[8]
+0.7055633808330174 * k[9]
+1.0 * k[10]
-1.0 * k[11]
-1.0 * k[12]
-1.0 * k[13]
-0.05873554635325629 * k[14]
+1.0 * k[15]
+1.0 * k[16]
 +       0.4514
```

```
-1.0 * k[11]
-1.0 * k[12]
-1.0 * k[13]
-0.05873554635325629 * k[14]
+1.0 * k[15]
+1.0 * k[16]
 +       0.4514



Number of kernel evaluations: 153 (92.312% cached)

Time taken to build model: 0.06 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.9769
Mean absolute error                   1918.5316
Root mean squared error               2506.8769
Relative absolute error                 24.9962 %
Root relative squared error             25.0588 %
Total Number of Instances               17
```

Figura 17. Modelo Defunciones-No Fetales

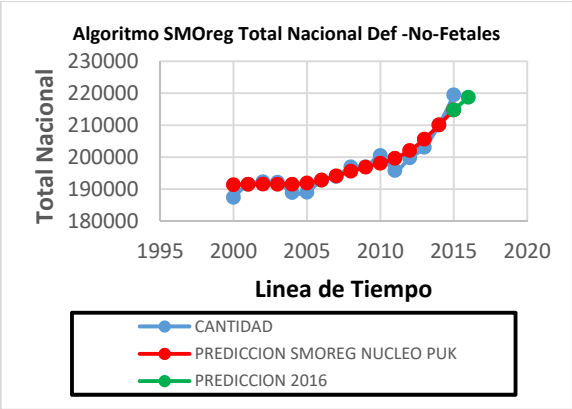


Figura 18. Algoritmo SMOreg Defunciones –no fetales

Tabla 6.
Prediccion - 5 años total general

AÑO	2017	2018	2019	2020	2021
TOTAL	203333	202783	202691	202891	205054

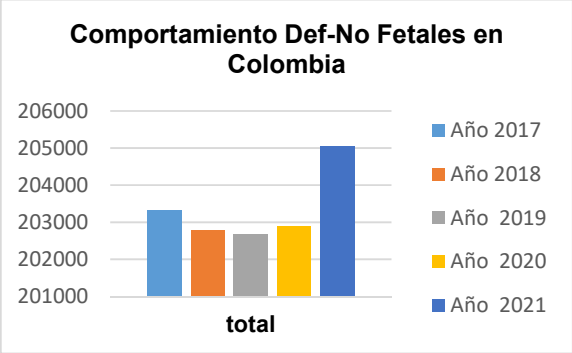


Figura 19. Defunciones -No-Fetales en 5 años

VII. CONCLUSIONES

- El estudio realizado demuestra que el algoritmo SMOreg basado en métodos de análisis de regresión permite el entrenamiento automatizado de modelos predictivos con muy buenos resultados.
- El modelo predictivo SMOreg obtenido después del proceso de evaluación generó como resultado una alta efectividad al obtener porcentajes de error entre el 2% y el 6% en la predicción de los nacimientos, así como porcentajes de error entre el 3% y el 9% en la predicción de las defunciones fetales y no fetales.
- La herramienta WEKA haciendo uso de técnicas de minería de datos permite ajustar modelos predictivos para evaluar el comportamiento de las estadísticas de nacimientos y defunciones en todos los departamentos de Colombia.
- Con la integración del modelo al plugin forecasting en la herramienta WEKA se logran obtener mejores resultados y predecir el comportamiento de los nacimientos y las defunciones en Colombia para los próximos 5 años.

VIII. REFERENCIAS

- [1] Omar González Amor "Predicción de parámetros de energía eólica utilizando modelos de regresión" [en línea], [Consulta: 11 Jul 2018]. Disponible a: http://dspace.uclv.edu.cu/bitstream/handle/123456789/7323/Tesis%20OGA-FINAL.pdf?sequence=1&isAllowed=y
- [2] Yvonne Gala García "Algoritmos SVM para problemas sobre big data" [en línea], [Consulta: 13 Jul 2018]. Disponible a: https://repositorio.uam.es/bitstream/handle/10486/14108/66152_Yvonne_Gala_Garcia.pdf?sequence=1
- [3] Carlos Blanco González, Fernando García Diez "Minería de Datos: Predicción de las condiciones meteorológicas" [en línea], [Consulta: 13 Jul 2018]. Disponible a: https://docplayer.es/21568017-Mineria-de-datos-prediccion-meteorologica.html
- [4] Pablo Casas Muñoz, Roberto García Sánchez "Predicción meteorológica usando WEKA" [en línea], [Consulta: 20 Jul 2018]. Disponible a: https://docplayer.es/66268639-Prediccion-meteorologica.html

IX. BIBLIOGRAFIA

- Carmona Suarez, E. (11 de JULIO de 2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM).* Obtenido de http://www.ia.uned.es/~ejcarmona/publicaciones/%5B2013-Carmona%5D%20SVM.pdf
- Dumais, S. (5 de noviembre de 2001). *Support Vector Machines.*

Obtenido de https://www.microsoft.com/en-us/research/project/support-vector-machines/

Optimización mínima secuencial. (3 de noviembre de 2017). Obtenido de https://en.wikipedia.org/wiki/Sequential_minimal_optimization

- Platt, J. (14 de abril de 1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.* Obtenido de https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/
- http://cleverdata.io/que-es-machine-learning-big-data/
- https://e-archivo.uc3m.es/bitstream/handle/10016/9912/Memoria%20PFC_FVP%20FINAL.pdf
- https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico
- http://revistaseden.org/files/14-CAP%2014.pdf
- https://machinelearningmastery.com/regression-machine-learning-tutorial-weka/
- http://dspace.uclv.edu.cu/bitstream/handle/123456789/7323/Tesis%20OGA-FINAL.pdf?sequence=1&isAllowed=y
- http://jesusnubiola.com/data/tutorialWeka.pdf
- https://knowledgesociety.usal.es/sites/default/files/MANUAL%20WEKA.pdf