Case report

REVISTA Ingeniería

Automatic Classification of Public Investment Megaprojects in Colombia from a Technical, Organizational and Environmental approach

Clasificación Automática del Avance de Megaproyectos de Inversión Pública en Colombia, desde un Enfoque Técnico, Organizacional y Ambiental

Hugo Ernesto Gutiérrez Vanegas[1] ***, Miguel Alberto Melgarejo Rey***[*2]

[1]TIVIT, [2]Universidad Distrital Francisco José de Caldas, LAMIC

[*]Correspondence: mmelgarejo@udistrital.edu.co

Recibido: 30/01/2017. Modificado: 17/06/2017. Aceptado: 31/07/2017.

Abstract

Context: The TOE (Technical, Organizational, and Environmental) framework for the analysis of large scale projects is considered as the basis for the development of megaproject progress classification in accordance with the needs of the national planning agency in Colombia.

Method: Classification of a megaproject progress is supported in the selection of several features taken from the TOE. These feature set is used to configure a database from the projects registered in the project-surveillance platform of the national planning agency in Colombia. The database is used to train two classification models. Information about 3200 projects from 2008 to 2012 was used, covering four economic sectors (Environment and sustainable development, Energy and mining, Health and social care and transportation). Debugging of the database was carried out by an analytic and quantitative approach. Model training and validation were computed with 70 % and 30 % of data respectively.

Results: Obtained models have similar performances beyond 70 % in precision and agree in relevant input features.

Conclusions: This work is a starting point to develop an automatic tool that can be used by the national planning agency of Colombia in the a-priori evaluation of delays in public investment Megaprojects.

Keywords: Megaprojects, Complexity, Management, Neural networks, Support vector machines.

Language: Spanish

Cite this work as: H. Gutierrez, M. Melgarejo, "Automatic Classification of Public Investment Megaprojects in Colombia from a Technical, Organizational and Environmental approach", Ingeniería, vol. 22, no. 3, pp. 377-395, 2017.

DOI: https://doi.org/10.14483/23448393.11483

Resumen

Contexto: Se considera el marco técnico, organizacional y ambiental (TOE, por sus siglas en inglés) para el análisis de proyectos de gran escala como contexto para el desarrollo de clasificadores de avance de megaproyectos, según las necesidades del Departamento Nacional de Planeación, Colombia.
Método: Se establecen algunas características para la clasificación del avance de proyectos de inversión pública, tomadas del marco TOE; a partir de estas, se construye una base de datos que se utiliza para entrenar dos clasificadores del avance de los proyectos reportados en la plataforma de seguimiento de proyectos de inversión del departamento de planeación nacional. Se empleó la información de cerca de 3200 proyectos registrados entre el 2008 y 2012, correspondientes a cuatro sectores económicos (medio ambiente y desarrollo sostenible, minas y energía, salud y protección social y transporte). La base de datos fue depurada siguiendo un enfoque analítico y cuantitativo. Se empleó el 70 % de los datos para entrenamiento y el 30 % para validación.
Resultados: Se obtienen algunos modelos con tasas de clasificación superiores al 70 %, lo que valida la elección de características a partir del análisis del marco TOE.
Conclusiones: Este trabajo es un punto de partida para la configuración de una herramienta que pueda ser usada por el departamento nacional de planeación en la evaluación a priori del retraso de megaproyectos de inversión pública.
Palabras clave: Complejidad, Gestión, Máquinas de vectores de soporte, Megaproyectos, Redes
Idioma: Español

1. Introducción

Los proyectos de ingeniería a gran escala tienden al fracaso a causa de sobrecostos y retrasos, debido al incremento de su complejidad y la poca importancia que se le atribuye a esta; lo anterior ha impulsado nuevos paradigmas de estudio de megaproyectos y la inclusión de métodos de sistemas blandos para modelar la gestión y posibles resultados de estos [1], [2].

Este trabajo se presenta como un estudio de interés a la hora de determinar algunas características relevantes para anticipar el posible avance de un proyecto de inversión pública, a partir de un marco que considera aspectos técnicos, organizacionales y ambientales [5]; las características seleccionadas, de acuerdo con este marco, alimentan dos modelos de clasificadores, cuyos desempeños son similares para este problema en cuestión. Este estudio podría ser el punto de partida para la definición de una herramienta de soporte generalizada en la gestión de proyectos a gran escala del Departamento Nacional de Planeación (DNP).

En primera instancia, se presenta un marco de referencia que reúne algunos conceptos necesarios para el entendimiento del trabajo realizado, recopilando algunos aspectos relevantes frente al marco técnico, organizacional y ambiental; seguidamente, se describe la obtención, selección y acondicionamiento de una base de datos a partir de los proyectos contenidos en la plataforma de seguimiento a proyectos de inversión (SPI). Luego, se describe el entrenamiento de clasificadores de avance de proyectos, mediante el entrenamiento de dos técnicas de inteligencia computacional: redes neuronales [19] y máquinas de vectores de soporte [20].

Por último, se presenta un análisis comparativo de los resultados de clasificación, de acuerdo con un conjunto de estadísticos utilizados en la evaluación de desempeño de clasificadores como son: precisión, sensibilidad, especificidad y curvas ROC. Se propone también una discusión acerca de los resultados obtenidos y el trabajo futuro.

2. Formulación del problema

2.1. Proyectos: cambio de paradigma

Tradicionalmente, un proyecto se entiende como un emprendimiento con un comienzo y un fin, dirigido por personas quienes buscan lograr unas metas establecidas teniendo restricciones de costo, agenda y calidad [1].

Los proyectos podrían entenderse como sistemas complejos [4], [5], considerándose así un marco de trabajo para tratar el problema común de fracasos en los proyectos, principalmente en términos de sobrecostos, retrasos e inefectividad en las metas alcanzadas. La complejidad de proyectos se caracteriza por dos dimensiones: (a) la complejidad estructural compuesta por el tamaño o número de elementos en el proyecto junto con la interdependencia entre estos; (b) La incertidumbre vista en las metas del proyecto y los medios definidos para lograr estas metas [3].

La Figura 1 presenta un contraste entre un proyecto visto como un sistema secuencial, compuesto por paquetes de actividades independientes y un proyecto visto como un sistema complejo, compuesto por elementos interdependientes. El cambio de paradigma en cuanto al entendimiento de un proyecto y su gestión implica un acercamiento a la noción de complejidad, la cual se manifiesta no solo en los múltiples actores, factores o áreas involucrados sino también en el amplio rango de posibles relaciones que se dan entre estos [3], [4].

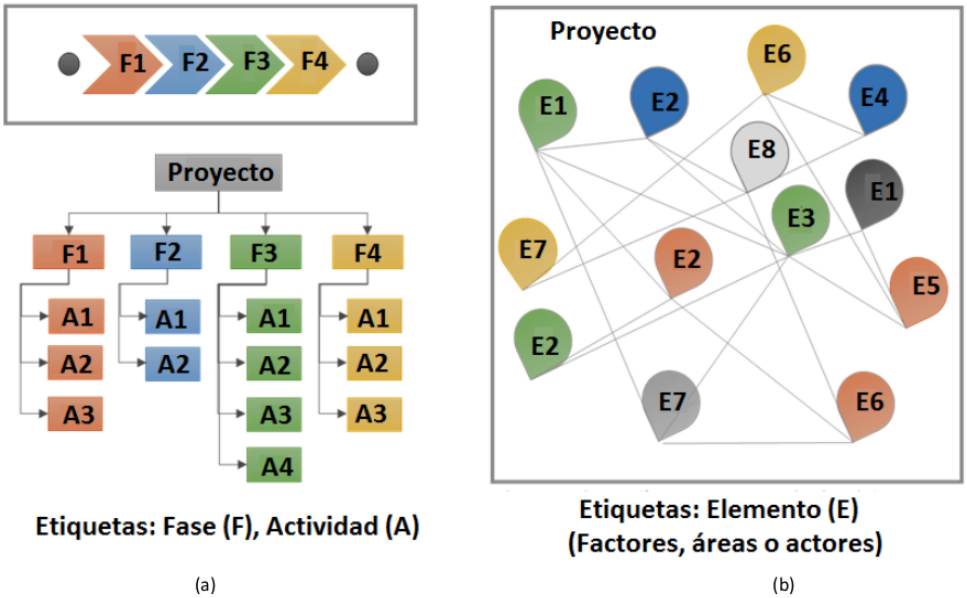

Figura 1. a) Proyecto visto como un sistema secuencial. b) Proyecto visto como un sistema complejo. Fuente: elaborado con base en [3].

Con el fin de contribuir en la gestión de proyectos de gran escala, se contempla en este trabajo el marco de referencia técnico, organizacional y ambiental (TOE, por sus siglas en inglés) [5]; este marco de referencia fue desarrollado desde el análisis de dos fuentes de conocimiento: (a) consultas bibliográficas de distintos autores que han abordado la complejidad de proyectos, y (b) entrevistas a gerentes e integrantes de seis megaproyectos. El resultado de esta combinación es un conjunto de 50 elementos distribuidos en tres categorías (TOE), las cuales se muestran en la Tabla I.

Con el fin de detallar un poco más los aspectos tratados por el TOE sin entrar a realizar una revisión exhaustiva de estos, se presenta la Figura 2(a), en la cual se muestra la distribución porcentual de los elementos del TOE de acuerdo con las categorías mencionadas en la Tabla I. Nótese que la categoría que reúne la mayor cantidad de elementos es la organizacional, sin embargo, sumados el componente técnico y ambiental superan a la primera. De esta forma, la distribución de los elementos sobre las categorías podría considerarse uniforme en términos prácticos, así el marco no muestra especial inclinación hacia alguna categoría en particular.

Tabla I. Aspectos cualitativos del marco TOE

Técnico	Organizacional	Ambiental
Metas	Tamaño	Involucrados
Alcance	Recursos	Locación
Actividades	Equipo de proyecto	Condiciones mercado
Experiencia	Confianza	

En la Figura 2(b), 2(c) y 2(d), se puede observar que en cuanto a los aspectos técnicos se refiere, los elementos relacionados con las tareas del proyecto son un elemento dominante; esto sugiere que la cantidad y naturaleza de las tareas involucradas en un proyecto de gran escala, contribuyen determinantemente en la configuración de su complejidad. En la categoría organizacional, los elementos que dominan están relacionados con el tamaño del proyecto y los recursos disponibles, evidenciándose que tanto las restricciones económicas como la cantidad de agentes involucrados son atributos moldeadores de la complejidad del proyecto en concordancia con la Figura 1(b). Por último, en la categoría ambiental se nota que los involucrados o afectados por el proyecto (i.e. *stackholders*), constituyen un aspecto relevante dado que su participación condiciona la complejidad del proyecto en cuanto al incremento de relaciones y el aumento de restricciones.

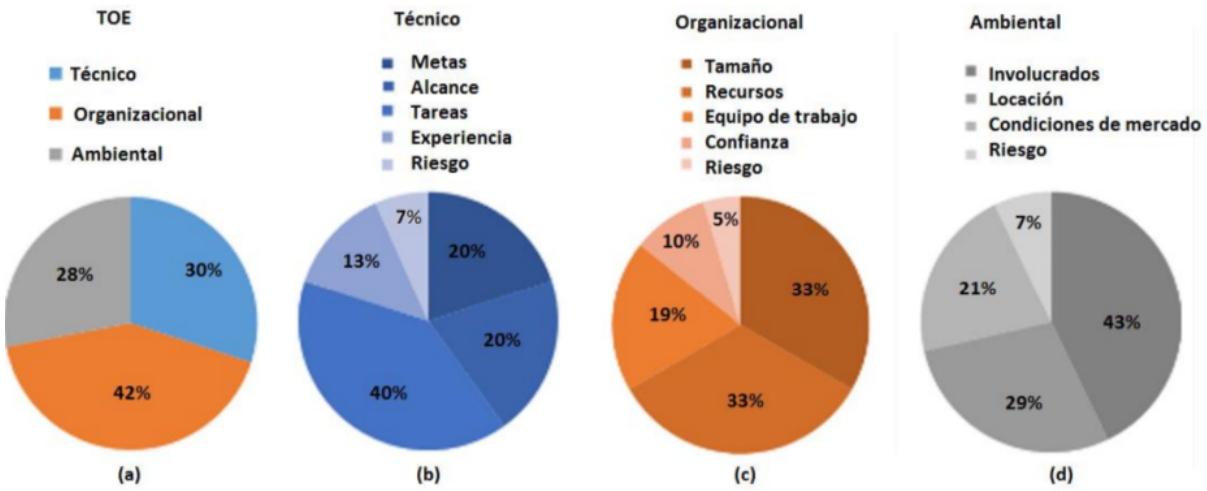

Figura 2. Distribución porcentual de los elementos marco TOE. a) Porcentajes de distribución general de los 50 elementos TOE. b) Distribución de elementos de tipo técnico. c) Distribución de elementos de tipo organizacional. d) Distribución de elementos de tipo ambiental.

2.2. Inteligencia computacional en la gestión de proyectos

Según el marco referencial consultado para este trabajo, se han encontrado tres líneas de estudio en los temas de análisis y gestión de proyectos a gran escala desde la inteligencia computacional: (a) técnicas de inteligencia computacional para la selección de proyectos según estructura y escala tecnológica [6]–[8]; (b) algoritmos computacionales para problemas de agendamiento y costo en proyectos específicos [9]–[11], y (c) sistemas de predicción del éxito de proyectos basados en redes neuronales [16]. Sin embargo, no se encontraron referencias que aborden la estimación del avance de megaproyectos de inversión pública desde estas técnicas.

Las posibilidades de usar la inteligencia computacional en el marco de la evaluación de proyectos serían muy amplias al igual que en cualquier otro dominio de aplicación. Los teoremas NFL (*Non Free Lunch*) [17], [18] previenen establecer una inclinación *a priori* por algún algoritmo en particular; de aquí surge la necesidad de llevar a cabo una comparación entre técnicas cuando se pretende abordar un nuevo problema.

Particularmente, se consideran dos técnicas de clasificación automática de uso habitual desde la inteligencia computacional: redes neuronales y máquinas de vectores de soporte. Se presentará una breve introducción a las mismas invitando al lector a que profundice a través de la consulta de [19] y [20]. En cuanto al marco de interés de este trabajo, se han reportado en la literatura modelos de gestión de proyectos que incluyen redes neuronales y SVM (máquina de soporte vectorial), a manera de ejemplo se mencionan algunos: selección óptima de proyectos de innovación tecnológica aplicando árboles de decisión neurodifusos [6], agendamiento de proyectos según teoría de restricción de recursos con técnica de red neuronal basado en modelo de colonias de hormigas [10], sistemas de selección en portafolio de proyectos con redes neuronales y sistemas difusos [8], [16], predicción de la duración de proyectos con SVM [21], [22] e identificación del riesgo en proyectos por medio de una combinación entre SVM y algoritmos de colonias de hormigas [23].

2.2.1. Redes neuronales

Una red neuronal está constituida por neuronas interconectadas y arregladas en tres capas (esto último puede variar). Las neuronas son funciones no lineales de múltiples entradas y una sola salida, cuya estructura está conformada por una combinación lineal de las entradas, seguidas de una función de activación donde aparece su carácter no lineal. Los parámetros de la combinación lineal se denominan los pesos de la neurona.

Los datos de entrada son preprocesados por una primera capa de neuronas, la cual produce un conjunto de niveles de disparo; estos niveles ingresan a la capa intermedia donde son combinados por medio de otras neuronas, generando nuevamente un abundante conjunto de niveles de disparo. La capa de salida combina los niveles de disparo intermedios para producir una única salida que representa en el caso particular de un clasificador una aproximación de la clase correspondiente a los datos suministrados en la entrada.

Las redes neuronales se entrenan con el propósito de lograr una generalización en la salida a partir de datos conocidos de entrada [12]. Una red neuronal debe aprender a calcular la salida correcta

para cada vector ejemplo de entrada; este proceso de aprendizaje es denominado proceso de entrenamiento o acondicionamiento de la red neuronal y el conjunto de datos sobre el cual se basa este proceso es llamado conjunto de datos de entrenamiento.

Se reconoce hoy en día principalmente dos métodos de aprendizaje con los que se pueden adaptar los pesos de una red neuronal en aras de lograr que esta trabaje como un clasificador, estos métodos son: (a) aprendizaje supervisado, el cual se caracteriza porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (i.e. un supervisor) que determina la respuesta que debería generar la red a partir de una entrada determinada; (b) aprendizaje no supervisado, las redes con aprendizaje no supervisado no requieren influencia externa para ajustar los pesos de las conexiones entre sus neuronas, la red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta.

2.2.2. Máquinas de vectores de soporte

La máquina de vectores de soporte (SVM, por sus siglas en inglés), es un modelo de aprendizaje supervisado que procesa ejemplos numéricos reales y reconoce sobre estos patrones separados geométricamente por un hiperplano. Un clasificador SVM básico utiliza funciones binarias lineales de tipo no probabilístico; de esta manera, dado un conjunto de ejemplos de entrenamiento, se inicia una representación de los ejemplos mediante un mapeado de puntos en un espacio N dimensional [13]. Nuevos ejemplos son referenciados a dos posibles grupos en tal espacio. A medida que se va ejecutando el entrenamiento, la topología de los agrupamientos se va ajustando.

Para el entrenamiento de una SVM se suelen seguir tres pasos: (a) mapear los puntos de entrenamiento a un espacio vectorial de mayor dimensión, (b) construir un hiperplano que separe los puntos en sus clases respectivas y (c) clasificar un punto nuevo de acuerdo con su ubicación con respecto al hiperplano. Diferentes vectores de soporte pueden generar un mismo hiperplano; así, por cuestiones de estandarización, dichos vectores deben escalarse de manera que la distancia entre el hiperplano y el patrón más cercano a este sea de uno. Se busca en el entrenamiento que el hiperplano obtenga la forma más confiable, es decir que exista la máxima distancia posible entre este y los patrones.

El aprendizaje en la SVM se lleva a cabo considerando que en su gran mayoría los problemas no son linealmente separables; por tanto, se busca hacer una transformación de los ejemplos de entrenamiento a un espacio vectorial de alta dimensión donde sea posible una solución mediante una separación lineal. Tal transformación se pude implementar por medio de un conjunto de funciones parametrizables que deben conformar un kernel, la elección del kernel depende en gran medida del problema de aplicación. Dentro de los kernels comúnmente usados se encuentran los lineales, polinomiales, radiales y sigmoides.

2.3. La plataforma SPI

Para este trabajo, el DNP ha puesto a disposición un registro de datos de proyectos que involucran presupuesto general de la nación; dicho registro de datos se encuentra alojado en la plataforma seguimiento a proyectos de inversión (SPI) [14], [15]. SPI es una ventana directa para ver los lo-

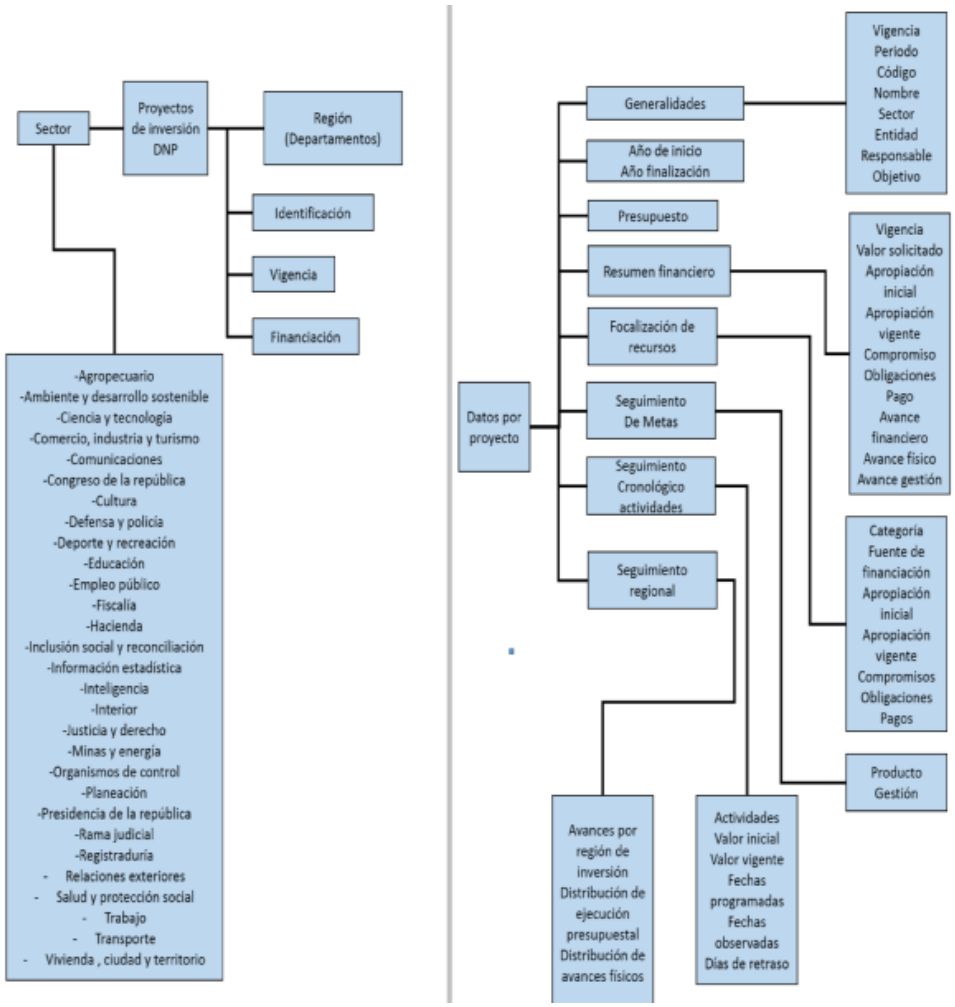

Figura 3. Descripción del SPI: (a) Clasificación de los proyectos, (b) Datos registrados.

gros y resultados generales en materia de inversión pública de entidades que planeen, ejecuten y evalúen proyectos de inversión; también es definida cómo una herramienta de utilidad al ciudadano colombiano para verificar las políticas del Plan Nacional de Desarrollo y la gestión del Gobierno Nacional.

La plataforma SPI ofrece información como: (a) los objetivos, presupuestos anuales, logros y metas de los proyectos más importantes del gobierno nacional; (b) una perspectiva de avance físico-financiero y cronológico de todos los proyectos de inversión pública, y (c) la evolución físico-financiera de los proyectos bajo múltiples lentes como por entidad, por sector, por departamento, por proyecto e incluso por estrategias transversales como Red Juntos, TIC o Ciencia y Tecnología [15].

Dentro de esta herramienta, los proyectos son clasificados de acuerdo con su región, origen de su financiación y sector de aplicación, tal como se muestra en la Figura 3(a). Nótese que dentro del aplicativo se reconocen actualmente más de veinte sectores de aplicación; en este sentido, se evidencia una diversidad latente en los proyectos registrados en cuanto a las tres dimensiones identificadas en el marco TOE.

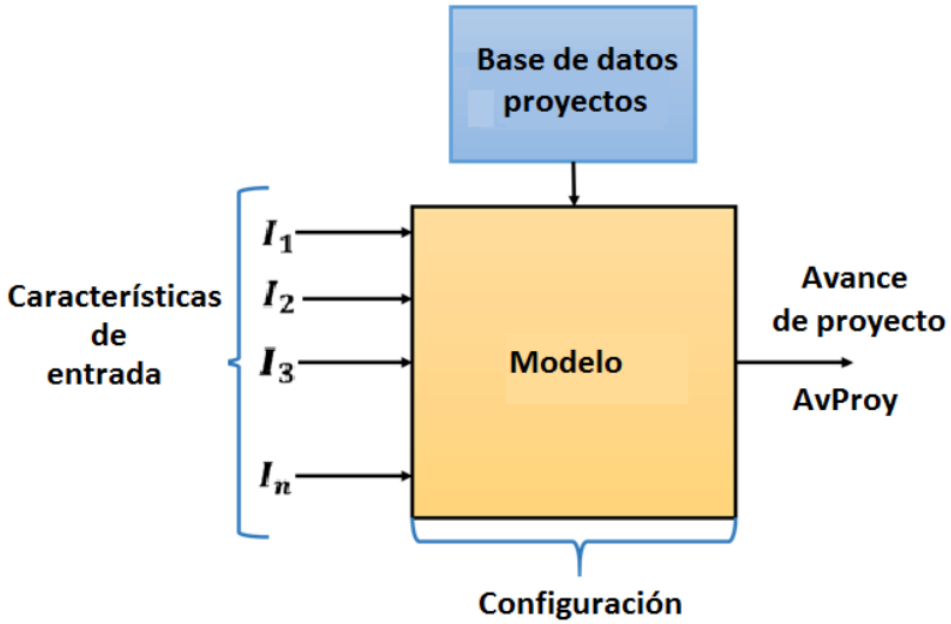

Figura 4. Representación del modelo de clasificación de avance de proyectos.

La plataforma SPI contiene registro de datos de proyectos entre 2008 y 2012 que involucran Presupuesto General de la Nación (PGN). El total bruto de proyectos es de 7934 divididos en veintinueve sectores y teniendo cerca de 43 características por proyecto, tal como se puede apreciar en la Figura 3(b). El indicador avance proyecto es prioritario en el sentido de que permite medir de manera porcentual el éxito anual por proyecto según metas logradas.

2.4. Pregunta de investigación

Según el alcance de este trabajo, el cual está basado en la necesidad presente en el DNP, según su área de inteligencia de negocios y la disponibilidad de los datos, se requiere un modelo computacional que clasifique el indicador avance a partir de algunos datos de entrada al inicio de cada proyecto; lo anterior con el fin de evaluar *a priori* el potencial retraso que un megaproyecto podría llegar a tener dadas sus características. En ese sentido, se dirige entonces la pregunta de investigación detrás de este trabajo: ¿cómo puede ser la clasificación, basada en inteligencia computacional del avance de proyectos, que considere características del marco TOE y una base de datos históricos de proyectos ya ejecutados SPI-DNP? La Figura 4 pretende aclarar la pregunta de investigación con la que se busca definir las características de entrada, la base de datos y la configuración de un modelo que permita la clasificación de avance de proyectos.

3. Materiales y métodos

El desarrollo de este trabajo partió de la obtención y acondicionamiento de una base de datos según los registros provistos en SPI y su posible asociación con los elementos aportados por el TOE; posteriormente se definieron dos escenarios para efectuar la clasificación de avance de proyectos mediante las técnicas de inteligencia computacional presentadas en la sección anterior.

3.1. Base de datos

En la plataforma SPI originalmente se contó con 7934 proyectos pertenecientes a veintinueve sectores económicos. El primer filtro de selección fue acorde con los proyectos pertenecientes a los sectores que más relación tuvieran con la ingeniería, estos fueron: transporte (960), ambiente y desarrollo sostenible (1027), salud y protección social (590) y minas y energía (554). La Tabla II presenta la distribución de la cantidad de proyectos por año para cada uno de los sectores seleccionados.

Tabla II. Distribución de proyectos por año según sectores seleccionados

	Año					
Sector	**2008**	**2009**	**2010**	**2011**	**2012**	**Total**
Ambiente y desarrollo sostenible	237	241	228	182	139	1027
Minas y energía	116	127	122	126	63	554
Salud y protección social	125	119	124	128	94	590
Transporte	183	209	175	178	215	960

Fuente: elaboración propia

Para el acondicionamiento de la base de datos sobre el conjunto de características definidas en la Tabla III, se efectuaron cinco pasos: (a) depuración de los indicadores SPI; (b) asignación de clases en característica de salida avance de proyectos; (c) balance de clases; (d) estudio correlación lineal, y (e) generación del espacio de características.

Tabla III. Características de entrada al modelo

No.	**Nombre**	**Descripción**	**Medida**	**Asociación con TOE**
1	Duración (Dur)	Duración del proyecto	Meses	TOE16: duración de proyecto
2	Apropiación vigente (ApV)	Presupuesto/capital apropiado al proyecto	Pesos	TOE18: Cantidad en inversión de capital (CAPEX)
3	Componente producto (CPr)	Grado de importancia y esfuerzo dedicado a metas producto(tangibles)	Porcentaje	TOE2: alineación de metas
4	Componente gestión (CGe)	Grado de importancia y esfuerzo dedicado a metas gestión(intangibles)	Porcentaje	TOE2: alineación de metas
5	Número de metas producto (NMPr)	Cantidad de metas tangibles del proyecto	Metas	TOE1: número de metas TOE3: claridad de metas
6	Número de metas Gestión (NMGe)	Cantidad de metas intangibles del proyecto	Metas	TOE1: número de metas TOE3: claridad de metas

3.1.1. Depuración de los indicadores SPI

La depuración de los indicadores SPI se realizó sobre una confrontación entre estos *versus*, los elementos TOE con el fin de establecer un primer conjunto de elementos útiles de trabajo en aras de definir un primer juego de características de acuerdo con la Figura 4. La asociación de los datos de proyectos SPI con los elementos del marco TOE se desarrolló de la siguiente manera.

- Análisis de datos SPI y TOE: se extrajeron los datos SPI y una breve descripción de estos al igual que de los elementos TOE en aras de su entendimiento.
- Pre procesamiento de datos SPI: algunos indicadores SPI fueron procesados para determinar datos de mayor relevancia con respecto al marco de referencia TOE. Estos casos fueron año inicio y año finalización (datos SPI9 y SPI10), los cuales al ser sustraídos entre sí revelaron la duración completa de cada proyecto, y seguimiento regional de metas producto y metas gestión (dato SPI30).
- Conservación de datos SPI para la identificación de los proyectos: los datos SPI1 y SPI3, aunque no tienen relación con el marco TOE, fueron extraídos para facilitar la identificación y temporalidad de cada proyecto durante el presente estudio.

De este ejercicio se obtuvo el siguiente conjunto de elementos asociativos entre el SPI y el TOE, del cual a su vez se define un conjunto de seis características de entradas, tal como se presenta en la Tabla III.

- Sector (SPI5): permitió la selección de proyectos afines al área de ingeniería y poder contrastarlos con el marco TOE. Se aclara que este dato no contiene información detallada para determinar *riesgos técnicos* (TOE15) o *nivel de competencia* (TOE 49).
- Avance de proyecto (SPI9): se relaciona con *alineación de metas* TOE2. Representa la salida objetivo de clasificación. Se mide a manera de porcentaje.
- Año de inicio (SPI10): relación directa con duración de proyecto TOE16.
- Año de finalización (SPI11): relación directa con duración de proyecto TOE16.
- Apropiación vigente (SPI16): indica el CAPEX o cantidad de inversión de capital TOE18. Representa el presupuesto o capital apropiado para el proyecto en el año en curso.
- Descripción (SPI30): se puede evidenciar aquí relación con número de metas TOE1, alineación de metas TOE2, claridad de metas TOE3. Representa el tipo de meta definida por el proyectista ya sea de gestión o producto. Se requiere una interpretación de texto por parte de un experto para lograr establecer la cantidad. En el caso de meta producto se hace referencia a la consolidación material de entregables, mientras que las metas gestión hacen referencia al cumplimiento de actividades dentro del proyecto.
- Meta total (SPI31): en relación con TOE2, permite determinar el grado de importancia y esfuerzo dedicados a los tipos de meta, bien sea producto o gestión.

- Distribución de ejecución presupuestal (SPI46): es un dato orientado a la división del presupuesto general de la nación. Se usará como una variable categórica para hacer análisis parcial del comportamiento de los modelos de clasificación. Cubre desde elementos como número de locaciones (i.e. territorios administrativos involucrados) TOE22, condiciones climáticas TOE44 y riesgos ambientales TOE50.

3.1.2. Asignación de clases

De acuerdo con algunas discusiones con los expertos del DNP, determinar el avance de un proyecto desde el inicio de cada año no implica establecer un porcentaje sin incertidumbre. Una primera vía sería la clasificación de aquellos proyectos que tendrán un avance significativo con respecto a aquellos que no lo tendrán; así, una posible solución se podría establecer a manera de un ejercicio de clasificación en el que se distinga proyectos con un buen avance frente aquellos con un avance pobre. La asignación de categorías como "alto" y "bajo" se estableció a partir de las discusiones con los expertos del DNP; no existe a la fecha como tal una asignación formal de estas categorías según el porcentaje de avance de proyecto. Por tanto, para propósitos de este trabajo se realiza la propuesta de clases descrita en la Tabla IV. La franja intermedia fue pensada como un intervalo que permitiera agrupar los proyectos con un avance intermedio o que no fue cercano a las bandas límite del 30 % y 70 %

Tabla IV. Descripción de clases

Avance de proyecto	Clase	Descripción
Más del 70 %	1	Proyecto con un avance mayor al 70 % de lo planeado al cierre del año de vigencia.
Entre el 30 % y el 70 %	0	Proyecto con un avance entre el 30 % y 70 % de lo planeado al cierre del año de vigencia
Menos del 30 %	-1	Proyecto con un avance menor al 30 % de lo planeado al cierre del año de vigencia

3.1.3. Equilibrio de clases

De acuerdo con las clases propuestas, se realizó una revisión de los datos que contienen estas. La revisión consiste en verificar que presente una similitud en cuanto al número de proyectos clase 1 con respecto a los proyectos de clase -1. De ser posible, el entrenamiento de los modelos clasificadores tendría un mejor desempeño puesto que habría un balance de ejemplos. En aras de ajustar el desbalance, se revisó el avance de los proyectos que estuvieran detrás de este fenómeno; así se exploró el conjunto de datos para detectar aquellos proyectos que están fuera de tendencia en cuanto a avance se refiere. Se detectaron 279 proyectos en SPI con avances entre el 105 % y 16082 %; estas cantidades sorprenden dado que un avance de más del 100 % no tiene sentido. Tales magnitudes se deben principalmente a errores humanos al momento de introducir la información al SPI. Estos proyectos generan en parte el desequilibrio entre las clases, por lo que fueron tratados como puntos fuera de tendencia y removidos de la base de datos.

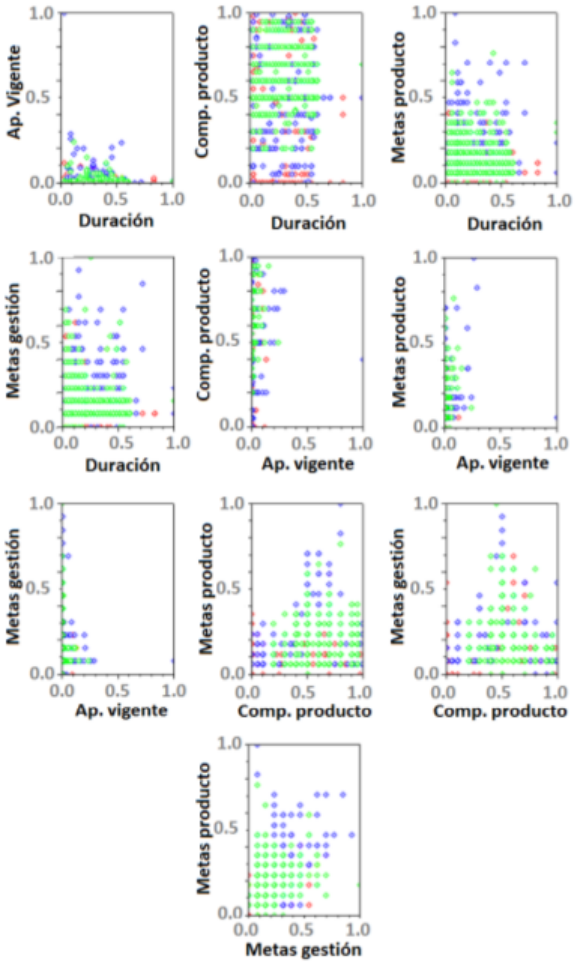

Figura 5. Representación por secciones bidimensionales del espacio de características obtenido.

3.1.4. Estudio de correlación lineal

De acuerdo con la aplicación del coeficiente de correlación lineal de Pearson entre las características seleccionadas luego del balance clases (Tabla V), se detecta que las variables componente producto (CPr) y componente gestión (CGe) son complementarias en su gran mayoría para lograr un 100 %. Como ejemplo se detecta que, si un proyecto tiene como componente producto 45 %, el dato componente gestión es 55 %. Puesto que esta situación se presenta en alrededor de un 95 % de casos, se sustraerá la variable componente gestión y se mantendrá la variable componente producto en representación de ambas.

Tabla V. Coeficiente de correlación de Pearson entre características

Campos	Dur	ApV	CPr	NMPr	NMGe
Dur	1	0.025	0.250	0.001	-0.043
ApV	0.025	1	0.057	0.148	-0.020
CPr	0.250	0.057	1	0.187	0.121
NMPr	0.001	0.148	0.187	1	0.537
NMGe	-0.043	-0.020	0.121	0.537	1

3.1.5. Generación del espacio de características

Finalmente, la base de datos se desplegó a modo de espacio de características representado por secciones bidimensionales, tal como se muestra en la Figura 5. Esta representación se llevó a cabo

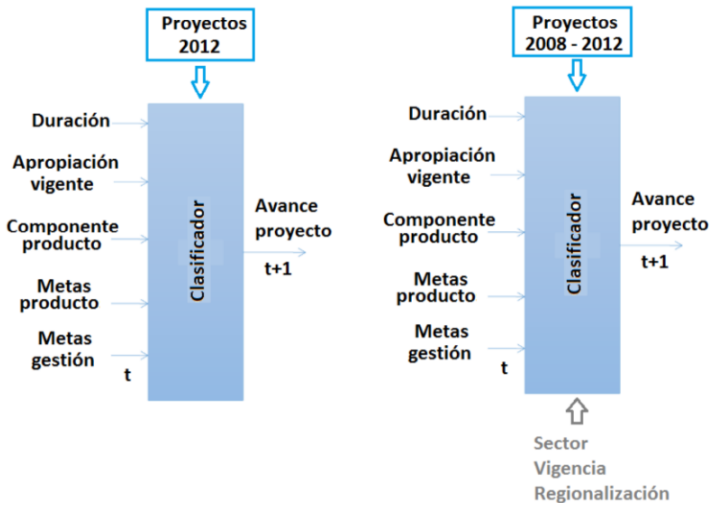

Figura 6. Representación de escenarios base de datos. a) Solo proyectos 2012. b) Proyectos 2008 a 2012.

con el fin de aproximarse a la naturaleza del problema de clasificación. Se anotaron las siguientes observaciones:

- El espacio apropiación vigente *versus* duración, aunque la mayor parte de los datos se encuentran en la parte inferior del plano, no es fácilmente separable la clase 1 (color verde) de la clase -1 (color rojo).

- Para los pares de características componente producto *versus* duración, metas producto *versus* duración, metas gestión *versus* duración, metas producto *versus* componente producto y metas producto *versus* metas gestión, los datos se encuentran distribuidos en todo el plano.

- En los pares de características componente producto *versus* apropiación vigente, metas gestión *versus* apropiación vigente y metas producto *versus* apropiación vigente, se puede distinguir la forma de unas pequeñas barras en los datos orientados al lado izquierdo de cada plano; sin embargo, no son fácilmente diferenciables las clases ya que hay datos de ambas superpuestos.

- Por último, el par metas gestión *versus* componente producto, aunque denota cierta forma triangular en sus datos, existe una buena cantidad de ellos que se encuentran mezclados en los bordes izquierdo y derecho del plano visto.

Las anteriores observaciones sugieren que el problema de clasificación de avance proyecto a partir de las características seleccionadas desde el TOE en relación con el SPI se muestra como un problema de separación no lineal.

3.2. Escenarios de bases de datos de proyectos

Desde las discusiones con los expertos del DNP, se identificó que la cultura de registro de proyectos es un proceso de aprendizaje por parte de los proyectistas; así, de los primeros años de registro se sabe que pueden existir imprecisiones en los datos alojados en el SPI. Este aspecto

mejora conforme aparecen nuevos registros de proyectos en los años siguientes; por tanto, se consideró plantear dos escenarios para el desarrollo de la clasificación. En la primera se buscó reducir la incertidumbre frente al registro de atributos en el SPI, considerando los proyectos más recientes; sin embargo, se hace evidente una posible pérdida de información por contar con una muestra reducida de proyectos. En el segundo caso, se exploró la situación contraría en la cual, la muestra de proyectos aumenta, pero se reduce la fiabilidad de la información contenida por proyecto.

La Figura 6 representa gráficamente los dos escenarios propuestos en relación con el modelo de clasificación de la Figura 4 .El primero de ellos consiste en tomar solo los proyectos del 2012, conjunto de proyectos más reciente que conservaría una única temporalidad de registro y ejecución de cada proyecto. En este primer escenario, al estar centrado en el último año de observación, se considera un grupo selecto de proyectos sobre los cuales se ha realizado en registro juicioso de sus atributos en el SPI. Como segundo escenario se propone tomar el conjunto de datos de proyectos desde el 2008 hasta el 2012. Esto permite tener en cuenta una mayor cantidad de proyectos. Este escenario se propone en el sentido de contar con una muestra más amplia de proyectos, pero sobre la cual existe incertidumbre frente al registro de sus atributos.

3.3. Entrenamiento y prueba de modelos de clasificación

Para los dos escenarios de base de datos, se desarrolló una metodología de experimentación para un modelo de clasificación basado en redes neuronales y otro modelo basado máquinas de vectores de soporte. El modelo basado en redes neuronales consistió en una arquitectura *Feed-forward* implementada con ayuda de la caja de herramientas ANN Toolbox del *software* Scilab V.5.5.0. Como parámetros fijos se tuvieron 100 *epochs* y tasa de aprendizaje 0,1; como parámetros variables se realizaron todas las combinaciones posibles de entradas (26 configuraciones en total) y variaciones en el número de neuronas de capa oculta según potencia de 2 (2, 4, 8 y 16 neuronas). Para cada configuración resultante se realizaron 100 pruebas de entrenamiento y validación, dada la aleatoriedad presente en la inicialización de los pesos de las redes. La Figura 7(a) representa la configuración de experimentación de las redes neuronales.

El modelo basado en máquinas de vectores de soporte consistió en un algoritmo de entrenamiento SMO provisto en la caja de herramientas LIBSVM Toolbox del *software* Scilab V5.5.0. Como parámetros variables se realizaron todas las combinaciones posibles de entradas (26 configuraciones en total) y variaciones en el tipo de configuración de función de kernel (lineal, polinomial, base radial y sigmoide). Para cada configuración resultante se realizaron 100 experimentos dada la aleatoriedad en la inicialización de los parámetros de los kernels. La Figura 7(b) representa la configuración de experimentación de las máquinas de vectores de soporte.

En resumen, por cada escenario y tipo de modelo se realizaron 10400 experimentos según (1) y (2). En cada uno de los experimentos se segmentó la base de datos de manera aleatoria, garantizando un 70 % para entrenamiento y el 30 % para prueba.

$$Total\_config = No.config.caract\_entrada * No.config\_modelo \tag{1}$$

$$Total\_experimentos = 100 * experimentos * Total\_config \tag{2}$$

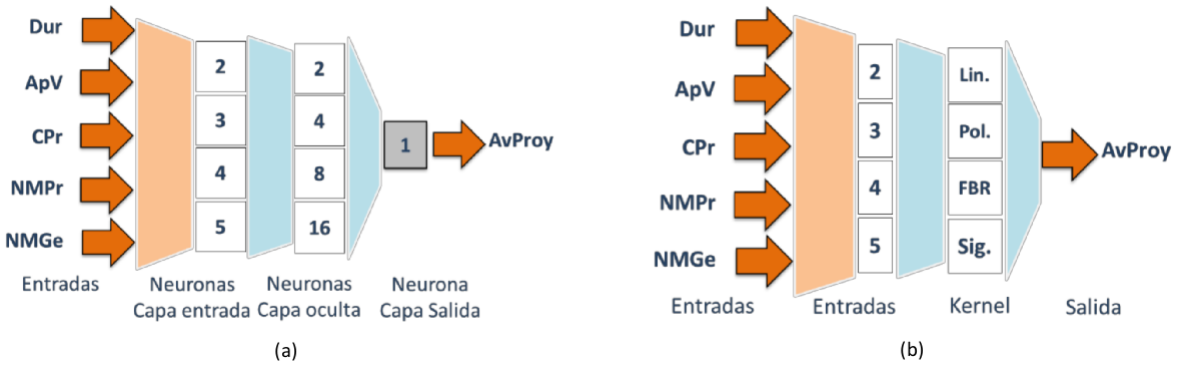

Figura 7. (a) Representación de configuración de la red neuronal, (b) Representación de configuración de Máquinas de Vectores de Soporte.

3.4. Estadísticos de interés

Para cada uno de los experimentos realizados se tomaron mediciones de precisión P, especificidad E y sensibilidad S, ecuaciones (3), (4) y (5). Las tres en conjunto permiten evaluar el desempeño del clasificador teniendo en cuenta los verdaderos positivos V_p y verdaderos negativos V_n dado el conocimiento en los datos de los Ejemplos positivos E_p y los Ejemplos negativos E_n.

$$P = \frac{V_p + V_n}{E_p + E_n} \tag{3}$$

$$E = \frac{V_n}{E_p + E_n} \tag{4}$$

$$S = \frac{V_p}{E_p + E_n} \tag{5}$$

También fueron usadas las curvas ROC como una herramienta gráfica importante para determinar el desempeño del clasificador a través de la relación existente entre la tasa de verdaderos positivos y la tasa de falsos positivos (Figura 8). Particularmente se consideró caracterizar el área bajo la curva ROC (AUC).

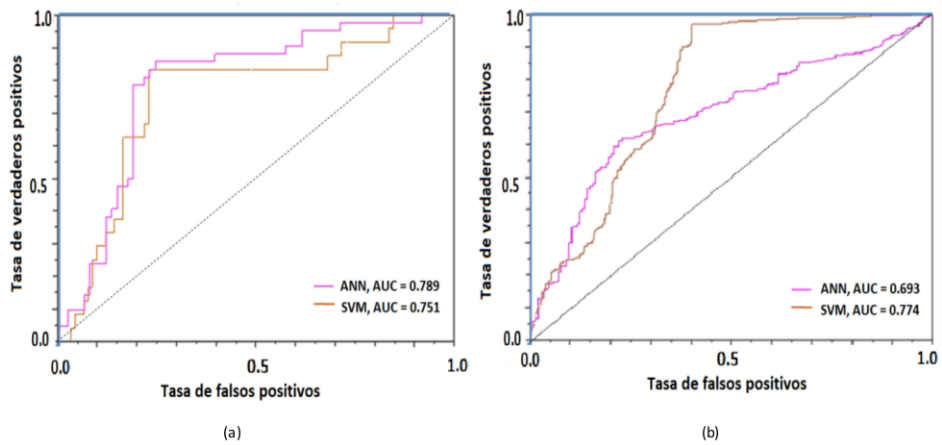

Figura 8. Curvas ROC mejores modelos ANN y SVM: (a) escenario 1, (b) escenario 2.

4. Resultados

Se presenta aquí los resultados del estudio comparativo de clasificación de avance de proyectos mediante ANN y SVM en los escenarios generados a partir de la base de datos.

4.1. Primer escenario

La Tabla VI muestra las mediciones de precisión, especificidad, sensibilidad y AUC para las dos mejores configuraciones de redes neuronales y SVM encontradas. La Figura 8(a) presenta las correspondientes curvas ROC para estos dos clasificadores. De aquí se puede observar que el modelo basado en una red neuronal supera al basado en una SVM, puesto que presenta una mejor curva ROC (i.e. mayor área) y mayor precisión. Puntualmente la ANN supera en precisión en un 1,12 % relativo al SVM. Se observa que funcionalmente estos dos modelos conservaron en común las características de entrada: duración, apropiación vigente, componente producto y número de metas gestión. Solo la red neuronal incluye en este experimento de manera adicional la entrada número de metas producto.

Los modelos estructuralmente hablando no serían comparables; pero al menos en términos de la parsimonia frente al número de entradas, los clasificadores resultan similares. En el caso particular de la ANN, la red opera con cinco neuronas en la capa de entrada, dos en la capa oculta y una neurona en la capa de salida, para un total de ocho funciones no lineales. En el caso de la SVM, esta opera con un kernel lineal sobre un conjunto de 220 vectores de soporte.

Tabla VI. Resultados del primer escenario.

Modelo	*P(%)*	*1-E*	S	*ROC(AUC)*
ANN	78.261	0.247	0.833	0.789
SVM	77.391	0.264	0.917	0.751

Tabla VII. Resultados del segundo escenario.

Modelo	*P(%)*	*1-E*	*S*	*ROC(AUC)*
ANN	70.028	0.246	0.682	0.693
SVM	77.871	0.048	0.725	0.774

4.2. Segundo escenario

Para el segundo escenario se podría declarar como ganador el modelo basado en máquina de vectores de soporte puesto que presenta una mejor curva ROC (11.68 % más que ANN) y precisión frente al modelo finalista basado en ANN. Puntualmente el SVM supera en AUC en un 11,68 % y en precisión en un 11,19 % relativo a la ANN. La Tabla VII resume las mediciones de precisión, especificidad, sensibilidad y AUC para los dos mejores experimentos, mientras que la Figura 8(b) presenta las correspondientes curvas ROC.

Se observa que funcionalmente estas dos configuraciones conservaron en común las características de entrada: duración, apropiación vigente y componente producto. Solo la red neuronal incluye de manera adicional la entrada número de metas gestión. La mejor ANN encontrada presenta la misma configuración de la red neuronal obtenida en el primer escenario. Mientras que en el caso de la SVM, el kernel que mostró el mejor desempeño fue el radial con un mayor número de vectores de soporte, 1260 en total.

5. Discusión

El desarrollo metodológico presentado en este trabajo involucró la conceptualización de un megaproyecto desde el marco de referencia TOE, el cual pudo ser enlazado analíticamente a una plataforma existente de registro de proyectos de inversión como SPI-DNP. El ejercicio desarrollado reconoció atributos del SPI que se relacionan directa e indirectamente con el TOE. En este sentido, se puede evidenciar que la formulación teórica que hace el TOE de un megaproyecto es acertada en términos prácticos. Por otro lado, este ejercicio de análisis desembocó en el planteamiento de unos aspectos que configuran una intersección entre el TOE y la plataforma SPI. De allí se identificó un conjunto de características cuantitativas sobre las cuales se fundamentó un posible modelo de clasificación del avance de un proyecto dado el conocimiento a-priori de sus características.

El conjunto de características: duración, apropiación vigente y componente producto podrían ser indispensables para estudios futuros de clasificación de avance de proyectos SPI; sin embargo, es posible que se requieran más elementos SPI que permitan describir de mejor manera los proyectos en estudio, esto manteniendo como referencia el marco TOE. Una exploración de características contenidas en SPI de base lingüística podría complementar el modelo de clasificación propuesto en este trabajo.

El modelo clasificador del segundo escenario (base de datos proyectos 2008 a 2012) presentó un resultado de clasificación comparable frente al primer escenario (77,871 % frente a 78,261 % respectivamente. Diferencia relativa=0,5 %). Para estudios futuros la temporalidad de registro de los datos podría omitirse, según se propone en el segundo escenario. Podría realizarse un ajuste de los umbrales de las clases 1 (avproy$>$70 %) y -1 (avproy$<$30 %), permitiendo quizás mejorar el desempeño de clasificación y curvas ROC.

6. Conclusiones

Desde el marco referencial considerado, se presenta una primera propuesta metodológica para la clasificación de avance de proyectos mediante modelos de inteligencia computacional, y un marco de referencia de análisis de megaproyectos como es el TOE; en este sentido, el desarrollo analítico seguido en este trabajo ha permitido la identificación de características relevantes para la clasificación del avance de megaproyectos de inversión pública al menos para el caso colombiano. Aspectos como la duración del proyecto, la apropiación de recursos de inversión al igual que la naturaleza de los productos desarrollados en el proyecto serían claves para anticipar los posibles retrasos que pueda llegar a tener un megaproyecto de inversión.

Este trabajo podría ser el punto de partida para el DNP en el estudio de los proyectos que se registran en SPI dado que se planteó una propuesta de problema de clasificación de avance proyectos usando SPI-DNP como fuente de información. El núcleo del trabajo ha sido una comparación de modelos de clasificación basados en ANN y SVM, cuyas entradas se han definido con soporte en el marco TOE. En términos generales los modelos obtenidos desde estas dos perspectivas presentan indicies de desempeño similares. El espectro de posibilidades de modelado en este sentido puede llegar a ser mucho más amplio, sin embargo, desde los teoremas NFL no se puede llegar a favorecer

alguna técnica en particular de clasificación. La evidencia acumulada en este trabajo simplemente debe tomarse como un caso puntual sin pretensiones de generalización alguna en relación con el problema de la anticipación del retraso de un megaproyecto.

En cuanto a trabajo futuro, se refiere se identifican algunas posibilidades: realizar una revisión experimental de efectos sobre las curvas ROC al variar la definición de las clases desde un conjunto de expertos tratando incluir los efectos de la incertidumbre que esto acarrearía. Igualmente se propone considerar el TOE no como un marco de verificación, sino como una herramienta para la generación de política pública en cuanto al registro de megaproyectos se refiere. Esto involucraría diseñar una nueva plataforma SPI soportada conceptualmente en el TOE. Finalmente, en términos de la aplicación de la inteligencia computacional se considera abordar el problema desde técnicas que permitan incluir conocimiento experto cualitativo o lingüístico además de datos registrados.

Referencias

[1] T. Williams, "How Do Organizations Learn Lessons from Projects — And Do They?". *IEEE Transactions on Engineering Management*, vol. 55, no. 2, pp. 248–266, 2008. ↑378, 379

[2] E. Maaninen-Olsson and T. Müllern, "A contextual understanding of projects—The importance of space and time". *Scandinavian Journal of Management*, vol. 25, no. 3, pp. 327–339, 2009. ↑378

[3] T. Williams, "Assessing and Moving on From the Dominant Project Management Discourse in the Light of Project Overruns". *IEEE Transactions on Engineering Management*, vol. 52, no. 4, pp. 497–508, 2005.↑379

[4] L.E. Bohórquez, "La comprensión de las organizaciones empresariales y su ambiente como sistemas de complejidad creciente: rasgos e implicaciones". *Ingeniería*, vol.21, no. 3, pp. 363-377, 2016.↑379

[5] M. Bosch-Rekveldt, Y. Jongkind, H. Mooi, H. Bakker, and A. Verbraeck, "Grasping Project Complexity in Large Engineering Projects: The TOE (Technical, Organizational and Environmental) Framework". *International Journal of Project Management*, vol. 29, no. 6, pp. 728– 739, 2011. ↑378, 379, 380

[6] H. Jin, J. Zhao, and X. Chen, "The Application of Neuro-Fuzzy Decision Tree in Optimal Selection of Technological Innovation Projects". *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, pp. 438–443, Jul. 2007. ↑381

[7] C.-C. Huang, P.-Y. Chu, and Y.-H. Chiang, "A fuzzy AHP Application in Government-Sponsored R&D Project Selection". *The International Journal of Management Science*, vol. 36, no. 6, pp. 1038–1052, 2008. ↑381

[8] K. Khalili-Damghani, S. Sadi-Nezhad, F. H. Lotfi, and M. Tavana, "A Hybrid Fuzzy Rule-Based Multi-Criteria Framework for Sustainable Project Portfolio Selection". *Journal of Information Sciences*, vol. 220, pp. 442–462, 2013.↑381

[9] N. R. Shankar, P. P. B. Rao, S. Siresha, and K. U. Madhuri, "Critical Path Method in a Project Network using Ant Colony Optimization". *International Journal of Computational Intelligence Research*, vol. 7, no. 1, pp. 7–16, 2011. ↑381

[10] Y. Wang, "Resource-Constrained Multi-Project Scheduling Based on Ant Colony Neural Network". *The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding*, pp. 179–182, 2010. ↑381

[11] A. H. L. Chen and C.-C. Chyu, "A Memetic Algorithm for Maximizing Net Present Value in Resource-Constrained Project Scheduling Problem". *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pp. 2396–2403, Jun. 2008 ↑381

[12] M.T.Musavi, K. H. Chan, D. M. Hummels, and K. Kalantri. "On the Generalization Ability of Neural Network Classifiers". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 6, pp 659-663, 1994.↑381

[13] M. Pal and G. Foody. "Feature Selection for Classification of Hyperspectral Data by SVM". *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 48, No. 5, pp. 2297-2307, 2010↑382

[14] Departamento Nacional de Planeación. 2013. [En línea]. Disponible en: https://www.dnp.gov.co/ ↑382

[15] Seguimiento a Proyectos de Inversión (SPI). Departamento Nacional de Planeación. 2013. [En línea]. Disponible en: https://spi.dnp.gov.co/ ↑382, 383
[16] F. Costantino, G. Gravio, F. Nonino. "Project Selection in Project Portfolio Management: An Artificial Neural Network Model Based on Critical Success Factors". *International Journal of Project Management*, Vol. 33, No 8, pp 1744-1754,2015 ↑381
[17] D. Wolpert. "The Lack Of *A Priori* Distinctions Between Learning Algorithms". *Neural Computation*, Vol. 8, No. 7, pp. 1341–1390,1996.↑381
[18] D. Wolpert and W.Macready. "No Free Lunch Theorems For Optimization". *IEEE Transactions on Evolutionary Computation*, Vol. 1, No. 1, pp. 67–82, 1997. ↑381
[19] R. Duda, P. Hart and D. Stork, *Pattern Classification*, John Wiley & Sons, 2001. ↑378, 381
[20] I. Steinwart and A. Christman, *Support Vector Machines*, Springer, 2008.↑378, 381
[21] L. Haitao and Z. Xiaofu, "Introducing a New Method to Predict the Project Time Risk". *2009 International Conference on Information Management, Innovation Management and Industrial Engineering*, no. 1, pp. 27–30, 2009 ↑381
[22] S. Petruvesa, V. Zileska and V. Zujo, "Predicting construction Project Duration with Support Vector Machine". *International Journal of research in Engineering and Technology*, Vol 11, No. 2, pp. 12-24, 2013.↑381
[23] HC Yin and YS Chen". A Novel Machine Learning Model For Risk Management". *Proceedings oft he first Asia-pacific conference on global business, economics, finance and social sciences, Singapore*, August, pp. 1-15, 2014. ↑381

Hugo Ernesto Gutiérrez Vanegas

Ingeniero electrónico de la Universidad Distrital Francisco José de Caldas; coordinador Técnico de Servicios TI para Colombia en la empresa TIVIT; durante sus estudios de pregrado fue miembro del grupo de investigación Laboratorio de Automática e Inteligencia Computacional LAMIC y miembro voluntario del Instituto de Ingenieros Eléctricos y Electrónicos IEEE Colombia.
Correo electrónico: hugoegutierrezv@gmail.com

Miguel Alberto Melgarejo Rey

Ingeniero electrónico, Universidad Distrital Francisco José de Caldas; magíster en Ingeniería Electrónica y Computadores, Universidad de los Andes; doctor (c) en ingeniería, Pontificia Universidad Javeriana; miembro del grupo de investigación Laboratorio de Automática e Inteligencia Computacional LAMIC; profesor asociado, facultad de ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá DC. Senior member, Instituto de ingenieros Eléctricos y Electrónicos, IEEE.
Correo electrónico: mmelgarejo@udistrital.edu.co