TÍTULO

MACHINE METHODS LEARNING FOR THE CLASSIFICATION OF THE CORRUPTION RISK BY ADDITION IN THE PROCUREMENT OF PUBLIC WORKS THROUGH BIDDING. CASE STUDY IN COLOMBIA

MÉTODOS DE APRENDIZAJE AUTOMÁTICO PARA CLASIFICACIÓN DE RIESGO DE CORRUPCIÓN POR ADICIÓN EN CONTRATACIÓN DE OBRAS PÚBLICAS MEDIANTE LICITACIÓN. CASO DE ESTUDIO EN COLOMBIA

AUTORA

Fidelina Isabel Villa Pedroza

Esta edición electrónica ha sido realizada en 2022

Tutor	Dr. D. Antonio Javier Tallón Ballesteros
Instituciones	Universidad Internacional de Andalucía ; Universidad de Huelva
Curso	*Máster en Economía, Finanzas y Computación (2020/21)*
©	Fidelina Isabel Villa Pedroza
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	2021

Machine methods learning for the classification of the corruption risk by addition in the procurement of public works through bidding. Case study in Colombia

by

Fidelina I. Villa Pedroza

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalusia

November 2021

Métodos de aprendizaje automático para clasificación de riesgo de corrupción por adición en contratación de obra públicas mediante licitación. Caso de estudio en Colombia.

Fidelina Villa Pedroza

Máster en Economía, Finanzas y Computación

Supervisor: Antonio J. Tallón Ballesteros
Universidad de Huelva y Universidad Internacional de Andalucía

2021

Abstract

This work presents a methodology to predict the corruption risk caused by adding value to public works contracts awarded by tender in Colombia. Public data from the electronic contracts' platform were used to classify which contracts had this modification.

In terms of results, basic classification techniques such as Naïve Bayes were compared to more sophisticated algorithms such as Random Forest, trained, and validated with metrics such as precision, sensitivity, F-measure and ROC curve. The classification models achieved aimed to manage the investment of public expenditures and is an example of the use of open government data.

JEL classification H53, H57, H57, C82

Key words: Public works contracting, machine learning, SECOP II, Random Forest, Synthetic Minority Over-sampling TEchnique (SMOTE), Correlation-based Feature Selection (CFS).

Resumen

Este trabajo presenta una metodología para la predicción de riesgo de corrupción causado en adición de valor en los contratos de obra públicas adjudicados por licitación en Colombia. Se utilizaron datos públicos de la plataforma de contratos electrónicos que permitieron clasificar qué contratos tuvieron esta modificación.

En términos de resultados, se compararon las técnicas de clasificación básicas como Naïve Bayes hasta algoritmos más sofisticados como *Random Forest*, entrenados y validados con métricas como precisión, sensibilidad, medida F y curva ROC. Los modelos de clasificación obtenidos pretenden gestionar la inversión de los gastos públicos y es un ejemplo del aprovechamiento de los datos abiertos del gobierno.

Palabras claves: Contratación de obras públicas, aprendizaje automático, SECOP II, Random Forest, *Synthetic Minority Over-sampling TEchnique* (SMOTE), *Correlation-based Feature Selection* (CFS)

Tabla de Contenido

Lista de Tablas

Lista de Figuras

1 Introducción.

El gasto público es un tema que se relaciona entre el gasto del gobierno y el crecimiento económico. Dicho gasto en un país se construye con los valores monetarios que se crean de la producción de ingresos y la satisfacción de las necesidades básicas de la población suministrando bienes y servicios. El gasto público enfocado a las obras de infraestructura inciden en el crecimiento económico a largo plazo porque su propósito es incrementar la productividad del capital y el trabajo (Salazar, 2020).

Durante el 2020, el gasto público generado por los países suramericanos se establece en promedio por debajo del 45% del Producto Interior Bruto (PIB) según el banco mundial[1]. Las cifras indican que Surinam y Brasil son los países que registran la mayor participación con porcentajes de 44,58% y de 42,73% del PIB, respectivamente. Por su parte, Perú, Bolivia, Guyana, y Chile registran niveles de participación inferiores al 30% del PIB, mientras que Colombia se encuentra por encima de este porcentaje con 31,88% del PIB. Esta cifra dimensiona la cantidad de dinero que se administra para el funcionamiento y desarrollo del país invertido en diferentes rubros donde la infraestructura básica se sitúa en 1% del PIB (Melo & Ramos, 2017) y la compra pública representa el 8,6% del PIB correspondiente al 30% del gasto público (Ruiz, 2020).

La ineficiencia en el manejo de los recursos públicos según el informe del Banco Interamericano de Desarrollo (BID)[2], muestra que los países latinoamericanos y del Caribe han malgastado alrededor de 220.000 millones de dólares, equivalente al 4,4% del PIB de la región. Se hace indispensable mejorar la implementación del análisis de costo-beneficio en las elecciones presupuestales y la presencia de entidades dedicados para la planificación estratégica con evaluaciones rigurosas del impacto sobre la asignación de recursos. En la búsqueda de una distribución presupuestal eficiente que asegure la trazabilidad y la transparencia de la Gestión contractual, los sistemas electrónicos son la alternativa de diferentes países para la asignación y control de los recursos públicos. En Argentina, la

[1]https://datos.bancomundial.org/indicator/GC.XPN.TOTL.GD.ZS?end=2020&name_desc=false&start=2020&view=bar

[2] https://www.iadb.org/es/noticias/gasto-publico-en-america-latina-registra-ineficiencias-de-44-del-pib-estudio-bid

adopción de sistema COMPR.AR[3] disminuyó la duración de los procesos de adquisición, los precios pagados por los organismos públicos y aumentó el número de oferentes involucrados en los procesos de contratación pública (Michele & Pierri, 2020). En Chile, la implementación de ChileCompra[4] es un caso de éxito en la aplicación de técnicas electrónicas en la contratación pública con aporte en la creación de emprendimiento en Chile, y la plataforma ha alcanzado un ahorro 3.933 millones de dólares, frente a los 108 millones que ha costado su creación y mantenimiento (García, 2016). Otros países con sistema gestión contractual electrónica son: México con CompraNet[5], Panamá con PanamaCompra[6] y Paraguay con E-Jogua[7].

En 2015, el gobierno de Colombia lanzó el SistEma de COmpra Pública (SECOP II)[8] donde se realizan las transacciones con los recursos públicos entre las Entidades Estatales (Compradores) y los Proveedores. Así mismo, le permite al público en general consultar la actividad contractual de las Entidades Estatales. En este sistema electrónico se registraron en 2020 compras por 54,74 billones de pesos (Becerra, 2021). El tipo de contrato de obra entre 2015 y el primer trimestre del 2021 tiene asignados 16 billones de pesos, donde el 84% fue contratado mediante la modalidad de licitación.

La posibilidad de corrupción dentro de un proceso de contratación pública con respecto a la relación entre agente y cliente se atribuye a la falta de información y transparencia del comportamiento del agente contratista (Michele & Pierri, 2020). La contratación pública es atractiva para cometer actos corruptos, esto, producto de las altas cantidades de dinero destinadas para ello, las facilidades para adquirirlo y las altas probabilidades que las conductas queden impunes (Scheller & Silva, 2017). En particular, las obras de infraestructura, es el área más expuesta por los montos adjudicados a los contratos, y los

[3] La página digital COMPR.AR es una herramienta para gestión de Compras y Contratos realizados por entidades gubernamentales, integrando compradores, proveedores y la comunidad.

[4] ChileCompra es la institución que administra el sistema de compras públicas de Chile basado en la transparencia, la eficiencia, la accesibilidad y la no discriminación.

[5] CompraNet es un sistema electrónico de información pública gubernamental en contrataciones públicas.

[6] PanamaCompra una herramienta de acceso público para las compras públicas y procesos de contratación.

[7] El Sistema e-Jogua permite a las unidades Compradoras gubernamentales realizar sus adquisiciones.

[8] https://colombiacompra.gov.co/ciudadanos/preguntas-frecuentes/secop-ii

tiempos de ejecución e implementación. Entre los modelos de corrupción reconocidos, sobresale la colusión entre agente público y proveedor (Bandiera, Prat, & Valletti, 2009), las ineficiencias en la ejecución de los contratos, modificaciones unilaterales, adiciones y prórrogas arbitraria (Scheller & Silva, 2017).

Según la organización Transparencia Internacional[9] (2021) en Índice de Percepción de la Corrupción, Colombia se ubicó en la posición 92 entre 180 países con 39 puntos en una escala de 100, lo cual lo ubicó en el grupo de países con un mayor índice de corrupción en toda América. De forma detallada entre enero de 2016 y julio de 2018 Colombia vinculó 327 hechos de corrupción concentrados mayoritariamente en los sectores de Educación (16%), Infraestructura y Transporte (15%) y Salud (14%) donde la Corrupción Administrativa ocurre en el 73% de los casos, siendo el proceso de contratación pública el más afectado con un 46% sobresaliendo con irregularidades como adjudicación irregular de contratos (29%), violación a los principios de transparencia, idoneidad y responsabilidad en la contratación estatal (17%), abuso de la figura de contratación directa (8%), detrimento patrimonial por incumplimiento del objeto contratado (8%), apropiación ilegal de recursos en los contratos (6%) y sobrecostos por irregularidades en celebración de contratos (6%) (Corporación Transparencia por Colombia, 2019). Es indispensable proteger los procesos de contratación de los corruptos controlando y protegiendo la inversión pública.

Este trabajo fin de master plantea obtener con técnicas de aprendizaje automático un modelo de clasificación que determine el riesgo de corrupción por adición en contratación de obras públicas, mediante licitación en Colombia. Concretamente, el objetivo es obtener un clasificador con buena capacidad predictiva para un problema real sobre contratación pública en Colombia que contiene 1401 registros con 60 atributos, comprendidos entre 2016 al 2021-T1 (primer trimestre) (4,25 años aproximadamente) llevando a cabo como paso previo las tareas adecuadas de minería de datos para dotar a dichos datos de mayor calidad. Finalmente, con la evaluación de métricas se establece un modelo de clasificación que determine qué contratos de obras con la modalidad de licitación tiene más probabilidad de presentar adición de valor en proceso de ejecución.

[9] Organización no gubernamental que promueve medidas contra crímenes corporativos y corrupción política en el ámbito internacional.

Los beneficios que tiene prevenir los hechos de corrupción principalmente, se derivan en la inversión equitativa del presupuesto estatal para dar solución a un mayor número de problemática. De igual forma, se pueden prevenir los gastos generados en procesos judiciales para la investigación de estos casos y salvaguardar que el objetivo de la contratación sea ejecutado.

La estructura de este documento consta de cinco partes, empezando con la introducción, seguida con la contextualización del trabajo en el marco referencial. Posteriormente, se expone de forma detallada la metodología que genera la sección de resultados que son discutidos y, por último, se presentan las conclusiones obtenidas.

2 Marco Referencial.

2.1 Marco teórico.

- **Ineficiencia:** en el ejercicio de cualquier actividad el incumplimiento de forma óptima de la misma. La ineficiencia del gasto de Colombia es del 5% del PIB que tiene gran proporción en las compras públicas porque son las más proclives a estar permeadas por la corrupción y los malos manejos (Chaves, 2018). El desarrollo o retroceso de un país depende de la gestión gerencial del gasto público en infraestructura de calidad, porque esta dinamiza el comercio, potencia la economía y multiplica el bienestar social (Celina, 2016).

  Un gasto ineficiente es necesario llevarlos a un gasto eficiente; en un país como Colombia para poder contribuir al crecimiento sin agudizar la desigualdad, el derroche de los recursos públicos en sobornos y presupuestos abultados llega a aproximadamente al 26% del costo de los proyectos (Izquierdo, Pessino, & Vuletin, 2018) y las adquisiciones públicas ofrecen más oportunidades para la corrupción, exponiendo a las compras públicas a diversos riesgos de despilfarro y mala administración.

- **Contratación pública:** es descrito por el portal de información del banco Interamericano de desarrollo[10] como el proceso que una administración, a través de cualquier modalidad de contratación, obtiene el uso de bienes y/o servicios, con fines gubernamentales. Los aportes de los pueblos son la fuente monetaria para la realización de diferentes tareas de los gobernantes en sus mandatos, donde se espera honradez, transparencia, compromiso, eficiencia y efectividad. En Colombia la Contratación Estatal está regida por el Estatuto General de la Contratación Estatal de la Ley 80 de 1993[11] reglamentado y complementado por la Ley 1150 de 2007[12] de medidas para la eficiencia y la transparencia, la Ley 1437 de 2011[13] con el código de Procedimiento Administrativo y de lo Contencioso Administrativo, la Ley 1474 de 2011[14] con las normas orientadas a fortalecer los mecanismos de prevención, investigación y sanción de actos de corrupción y la efectividad del control de la gestión pública, el decreto 19 de 2012[15] que suprime o reforma regulaciones, procedimientos y trámites innecesarios existentes en la Administración Pública y el decreto 734 de 2012[16] que reglamenta el Estatuto General de Contratación de la Administración Pública (ANDI, 2013). Es importante resaltar que, desde la elaboración de los documentos que contienen las condiciones que exige el Estado a los proponentes hasta la liquidación de un contrato celebrado con el Estado, están revestidos de importantes principios que irradian todo el régimen.

- **Contrato tipo obra:** en Colombia existen varios tipos de contratos, uno de ellos es obra, según la ANDI[17] (2013) es un proceso que tiene como objeto la construcción, mantenimiento, instalación, realización de trabajos materiales sobre bienes inmuebles. El contrato de obra está determinando por el acuerdo de voluntades que regule las obligaciones pactadas, una entidad pública como contratante, la finalidad de la

[10] INTradeBID es un portal con información sobre integración y comercio de América Latina y el Caribe, desarrollado por el Banco Interamericano de Desarrollo.

[11] http://www.secretariasenado.gov.co/senado/basedoc/ley_0080_1993.html

[12] http://www.secretariasenado.gov.co/senado/basedoc/ley_1150_2007.html

[13] http://www.secretariasenado.gov.co/senado/basedoc/ley_1437_2011.html

[14] http://www.secretariasenado.gov.co/senado/basedoc/ley_1474_2011.html

[15] http://www.secretariasenado.gov.co/senado/basedoc/decreto_0019_2012.html

[16] https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=46940

[17] Asociación Nacional de Empresarios de Colombia

contratación debe ser la construcción, mantenimiento, instalación, trabajo material sobre inmuebles y acordado por la convención de un pago.

En el contrato estatal de obra, Escobar (1999) resalta que este tipo de contratos están alineados a la composición real al inmueble, de los bienes proporcionados y, al mayor o menor valor de las labores materiales, frente a otros beneficios. En la celebración del contrato de obra debe estar regido por el ordenamiento jurídico vigente, presentando las modalidades de pago como: precio global, llave en mano, precios unitarios, administración delegada y costos reembolsable (Betancourt, 2018).

Según la Guía Colombia compra eficiente (2016) las normas del nivel nacional y territorial en la ejecución de obras públicas debe contar con Licencias y obligaciones ambientales, Licencias urbanísticas y cumplimiento de los planes de ordenamiento territorial, Normas de protección del patrimonio histórico y cultural, Desarrollo territorial y manejo de temas prediales, Asuntos tributarios, Movilidad, Servicios públicos domiciliarios, Manejo de comunidades. La estructura del proceso de contratación de obra pública se compone de 3 etapas presentadas en la Figura 1.

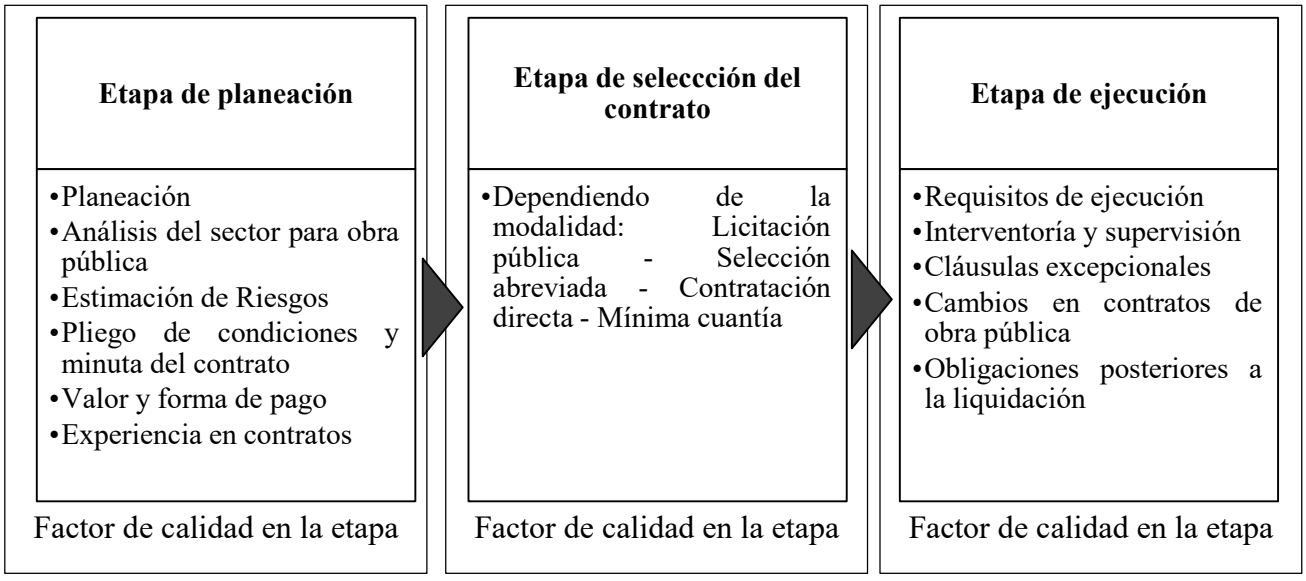


Figura 1. Marco estructural de contratación de obras públicas en Colombia. Fuente: elaboración propia.

En la planeación de la obra publica cada entidad estatal, debe tener definida la necesidad para poder ser incluida en el plan anual de adquisiciones con el respectivo estudio técnico que permita la viabilidad del proceso, dando paso al estudio del sector proveedor del contrato en forma institucional, gremial, estatal y particular. Con este análisis el estado conoce los indicadores financieros, capacidad organizacional, prevé número de proponentes y recopila información para el posterior análisis de riegos.

En la estimación del riesgo se tienen en cuenta aspectos de afectación en el ámbito territorial donde se realizará el proyecto, el monto y oportunidad de la obra, las eventuales variaciones, costos de mantenimientos, eventos que estén fuera del control de las partes, financiación del contratista, gestión predial y licitaciones. Con todas las características descritas en orden se redacta el pliego de condiciones y la minuta del contrato con su respectivo valor y forma de pago, que determina la exigencia en cuanto a la experiencia del ente que asumirá la ejecución del contrato.

Cuando se hace la selección de contratista, se debe elegir la modalidad del proceso donde se reglamentan: el convite de los interesados con la licitación pública, la selección abreviada por rango de la menor cuantía de la obra y es para el sector defensa y seguridad nacional, obras de carácter urgente o de la contratación que requiere, reserva por seguridad nacional con contratación directa y obras de costos pequeños con mínima cuantía (Colombia Compra eficiente, 2016).

Con la firma del contrato por las partes interesadas se establece la etapa de ejecución con los requisitos de iniciación: el paso de registro presupuestal, aprobación de las garantías del contrato, pago de obligaciones por el contratista y verificación de la capacidad de la interventoría y supervisión que hará la veeduría durante la ejecución de la obra. Dentro de la ejecución está la figura de las cláusulas excepcionales que evitan los paros en la obra y afectación del servicio con la interpretación unilateral, la modificación unilateral, la terminación unilateral y la caducidad.

En el momento de existir cambios sustanciales en el contrato como obras adicionales, suspender o reanudar la ejecución contractual, modificar precios con adiciones,

modificación de tiempo con prórrogas entre otras situaciones, es necesario justificar y documentar la modificación correspondiente. Por último, con la liquidación del contrato se adquieren la obligación del seguimiento de calidad, estabilidad, regulación de disposiciones o recuperación ambiental y el cierre del expediente con el cumplimiento del tiempo de garantía pactado.

Durante todo el proceso se reglamenta el factor de calidad en cada etapa compuesto por indicadores, medidas, método, especificaciones, sistema de control y características técnicas mínimas.

- **Índice de Transparencia de las Entidades Públicas ITEP:** según la procuraduría (2015) es una iniciativa de la sociedad civil que busca contribuir a la prevención de hechos de corrupción en la gestión administrativa del Estado. Este índice vigila el riesgo de la baja socialización de la información, la toma de decisiones sin reglamentación y la inoperancia de los controles de gestión. Especifica una calificación que va de 0 a 100 siendo 100 la mayor calificación con riesgo bajo y se obtiene de la evaluación de los factores de con un peso del 30% visibilidad, 40% institucionalidad, control y 30% sanción.

- ***Machine learning***: El aprendizaje estadístico con sus diversos modelos que lo conforman se agrupa en dos clases: no supervisados y supervisados. el aprendizaje no supervisado son los problemas donde se carece de una variable de respuesta que pueda supervisar el análisis y busca comprender las relaciones entre las observaciones; por el contrario, en el aprendizaje supervisado para cada observación de la medición hay una respuesta asociada y se busca ajustar un modelo que relacione la respuesta con los predictores para predecir con precisión la respuesta en el futuro. La naturaleza de la variable de salida determina los modelos para aplicar, donde una respuesta cuantitativa indica problema de regresión, mientras que una respuesta cualitativa a menudo se conoce como problema de clasificación (James, Witten, Hastie, & Tibshirani, 2014).

2.2 Revisión de la literatura

En la literatura se encuentra un amplio abanico de recursos académico generado por trabajos realizados de forma reciente sobre la creación de modelo que alerten posibles riesgos o fraudes. Díaz (2020) identifica patrones con algoritmos de *Machine Learning* para identificar casos que podrían indicar falta de competencia en contratos de vías primarias, en Colombia. En Chile, López (2019) crea un sistema de control y detección de fraudes en contratos de obras públicas basado en técnicas como regresión, arboles de decisión y SVM (*Support Vector Machine* en español máquinas de soporte vectorial).

En Brasil, existe literatura sobre la creación de un indicador de riesgo de cartel de contratación en licitaciones, relacionadas con la zona de ingeniería realizados con técnicas estadísticas y aprendizaje automático específicamente redes neuronales, SOM (*Self-Organizing Maps*) (Rosa & Zeviani, 2019). En Goiás (Brasil), se propuso con un método de aprendizaje supervisado programado en Python, un modelo para identificar sospechas de fraude en contratos de agencias públicas (Hailon & Amancio, 2020). Del mismo modo, en España hay literatura de un sistema de alerta temprana basado en un enfoque de red neuronal para predecir la corrupción en contratación pública dependiendo de las condiciones económicas de una región al momento de la predicción (López & Sanz, 2017)

En Colombia, hay modelos de aprendizaje para predecir qué contrato municipal resultará en investigaciones de corrupción por incumplimiento de contrato o ineficiencias en la implementación utilizando GBM (*Gradient Boosting*) y Lasso (Gallego, Rivero, & Martínez, 2020) y otro caso es la aplicación de herramientas de aprendizaje automático para predecir qué contratos en Bogotá, pueden presentar prórrogas o sobrecostos en su ejecución (Rodriguez, 2020).

El uso del *machine learning* en la detección del fraude tiene diferentes ámbitos de aplicación como en operaciones financieras legítimas y fraudulentas (Beltrán, 2017) y en la identificación actividades fraudulentas en transacciones de reembolso en el sector de telecomunicaciones. (Hussein & Miklos, 2011)

2.3 Marco contextual

SECOP II tiene registro en Colombia desde 2016, al primer trimestre 2021 (T1) de 763.392 contratos firmados con un valor total de 125 billones de pesos[18]. De forma general en esta plataforma electrónica se encuentra con una distribución de los contratos según su tipo de objeto de contratación, como se muestra en la Figura 2. Se trata, por tanto, de un caso de estudio con datos de un problema real.

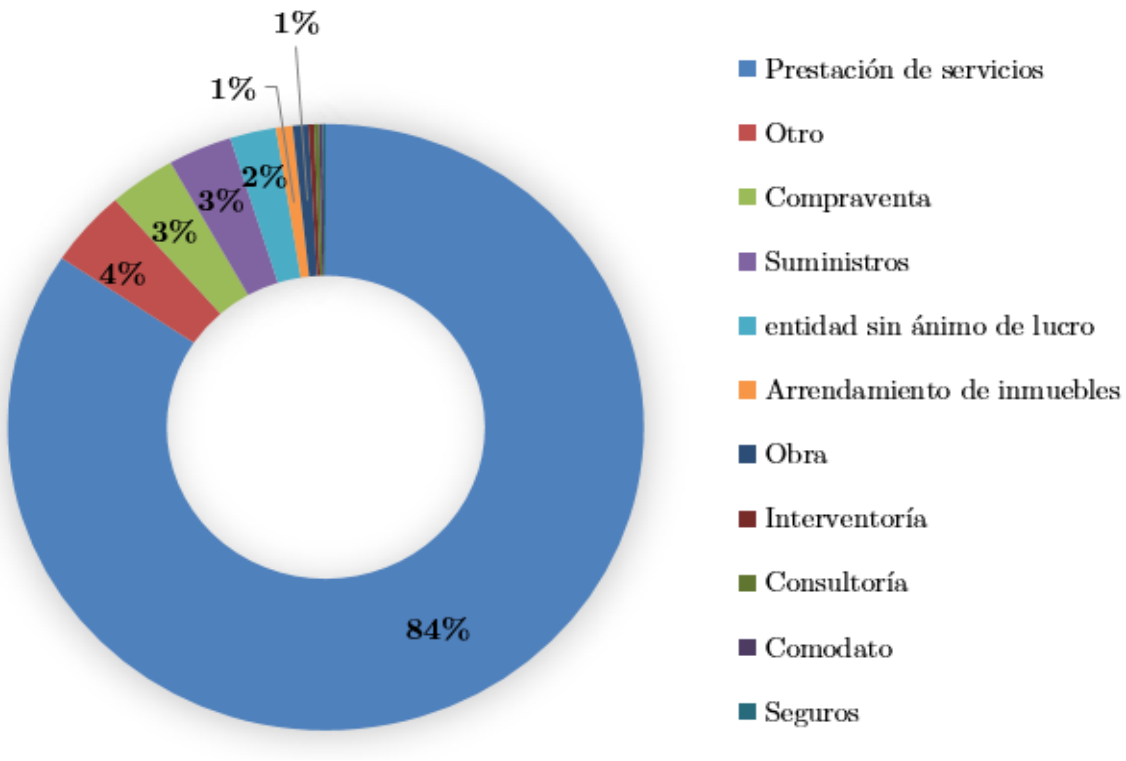


Figura 2. Tipos de contratos en SECOP II (2016-2021 T1). Fuente: elaboración propia.

En este grupo de contrato las obras públicas representan el 1% con un valor asignado de 17 billones de pesos. Donde el principal objetivo es el desarrollo eficiente de los recursos públicos para la implementación de infraestructura en el país. En la modalidad de contratación de obra la licitación es uno de los más utilizados con el 25% del total de los contratos entre 2016 y el

[18] Cálculos obtenidos de la base de datos abierta SECOP II

primer trimestre del 2021, y un valor acumulado de $14 billones de pesos, la distribución según la modalidad se muestra en la Figura 3.

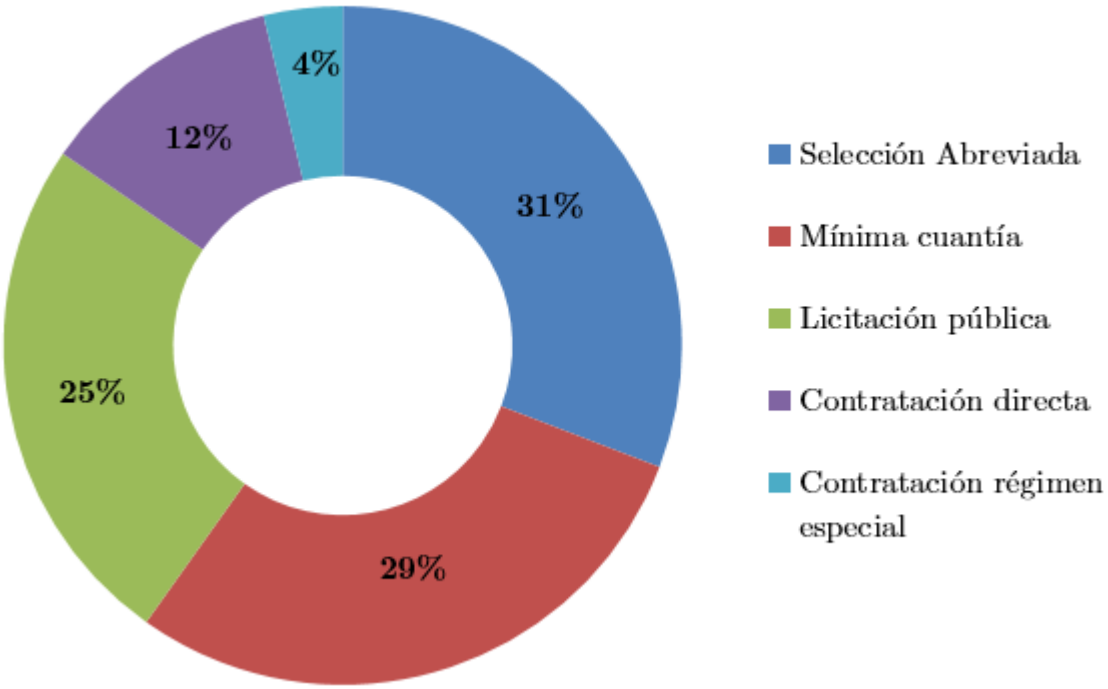


Figura 3. Modalidades de contratación en obras pública en SECOP II. Fuente: elaboración propia.

La licitación pública es un sistema de oferta dinámica, por medio de la cual se selecciona al proponente que haya ofrecido más por un bien o servicio, o de subasta inversa, es decir que se selecciona a aquel que ha ofrecido un menor valor por la ejecución del contrato (ANDI, 2013). Es importe analizar estos contratos que por sus altos costos son muy apetecidos por los diferentes proveedores de servicios, y hacer seguimiento con el objetivo de prevenir la corrupción que genera déficit económico reflejado en el ámbito político y social. Bajo esta modalidad de contratación Bogotá, es el mayor ofertante con el 79% del número de contratos, seguido por Antioquia y Caldas con un 4%.

La finalidad de las herramientas de *machine learning* es enfocarla en la creación de alerta de riesgo de posibles contratos con alza de precio, en los contratos con la figura de adiciones de valor que cambia el precio pactado inicialmente entre las partes interesadas. Los altos costos generan irregularidades en los procesos de contratación, sumadas a las pérdidas por

incumplimiento y a las condiciones de incertidumbre; estos costos extra son trasladados a un mayor precio de los bienes y servicios transados (Serrano, 2014).

3 Metodología.

Para el desarrollo de este trabajo se siguió el proceso descrito en la Figura 4. Se extrajo la información de la base de datos abierta del sistema electrónico de contratación pública de Colombia SECOP II, un portal que ofrece información oficial y actualizada del país. Los datos se obtuvieron en extensión csv con 59 atributos y una columna adicional relacionada con el índice de transparencia de cada departamento de Colombia que ofrece el contrato. En total la base de datos del desarrollo de los modelos contiene 60 atributos de interés con 1401 instancias que contiene la segmentación de los tipos de contrato “obra” y la modalidad de contrato “licitación pública”, una tabla con la descripción específica de cada atributo se muestra en la Tabla 8 (Anexos).

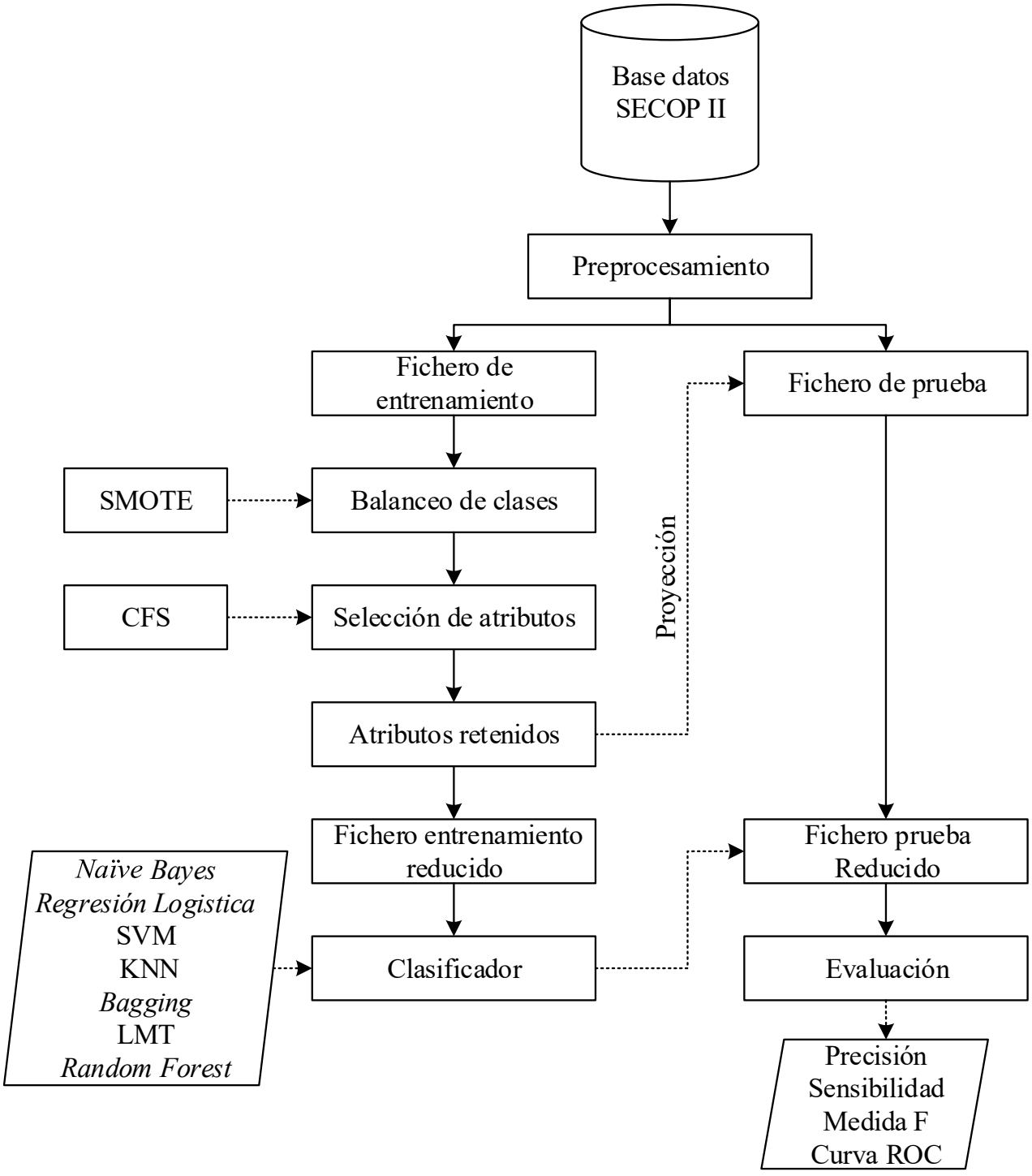


Figura 4. Metodología del trabajo. Fuente: elaboración propia.

El acotamiento de los datos está determinado entre las fechas del 2016 al primer trimestre del 2021, con 37 variables categóricas, 17 numéricas y 6 de tipo fecha. El evento que se estudia es la adición de valor a lo acordado al iniciar el contrato, esto se presenta durante la ejecución del contrato. Sin embargo, en el estudio se tiene en cuenta las variables presentes durante todas las etapas del contrato para crear un modelo de alerta temprana que contemple la información que se requiere en el transcurso de la contratación.

Por otro lado, el proveedor llamado contratado en los contratos analizados en este trabajo tiene tres figuras públicas de empresas que representan 1083 contratos de donde se generaron adición de valor en 33 de Consorcio[19] seguidos por 10 de Cooperativas[20] y 9 de Sociedad por acciones simplificada[21].

En los departamentos de Colombia contratantes, en este caso Bogotá, Caldas, Tolima y Antioquia reúnen un total de 1318 contratos, donde 60 tiene modificación por adición de valor. En la figura 5, se puede evidenciar de forma gráfica los lugares que hicieron contratación, donde el tamaño de la burbuja es proporcional al número de las ofertas, el color rojo representa los lugares donde se presentó el fenómeno que se busca predecir y el azul simboliza ausencia de adición de valor.

[19] Cuando dos o más personas en forma conjunta presentan una misma propuesta para obtener un contrato.

[20] Un mínimo de dos socios que deberán realizar la actividad corporativizada según la clase de cooperativa.

[21] Una o varias personas son responsables hasta el monto de sus respectivos aportes.

Figura 5. Mapa de departamentos con contratos tipo obra, tamaño de la burbuja referencia la cantidad de contratos y color rojo existencia de adición de valor. Fuente: elaboración propia.

El 5% del conjunto de datos de estudio presenta adición de valor, donde las tres entidades que realizan este tipo de modificación contractual son el instituto nacional de vías, la unidad administrativa especial de aeronáutica civil Aerocivil y el instituto de desarrollo urbano; estas a su vez son las tres organizaciones gubernamentales que han generado mayor número de contrato tipo obra en el periodo estudiado con 478. Como variable de respuesta se tiene una clasificación binaria donde 1 es contrato con adición y 0 con ausencia de esta modificación.

En el preprocesamiento de los datos se realizó la selección, limpieza y transformación de estos, donde de las 71 columnas originales se eliminan 12 que son utilizadas como identificador primario. Se sometió el conjunto de datos a una partición donde el 75% será el conjunto de entrenamiento para el modelo con 1050 instancias y el 25% restante del conjunto de prueba para evaluar los modelos a comparar con 351 instancias.

La base datos en estudio está constituida mayoritariamente por el 95% de contratos sin adición, donde es indispensable tratar el desbalanceo, se muestra en la Tabla 1; la cantidad de los datos componen cada clase.

Tabla 1. Distribución de los datos según la clase problema.

Contratos	Instancias	Porcentaje (%)
Sin adición	1335	95
Con adición	66	5
Total	1401	100

En los datos de entrenamiento, se realiza el balanceo que hace referencia a la distribución desigual entre la clase de interés (Hoyos, 2019). En este caso la clase minoritaria es el problema estudiado con la adición en el contrato, autores como Chawla, Japkowicz, & Kołcz (2004) y Yanminsun, Wong, & Kamel (2011) explican que el desbalanceo de datos es muy frecuente en la aplicación de clasificación en situaciones reales, lo que afecta directamente la probabilidad de que un caso sea asignado a una clase. El inconveniente del desbalanceo en los algoritmos de clasificación es por el entrenamiento con un número mayor de instancias de la clase mayoritaria, y cometen más errores cuando intenta clasificar ejemplos de la clase minoritaria (Pulgar, Rivera, Charte, & Jesus, 2018).

El sesgo generado por la presencia mayoritaria de la clase sin adición de valor en el contrato se aborda con el método de sobremuestreo*, Synthetic Minority Oversampling Technique* (SMOTE) donde Ha & Bunke (1997) fueron pioneros al introducir datos adicionales de entrenamiento entre los datos originales. Con este método, en el conjunto de entrenamiento se crean aleatoriamente instancias sintéticas de la clase minoritaria teniendo en cuenta los vecinos más cercanos (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Se debe tener en cuenta que el problema de este método es considerar la pérdida de rendimiento en la clasificación, dado que los datos no provienen de la distribución original (Chawla, Japkowicz, & Kołcz, 2004). En este trabajo se aplicó SMOTE con un K igual 1 y un porcentaje igual a 1000, cuyo resultado se compara en la Tabla 2;

posteriormente, se compara la aplicación de los modelos de clasificación en los datos originales y en los datos resultantes en la aplicación de este método.

Tabla 2. Distribución de los datos de entrenamiento según la clase problema y el método SMOTE.

Contratos	Fichero de entrenamiento original	Porcentaje (%)	Fichero de entrenamiento SMOTE	Porcentaje (%)
Sin adición	1001	95	1001	65
Con adición	49	5	539	35
Total	1050	100	1540	100

Con el conjunto de entrenamiento con SMOTE, se busca aplicar selección de atributos para llegar a un subconjunto de atributos, del conjunto original, buscando los atributos relevantes para una aplicación, logrando el máximo rendimiento y la comprensión de los resultados (Toledo, 2016). Sobre los efectos negativos que genera el tener un sinfín de atributos en modelo, Ruiz (2006) explica que son: tener mayor tiempo de ejecución por variabilidad estadística entre patrones de diferente clase, confusión en los algoritmos de aprendizaje por atributos irrelevantes y redundantes, el clasificador resultante es complejo y el alto costo de oportunidad y almacenamiento en la recopilación de información futura.

Para la creación de este trabajo se utiliza la selección de atributo tipo filtro, que es una aproximación indirecta planteada por Ben-Bassat en 1982 y usan heurísticos para determinar el subconjunto de atributos óptimo, dando como resultado rapidez en los cálculos (Armañanzas, 2004). Esta técnica se aplica con la correlación basada en la selección de características (CFS, *Correlation-based Feature Selection)* fue creado por Hall (1999) donde prima la incertidumbre simétrica, el atributo seleccionado debe tener una alta correlación con la clase y no debe estar correlacionada con otro atributo. El algoritmo de CFS se basa en una medida de bondad de la correlación de un atributo con la clase, definida como:

$$Evaluación\ (A_i) = \frac{k\bar{r}_{ci}}{\sqrt{k + k(k-1)\bar{r}_{ii'}}} \quad (1)$$

Donde A_i es el atributo evaluado, k es el número de atributos en el subconjunto, $\bar{r}_{ci}$ es la correlación media con la clase, y $\bar{r}_{ii'}$ es la intercorrelación media entre ellos (Khoirunnisaa, Pane, Wibawa, & Purnomo, 2018). La principal ventaja de este método es la rapidez. Sin embargo, aunque elimina variables redundantes, puede eliminar atributos que, en conjunto con otros pueden estar correlacionados con la clase (Toledo, 2016). Al aplicar este método en el conjunto de entrenamiento selecciono 15 atributos mostrados de forma específica en la Tabla 3. Para obtener el conjunto de prueba reducido se hizo la proyección de los atributos seleccionados sobre el conjunto de entrenamiento, que consiste en considerar únicamente para el conjunto de prueba los atributos seleccionados mediante el filtro basado en subconjuntos de atributos CFS.

Tabla 3. Atributos seleccionados por el método CFS.

Fecha de Fin del Contrato	ITEP	Modalidad de Contratación
Plazo contrato	Documento Proveedor	Municipio 2
Obligación ambiental	Valor Pendiente de Ejecución	Saldo CDP
Destino Gasto	Origen de los recursos	Número categorías Entidad
Contratos por proveedor	Saldo Vigencia	Habilitado pago adelantado

En el proceso de aplicación de los algoritmos de clasificación se hace en el conjunto de prueba original, con SMOTE y con selección de atributos, los algoritmos a comparar son: Naïve Bayes, regresión logística, SVM, KNN, *Bagging*, LMT y *Random Forest*. Conociendo que la finalidad de cada uno de los modelos es la predicción de valor de la variable salida Y a partir de valores de las variables de entrada X, tales que Y=f(X) + e, donde Y sea predicha con el mínimo error. A continuación, se procede a describir los algoritmos de aprendizaje automático supervisado que se usan en este TFM.

Naïve Bayes: es un modelo que asigna la probabilidad de la instancia nunca antes vista, perteneciente a cada clase, luego simplemente elije la clase más probable, de acuerdo a la probabilidad condicionada $Py_i(xi) = Pr\ (Y=y_i \mid X=x_i)$ donde la probabilidad que una muestra dada por x ($X=x_i$) sea de la clase y ($Y=y_i$) (James, Witten, Hastie, & Tibshirani, 2014). Usa la construcción de probabilidades Bayesianas para la construcción del modelo, partiendo de la independencia entre los predictores con la notación

$$P(y_i|x_i) = \frac{P(x_i|y_i).P(\ y_i)}{P\ (x_i)} \tag{2}$$

Se busca calcular $P(y_i|x_i)$, que es la probabilidad de que la instancia x esté en la clase y; $P(x_i|y_i)$, que es la probabilidad de generar la instancia x dada la clase y; $P(\ y_i)$, que es la probabilidad de ocurrencia de la clase y; $P\ (x_i)$ que es la probabilidad de que ocurra la instancia x siendo el mismo para todas las clases (Berrar, 2019).

Regresión logística: es un método de clasificación que estima valores directos binarios con una variable dependiente continua Y, que se distribuye normalmente en la población con una relación no lineal entre las variables. El planteamiento de este método estadístico explicado por Cox y Snell (1989) donde la probabilidad de pertenecer a una clase es generada por la distribución logística:

$$\pi = \frac{1}{1 + \ e^{-wx_i}} \tag{3}$$

En el modelo está implicado un vector de peso w relacionado con las variables explicativas y una x_i que maximizan la función de verosimilitud:

$$l = \prod_{i=1}^{N} p_i{}^{y_i}\ (1 - p_i)^{1-y_i} \tag{4}$$

Donde i es cada observación y N el total de la muestra, el objetivo es minimizar la desviación total, asignándole una probabilidad alta a los contratos con adición y una probabilidad baja a los contratos sin modificaciones (Robles, Cortés, & Barbadill, 2020). En los parámetros establecidos

el número máximo de iteraciones es 500, la heurística para la detención codiciosa mientras se realiza la validación cruzada del número de iteraciones es 50.

Máquina de soporte vectorial: es un modelo de clasificación creado por Vladimir Vapnik, donde el objetivo es predecir valores de la variable de salida. Los datos de entrada son tratados como un vector de dimensión p, para buscar un hiperplano que separe de forma óptima los puntos pertenecientes de una clase con respecto a otra, que han sido proyectados previamente en un espacio con mayor dimensionalidad. La separación óptima es lo que define a un modelo SVM, que se encarga de buscar aquel hiperplano que tenga la máxima distancia con los puntos más cercanos a él, y separando los puntos del vector que están a un lado y otro del hiperplano, para etiquetarlos con categorías distintas según su localización (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997).

Este modelo está formado por un *kernel* K (x_i, x_i') que cuantifica la similitud entre las observaciones. Dependiendo del *kernel* utilizado el procedimiento equivale a ajustar un clasificador de vector soporte en un espacio de mayor dimensión con una función de decisión establecida:

$$f(x) = \beta_0 + \sum_{ies} a_i K(x, x_i) \tag{5}$$

Dentro de los tipos de *Kernel* que se pueden utilizar para este propósito, las más comunes son: el polinómico que genera una frontera decisión no lineal más flexible siguiendo la ecuación (6) y la Gaussiana o radial donde el peso de los vectores soportes en evaluación decrece exponencialmente de acuerdo a su distancia euclídea elevada al cuadrado, ecuación (7) (James, Witten, Hastie, & Tibshirani, 2014).

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij}\, x_{i'j}\right)^d \tag{6}$$

$$K(x_i, x_{i'}) = exp\left(-\gamma + \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right) \tag{7}$$

Los parámetros del modelo se establecen con el número de pliegues para la validación cruzada en -1 significa usar datos de entrenamiento, los valores de la semilla han variado entre 1 y 30, el parámetro de complejidad C es 1, en *kernel* se usa *polykernel* y *RBFkernel*, los datos se normalizan, el parámetro de tolerancia es 0.001, para una estimación de probabilidad adecuada, se ajustan los modelos logísticos a las salidas y la épsilon para el error de redondeo es 10-12.

K-vecinos más cercanos (KNN): una vez localizados aquellos k vecinos para cada punto en los datos, se establece un sistema de votación, donde se considera el valor que toma la variable independiente para cada uno de los puntos y se devuelve como predicción el valor medio de dichos valores (Germán Morales, 2008). Para llegar a estas predicciones es muy importante seleccionar bien el valor de k, es decir el número de vecinos cercanos que vamos a utilizar, ya que esto determina a que grupo pertenece cada punto, con la ecuación:

$$d_j(x_0) = P(Y = j|X = x_0) = \frac{1}{K}\sum_{i\epsilon N_0} I(y_i, j) \qquad (8)$$

KNN es un modelo donde el aprendizaje se hace en las instancias de prueba, con el fichero de entrenamiento se determina el valor del hiperparámetro K y es un modelo no paramétrico que no asumen distribución que siguen los datos (James, Witten, Hastie, & Tibshirani, 2014). La forma utilizada en este trabajo para medir la cercanía entre los puntos es la distancia Euclidiana, con un numero de vecino óptimo de 5 determinado con validación cruzada en el conjunto de entrenamiento.

Bagging: es una técnica consiste en crear diferentes modelos, usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados, el cual fue desarrollado por Breiman (1996) donde plantea:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}\,(x) \qquad (9)$$

Usando B conjuntos de entrenamiento separados (árboles), y promedio los mismos para obtener un único modelo de aprendizaje estadístico de baja varianza (James, Witten, Hastie, &

Tibshirani, 2014). Este es un modelo que realiza remuestreo con reemplazo para minimizar la varianza, reduce el ruido de los valores atípicos y no mejora significativamente las funciones lineales (Santana, 2014). Los parámetros establecidos son: la semilla aleatoria entre 1 a 30, el número de ranuras de ejecución (subprocesos) a usar para construir el conjunto es 1, el tamaño de cada bolsa es 100 como porcentaje del tamaño del conjunto de entrenamiento y el número de iteraciones es 10.

LMT, Logistic Model Trees: proporciona una descripción muy buena de los datos. Consiste en una estructura de un árbol de decisión con funciones de regresión logística en las hojas. Como en los árboles de decisión ordinarios, una prueba sobre uno de los atributos es asociado con cada nodo interno (Charris, y otros, 2018). El árbol está basado en el modelo logístico que combina métodos de aprendizaje basados en el crecimiento de un árbol de clasificación estándar y la construcción de modelos de regresión para todos los nodos (Niels, Mark, & Eibe, 2003). LMT utiliza validación cruzada para encontrar una serie de iteraciones, para evitar el sobreajuste de datos de entrenamiento y el método de regresión logística lineal se usa para calcular las probabilidades posteriores de las hojas en el modelo LMT (García J. , 2020).

El LMT utiliza la validación cruzada para encontrar una serie de iteraciones de *LogitBoost* para evitar el sobreajuste de los datos de entrenamiento con mínimos cuadrados para cada clase Mi.

$$L_M(x) = \sum_{i=1}^{n} \beta_i \, x_i + \beta_0 \tag{10}$$

Donde βi es el coeficiente del i-ésimo componente del vector x, mientras que n es el número de factores. El método de regresión logística se utiliza para calcular las probabilidades posteriores de los nodos de hojas en el modelo LMT, donde D es el número de clases. (Chen, y otros, 2017)

$$P(M|x) = \frac{\exp(L_M(x))}{\sum_{M'=1}^{D} exp\,(L_{M'}(x))} \tag{11}$$

En el desarrollo del modelo se establece como parámetro, que el mínimo de instancias sea 15 validándose de forma cruzada y una sola vez para usar ese número en cada nodo del árbol, con el fin de mejorar el tiempo de ejecución.

Random Forest (RF): es un algoritmo predictivo que usa la técnica de *Bagging* para combinar diferentes árboles, donde cada árbol es construido con observaciones y variables aleatorias (Santana, 2014). Este método de aprendizaje conjunto propuesto por Breiman (2001) la predicción de RF se considera la mayoría no ponderada de los votos de la clase, utilizando la técnica de *bagging* para seleccionar muestras aleatorias de variables como el conjunto de datos de entrenamiento para la calibración del modelo. Para cada variable, la función determina el error de predicción del modelo y si los valores de esa variable se permutan a través de las observaciones fuera de la bolsa (Chen, y otros, 2017)

Al construir estos árboles de decisión se considera una división en un árbol, una muestra aleatoria de m predictores se eligen candidatos divididos del conjunto completo de p predictores. La división puede usar solo uno de esos m predictores (James, Witten, Hastie, & Tibshirani, 2014).

Las ventajas del *Random Forest* son: una excelente estimación en datos perdidos manteniendo la exactitud, tiene una eficiencia alta de detección de las interacciones de las variables (Breiman & Cutler, 2005), y en conjunto de datos de tamaño grande es un clasificador eficaz (Caruana, Karampatziakis, & Yessenalina, 2008). Sin embargo, entre sus desventajas se resalta el sobreajuste en ciertos grupos de datos ruidosos (Segal, 2014), tiene una salida mayor elaborada (Berthold, 2014), el método en variables categóricas con diferentes números de niveles parcializa favoreciendo a atributos con más niveles (Deng, Runger, & Tuv, 2011) y ayuda a grupos más pequeños con atributos correlacionados con similar relevancia para el rendimiento (Toloşi & Lengauer, 2011). Para la ejecución del algoritmo, se ha experimentado con 30 semillas diferentes y así promediar los resultados haciendo que se obtenga un rendimiento medio, evitando el escenario más optimista o más pesimista que se puede obtener si se prueba con una única semilla o bien con un reducido de ellas; los valores de la semilla han variado entre 1 y 30, el número de árboles generado es 100 y el número de ranuras de ejecución (subprocesos) que se utilizarán para construir el conjunto es 1.

Con el planteamiento de los clasificadores a comparar en el conjunto de entrenamiento se determinaron parámetros necesarios para aplicarlos en cada modelo, en los datos de prueba que tiene como medidores o evaluadores: la precisión, la sensibilidad, medida F y curva ROC (*Receiver Operating Characteristic Curve*).

En la compresión de esta métrica resulta de la matriz de confusión Figura 5. donde se visualiza y mide la capacidad de los algoritmos para clasificar un nuevo registro entre las dos clases establecidas. Los verdaderos positivos (VP) y verdaderos negativos (VN) representan la proporción de los registros del fichero de prueba que quedan perfectamente clasificados, mientras que los falsos positivos (FP) y los falsos negativos (FN) representan los casos clasificados erróneamente.

Tabla 4. Matriz de confusión.

		Clase predicha		
		Negativo	**Positivo**	
Clase real	**Negativo**	VN	FP	
	Positivo	FN	VP	**Precisión** $\frac{VP}{FP + VP}$
		Sensibilidad $\frac{VP}{FN + VP}$		

Precisión: es una métrica que mide la proporción de predicciones positivas correctas. Donde se tiene en cuenta los verdaderos positivos (VP) entre la suma de los falsos positivos (FP) y verdaderos positivos (VP).

$$\text{Precisión} = \frac{VP}{FP + VP} \tag{12}$$

Sensibilidad: es una métrica que mide la proporción de casos positivos detectados. Donde se tiene en cuenta los verdaderos positivos (VP) entre la suma de los falsos negativos (FP) y verdaderos positivos (VP).

$$\text{Sensibilidad} = \frac{VP}{FN + VP} \tag{13}$$

Medida F: es una métrica que permite comparar dos modelos de baja precisión y alta sensibilidad, utilizando la media armónica para castigar los valores extremos.

$$\text{Medida F} = \frac{2 * sensibilidad * Precisión}{sensibilidad * Precisión} \tag{14}$$

Curva ROC: mide de forma agregada el desempeño del modelo sobre todos los umbrales de clasificación al comparar la sensibilidad y la precisión, Figura 6. El valor de esta métrica está entre 0 y 1 donde un 90% al 100% es un clasificador excelente, 80% al 90% es un clasificador bueno, 70% al 80% es un clasificador aceptable, 60% al 70% es un clasificador pobre y 50% al 60% es un clasificador malo.

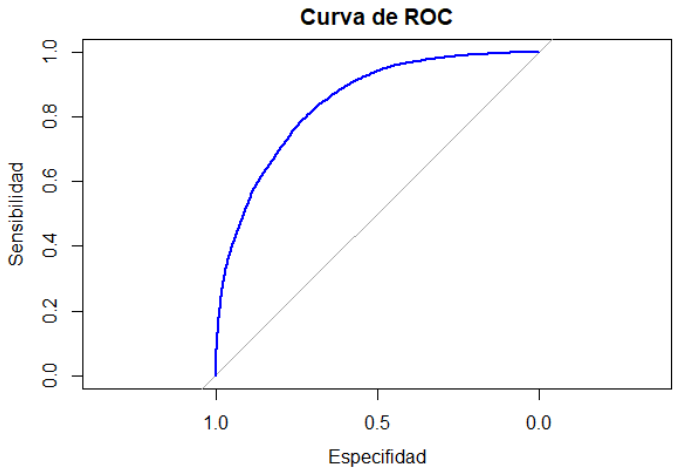


Figura 6. Representación gráfica de la curva ROC. Fuente: elaboración propia.

En este trabajo se han aplicado estos algoritmos utilizando como *software* Python 3.8 (Singh, 18) utilizando *Jupiter-Notebook* y la biblioteca Weka[22], en un ordenador con sistema operativo *Windows 10*, un procesador Intel(R) Core(TM) i5-6200U CPU 2.40 GHz y 4 GB de memoria RAM.

4 Resultados.

Con el seguimiento de la metodología descrita en la sección 3, se obtiene el cálculo de las distintas métricas en cada escenario, donde el fichero de entrenamiento no ha sido modificado para la primera batería de experimentos (escenario I), el fichero de entrenamiento es tratado el desbalanceo (escenario II) y en los ficheros hay selección de atributos (escenario III). La medición de la clase positiva, (predicción) en el escenario I se muestra en Tabla 5, donde los mejores modelos son RF y LMT con el mismo rendimiento en la métrica. Estos modelos solo difieren en la curva ROC donde el RF muestra superioridad. En importante resaltar que la variabilidad en los modelos estocásticos que cuentan como semillas aleatorias es relativamente baja.

[22] Código y datos en GitHub. Enlace: https://github.com/fisavilla/WekaInPython.git

Tabla 5. Promedio de resultados de precisión sobre el conjunto de prueba con los datos originales (Escenario I).

Datos brutos		SVM Kernel Polinómico	SVM Kernel RBF	Bagging	Random Forrest	Naïve Bayes	Regresión logística	KNN	LMT
Precisión	μ	94.0	94.2	90.5	95.9	94.6	94.3	94.2	95.9
	σ	0.001	0.012	0.010	0.007				
Sensibilidad	μ	94.9	95.2	94.6	95.7	35.0	95.4	95.4	95.7
	σ	0.001	0.001	0.002	0.001				
Medida F	μ	94.3	94.6	92.5	94.1	46.7	94.1	94.2	94.1
	σ	0.001	0.003	0.083	0.001				
ROC	μ	71.8	82.9	57.2	73.9	72.9	55.9	74.3	55.9
	σ	0.000	0.017	0.010	0.022				

En la base de datos sometida a balanceo la medición de la clase positiva (predicción) en Tabla 6, el mejor modelo es Naïve Bayes. En términos de sensibilidad y medida F, *Random Forest* repunta. De la misma forma los modelos estocásticos tienen una variabilidad baja entre los resultados obtenidos con las diferentes semillas.

Tabla 6. Resultados individuales de los datos con preprocesamiento mediante SMOTE (Escenario II).

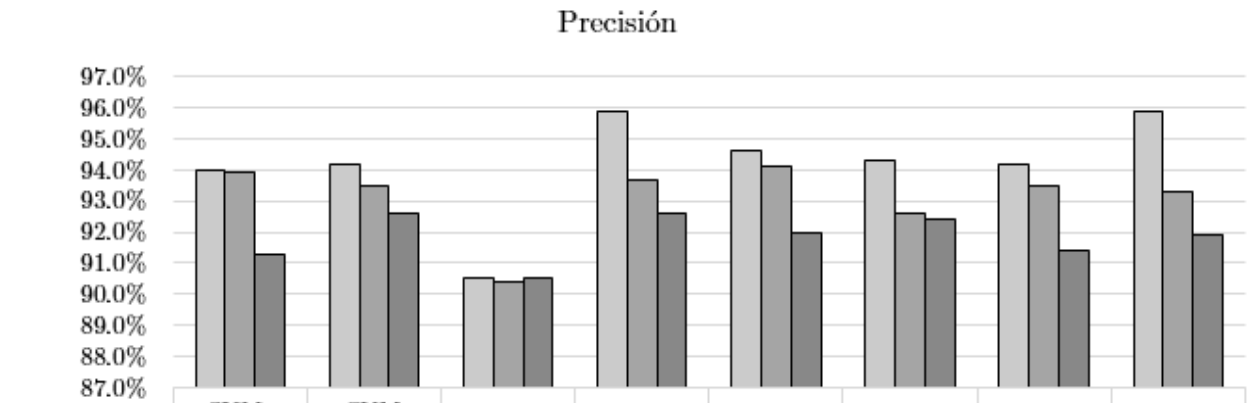
SMOTE		SVM Kernel Polinómico	SVM Kernel RBF	Bagging	Random Forest	Naïve Bayes	Regresión logística	KNN	LMT
Precisión	μ	93.9	93.5	90.4	93.7	94.1	92.6	93.5	93.3
	σ	0.001	0.002	0.005	0.010				
Sensibilidad	μ	93.2	94.6	92.9	95.2	43.0	94.6	93.2	94.3
	σ	0.001	0.001	0.003	0.028				
Medida F	μ	93.5	93.9	91.6	94.0	55.5	93.3	93.3	93.7
	σ	0.001	0.000	0.003	0.016				
ROC	μ	61.0	79.7	57.6	77.7	71.4	66.1	74.1	66.5
	σ	0.743	0.009	0.006	0.085				

Por último, con la base de datos con rebalanceo se hace la selección de atributos donde la medición de la clase positiva (predicción) en Tabla 7, resulta que el mejor modelo es *Random Forest.* En término de sensibilidad y curva ROC, LMT repunta y la regresión logística fundamento del LMT obtiene el mejor resultado en la medida F. De igual manera los modelos estocásticos ejecutados tienen una baja variabilidad.

Tabla 7. Resultados individuales de los datos con CFS a partir de los datos obtenidos por SMOTE (Escenario III).

CFS		SVM Kernel Polinómico	SVM Kernel RBF	Bagging	Random Forrest	Naïve Bayes	Regresión logística	KNN	LMT
Precisión	**μ**	91.3	92.6	90.5	**92.6**	92.0	92.4	91.4	91.9
	σ	0.003	0.001	0.004	0.003				
Sensibilidad	**μ**	93.4	90.6	93.2	93.7	33.9	94.3	90.9	**93.4**
	σ	0.003	0.007	0.004	0.006				
Medida F	**μ**	92.3	91.5	91.8	93.1	45.9	**93.2**	91.1	92.6
	σ	0.007	0.004	0.003	0.003				
ROC	**μ**	74.0	72.1	57.6	72.2	66.9	72.7	67.4	**74.2**
	σ	0.004	0.001	0.040	0.017				

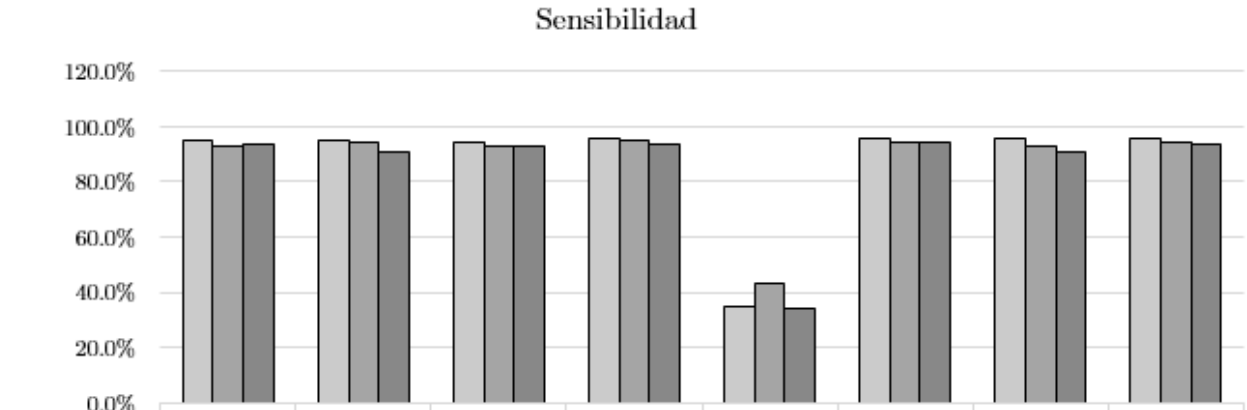
Al entrar a comparar los escenarios planteados, la métrica de la precisión juega un papel primordial porque con ella se tiene un indicador sobre la clasificación correcta de los contratos, que presentaran adición de valor. Dos árboles de decisión son los mejores modelos, comenzando por el *Random Forest* y seguido por el LMT, Figura 7. Donde el resultado superior indica que de 100 contratos 96 son clasificados correctamente, lo cual es un punto de partida para las organizaciones gubernamentales a la hora de analizar los riesgos de corrupción o ineficiencia en el dinero público. En los resultados, ninguno de los modelos presentó mejora con la aplicación de SMOTE y CFS. Esto significa que el desempeño del modelo no está afectado con la proporción de contratos que tiene o no presencia de adición de valor.



	SVM Kernel Polinomico	SVM Kernel RBF	Bagging	Random Forest	Naive Bayes	Regresión logistica	KNN	LMT
ORIGINAL	94.0%	94.2%	90.5%	95.9%	94.6%	94.3%	94.2%	95.9%
SMOTE	93.9%	93.5%	90.4%	93.7%	94.1%	92.6%	93.5%	93.3%
CFS	91.3%	92.6%	90.5%	92.6%	92.0%	92.4%	91.4%	91.9%

Figura 7. Resultados de los modelos con respecto a la métrica de precisión. Fuente: elaboración propia.

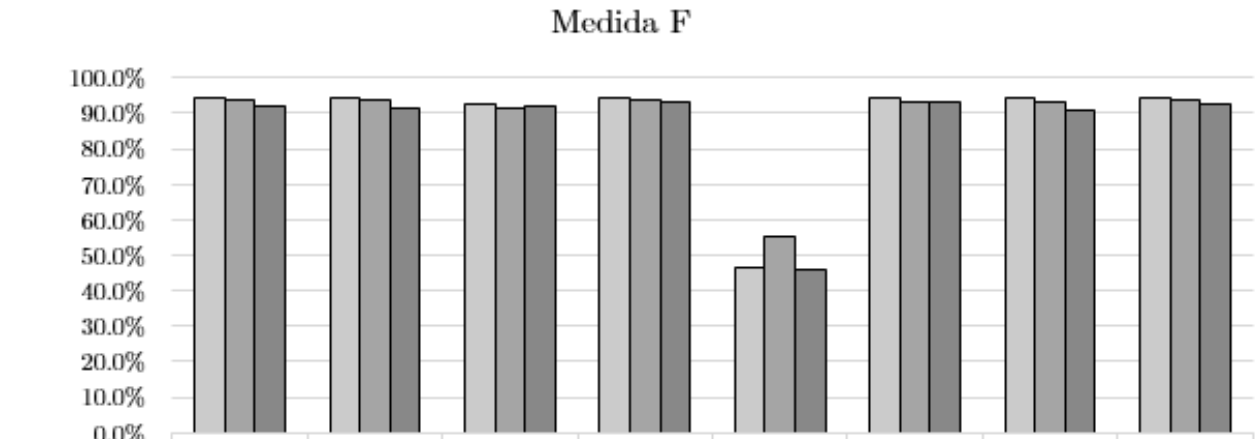
La métrica de sensibilidad presenta unos resultados excelentes en los modelos de los diferentes escenarios a excepción, de *Naïve Bayes* porque todos tiene valores superiores a 90%, figura 8. Donde al igual que en la precisión el mejor resultado lo obtuvo *Random Forest* en la base datos original, con una predicción del 95.9% de los positivos. En esta medida todos los modelos presentaron mejora en el resultado después del tratamiento del balanceo y selección de atributos frente al modelo original, pero en término de relevancia la predicción es la medida más importante.



	SVM Kernel Polinomico	SVM Kernel RBF	Bagging	Random Forest	Naive Bayes	Regresión logistica	KNN	LMT
ORIGINAL	94.9%	95.2%	94.6%	95.7%	35.0%	95.4%	95.4%	95.7%
SMOTE	93.2%	94.6%	92.9%	95.2%	43.0%	94.6%	93.2%	94.3%
CFS	93.4%	90.6%	93.2%	93.7%	33.9%	94.3%	90.9%	93.4%

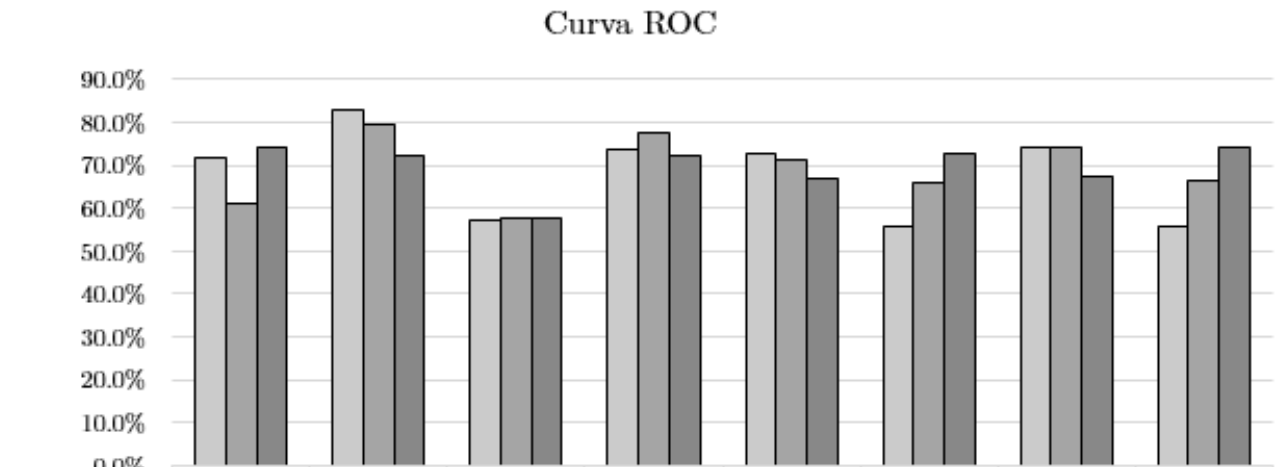
Figura 8. Resultados de los modelos con respecto a la métrica de sensibilidad. Fuente: elaboración propia.

La medida F, combina la precisión y sensibilidad en una sola métrica donde su utilidad es resaltar conjuntos de datos con clases desiguales. De forma general, los modelos tienen un resultado óptimo excluyendo Naïve Bayes con resultado pésimo por debajo del 50%, Figura 9. Por último, la curva ROC que genera la comparación de la sensibilidad y la precisión, Figura 10. En el valor de estas métricas SVM con *kernel* radial, seguido por KNN y *Random Forest* son los mejores modelos donde el resultado del clasificador es bueno y aceptable respectivamente en cada uno. En las Figuras 11, 12, 13 (Anexos). Se muestra las respectivas graficas obtenidas.



	SVM Kernel Polinomico	SVM Kernel RBF	Bagging	Random Forest	Naive Bayes	Regresión logistica	KNN	LMT
ORIGINAL	94.3%	94.6%	92.5%	94.1%	46.7%	94.1%	94.2%	94.1%
SMOTE	93.5%	93.9%	91.6%	94.0%	55.5%	93.3%	93.3%	93.7%
CFS	92.3%	91.5%	91.8%	93.1%	45.9%	93.2%	91.1%	92.6%

Figura 9. Resultados de los modelos con respecto a la medida F. Fuente: Elaboración propia.



	SVM Kernel Polinomico	SVM Kernel RBF	Bagging	Random Forest	Naive Bayes	Regresión logistica	KNN	LMT
ORIGINAL	71.8%	82.9%	57.2%	73.9%	72.9%	55.9%	74.3%	55.9%
SMOTE	61.0%	79.7%	57.6%	77.7%	71.4%	66.1%	74.1%	66.5%
CFS	74.0%	72.1%	57.6%	72.2%	66.9%	72.7%	67.4%	74.2%

Figura 10. Resultados de los modelos con respecto a la curva ROC. Fuente: elaboración propia.

5 Conclusiones.

En cuanto a lo abordado con anterioridad se logró plantear con técnicas de aprendizaje automático, la clasificación de contratos con riesgo de ineficiencia por adición de valor en obras públicas adjudicadas por licitación en Colombia. Basado en datos reales en el periodo (2016 - 2021T1), el planteamiento integró el aprendizaje con datos históricos, identificación de patrones, toma de decisiones con mínima intervención humana y elección del mejor predictor con la evaluación de métricas establecidas.

Dentro del análisis expuesto, *Random Forest* fue el modelo con mejores métricas en general, donde el resultado indica que de 100 contratos 96 son clasificados correctamente, ayudando así a las organizaciones gubernamentales a la hora de analizar los riesgos de corrupción o ineficiencia con el dinero público. El cual tiene beneficios derivados en la inversión equitativa del presupuesto estatal para dar solución a un mayor número de problemáticas. De igual forma, se puede prevenir los gastos generados en procesos judiciales para la investigación de estos casos y salvaguardar que el objetivo de la contratación sea ejecutado.

La precisión de las diferentes técnicas predictivas fue la métrica determinante a la hora de elegir el mejor algoritmo. Los tratamientos utilizados para el balanceo de las clases y la selección de atributos no genero incremento indicando que el desempeño del modelo no está afectado con la proporción de contratos que tiene o no presencia de adición de valor, caso contrario de la métrica de sensibilidad que mostró valores muy cercanos a la base original.

Dentro del análisis expuesto, *Bagging* fue el peor modelo en ajustarse a diferencia del algoritmo de Random Forest generado a partir de este. Modelos como regresión logística y LMT que tiene la misma relación de fundamento que tuvieron un comportamiento sin diferencia abismal. Por otro lado, en la máquina de soporte vectorial el *kernel* RBF generó un desempeño superior al polinómico.

La integración de la información generada por las bases de datos de la plataforma SECOP II y ITEP junto con un análisis preliminar, fue indispensable en el preprocesamiento donde la selección, limpieza y transformación generó un conjunto de datos con características fiables para la ejecución y posterior evaluación de los modelos. Se hizo uso de dos técnicas de preparación de datos relativas a dos líneas diferentes, como son la presencia de datos con desbalanceo en sus

clases, así como la selección de atributos. Hay que tener en cuenta que se aplicaron ambas secuencialmente, que no es demasiado habitual entre la comunidad de investigadores del área de minería de datos, lo cual da una idea de la complejidad del problema real abordado.

Referencias

Abdel-Sattar, M., & Aboukarima, A. M. (2021). Performance of the different kernel functions of SVR in the validation dataset with the default quantities used in the Weka software. *PLOS ONE*.

ANDI. (2013). *Asociación Nacional de Empresarios de Colombia*. Obtenido de Contratación Estatal: http://proyectos.andi.com.co/es/GAI/GuiInv/ConEst/ConEst/Paginas/default.aspx

Armañanzas, R. (2004). *Medidas de filtrado de selección de variables mediante la plataforma "Elvira".* paises vasco: Universidad de paises vasco.

Asociación Nacional de Empresarios de Colombia . (2013). *Contratación Estatal*. Obtenido de Asociación Nacional de Empresarios de Colombia : http://proyectos.andi.com.co/es/GAI/GuiInv/ConEst/ConEst/Paginas/TipCon.aspx

Bandiera, O., Prat, A., & Valletti, T. (2009). Active and Passive Waste in Government Spending: Evidence from a Policy Experiment. *American Economic Review*, 1278-1308.

Becerra, L. (13 de enero de 2021). Transacciones por Secop II en 2020 llegaron a $54 billones. *portafolio*. Obtenido de https://www.portafolio.co/economia/transacciones-por-secop-ii-en-2020-llegaron-a-54-billones-548220

Beltrán, I. (2017). Comparación de métodos de detección del fradude en operaciones de banca móvil. TFM. Huelva: Universidad de Huelva & Universida internacional de Andalucia.

Berrar, D. (2019). Bayes' theorm and Naïve Bayes classifier. *Elsevier*, 403-406.

Berthold, M. R. (2014). *Guide to Intelligent Data Analysis.* London: Springer .

Betancourt, J. (2018). *El fenómeno de la corrupción en los procesos de licitación pública en contratación estatal en Colombia.* Obtenido de El fenómeno de la corrupción en los procesos de licitación pública en contratación estatal en Colombia: https://repository.ucatolica.edu.co/bitstream/10983/15976/1/El%20fen%C3%B3meno%20de%20la%20corrupci%C3%B3n%20en%20los%20procesos%20de%20licitaci%C3%B3n%20p%C3%BAblica%20.pdf

Breiman, L. (1996). Bagging predictors. *Mach Learn 24*, 123–140.

Breiman, L. (2001). Random Forest. *University of California*, 1-33.

Breiman, L., & Cutler, A. (2005). *Random Forests*. Obtenido de Random Forests: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning (ICML).*, 1-8. Obtenido de An empirical evaluation of supervised learning in high dimensions.

Celina, Y. (21 de diciembre de 2016). La ineficiencia del ejercicio mata al presupuesto. *obras expansion*.

Charris, L., Henriquez, C., Hernandez, S., Jimeno, L., Guillen, O., & Moreno, S. (2018). Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos. *Revista I+D en TIC*, 26-34.

Chaves, M. (24 de septiembre de 2018). Malgasto de recursos públicos en Colombia llega a 4,8% del Producto Interno Bruto. *La republica*.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Chawla, N., Japkowicz, N., & Kołcz, A. (2004). Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explorations*, 1-6.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Ma, J. (2017). A comparative study of logistic model tree, random Forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *CATENA*, 147-160.

Colombia Compra eficiente. (2016). *Guía para Procesos de Contratación de obra pública.* Obtenido de Guía para Procesos de Contratación de obra pública: https://colombiacompra.gov.co/sites/cce_public/files/cce_documents/cce_guia_obra_publica.pdf

Colonnelli, E., & Prem, M. (2020). Corruption and Firms . *Social Science Research Network*, 1-85.

Corporación Transparencia por Colombia. (2019). *Así se mueve la corrupción Radiografía de los hechos de corrupción en Colombia 2016-2018.* Bogotá: Monitor ciudadano de la corrupción.

Cox, D., & Snell, E. (1989). *The Analysis of Binary Data.* New York: Chapman & Hall.

Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*, 293-300.

Díaz, J. F., & Correa, J. C. (2013). Comparación entre árboles de regresión CART y regresión lineal. *Dialnet, 6*(2), 175-195.

Díaz, M. (2020). *Limitaciones a la competencia en procesos de contratación pública de vías primarias.* Tesis de grado. Bogotá: Universidad de los Andes.

Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems, 9*, 155-161.

Escobar, A. (1999). *El contrato estatal de obra.* Bogotá: Juridicas G. Ibañez.

Gallego, J., Rivero, G., & Martínez, J. (2020). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, 1-18.

García, A. (2016). El sistema electrónico de contratación pública chileno: ChileCompra. *Revista de Administración Pública*, 363-388.

García, J. (2020). Modelo predictivo para la identificación de actividades de la vida diaria (adl) en ambientes indoor usando técnicas de clasificación basadas en machine learning. Tesis de

grado.Barranquilla: corporación universidad de la costa – CUC.

Germán Morales, J. M. (2008). Estrategia de regresión basada en el método de los k vecinos más cercanos para la estimación de la distancia de falla en sistemas radiales. *Revista Facultad de Ingeniería Universidad de Antioquia 45*, 100-108.

Ha, T., & Bunke, H. (1997). Off-line, handwritten numeral recognition by perturbation method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 535-539.

Hailon, B., & Amancio, J. (2020). *Utilizando inteligência artificial para identificação de suspeita de fraudes em contratos públicos nos municipios de goiás.* TCC. Goia: centro universitário de Anápolis.

Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning.* PhD Thesis. Hamilton: The University of Waikato.

Hernández Mota, J. L. (2009). La Composición del Gasto Público y el Crecimiento Económico. *Análisis Económico , Vol. XXIV, No. 55*, 77-102.

Hoyos, J. (2019). *Metodologia de clasificacion de datos desbalanceados basado en metodos de submuestreo.* Tesis.Pereira: Universidad Tecnologica de Pereira.

Hussein, I., & Miklos, V. (2011). Application of Anomaly Detection Techniques to Identify Fraudulent Refund. *SSRN*, 1-30.

Izquierdo, A., Pessino, C., & Vuletin, G. (2018). Mejor gasto para mejores vidas Cómo América Latina y el Caribe puede hacer más con menos. *Banco interamenricano de desarollo*, 1-66. Obtenido de https://flagships.iadb.org/sites/default/files/dia/chapters/DIA-2018-Capi%CC%81tulo-3-La-%28in%29eficiencia-del-gasto-pu%CC%81blico.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning with aplications in R.* United states: springer.

Khoirunnisaa, A., Pane, E., Wibawa, A., & Purnomo, M. (2018). Channel Selection of EEG-Based Cybersickness Recognition during Playing Video Game Using Correlation Feature Selection (CFS). *2nd International Conference on Biomedical Engineering (IBIOMED). .*

López, F., & Sanz, I. (2017). Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research*, 1-24.

López, J. (2019). *Uso de técnicas de machine learning para la detección de fraudes en los contratos de obras públicas.* santiago de chile: concurso OLACEFS 2019.

Matlab. (2018). *MathWorks*. (MathWorks) Recuperado el 2018, de https://la.mathworks.com/products/matlab.html?requestedDomain=

Melo, L., & Ramos, J. (2017). El gasto público en Colombia: Algunos aspectos sobre su tamaño,. *Borradores de economía*, 1-43. Obtenido de https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/be_1003.pdf

Michele, R., & Pierri, G. (2020). Transparencia y Gobierno Digital: El Impacto de COMPR.AR en Argentina . *Banco Interamericano de Desarrollo (BID)* , 1-27.

Niels, L., Mark, H., & Eibe, F. (2003). Logistic Model Trees. *ECML PKDD, the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Procuraduria nacional. (2015). *Transparencia en el ámbito nacional.* Obtenido de Transparencia en el ámbito nacional: https://www.procuraduria.gov.co/portal/media/file/%C3%8Dndice%20de%20Transparencia%20Nacional%20(ITN)%202013-%202014.pdf

Pulgar, F., Rivera, A., Charte, F., & Jesus, M. (2018). Analisis del impacto de datos desbalanceados en el rendimiento predictivo de redes neuronales convolucionales. *XVIII Conferencia de la Asociacion Espa ´ nola para la Inteligencia Artificial*, 1213 - 1218.

Robles, A., Cortés, P. M., & Barbadill, M. (2020). Aplicación de la regresión logística para la predicción de roturas de tuberías en redes de abastecimiento de agua. *revistadyo*, 85.

Rodriguez, S. (2020). *Predicción de inefecciencia en la contratación publica en bogotá.* bogotá: Universidad del Rosario.

Rosa, P., & Zeviani, W. (2019). *Descartelizando: Uso deMachine Learning e Estatística para Detecção de Indícios de Cartel em Processos Licitatórios.* Paraná: Universidade Federal do Paraná.

Ruiz, A. J. (2020). Inclusión de mujeres en las contrataciones públicas: la experiencia latinoamericana. . *ILDA y programa Open Up Contracting de Hivos* (págs. 1-43). Montevideo: Hivos people unlimeted.

Ruiz, R. (2006). *Heurísticas de selección de atributos para datos de gran dimensionalidad.* Sevilla: Universidad de Sevilla.

Salazar, C. (2020). Gasto público y crecimiento económico: Controversias teóricas y evidencia para México. *Economía UNAM, 17(50)*, 53-71.

Salzberg, S. (1994). Programs for Machine Learning by J. Ross Quinlan. *Kaufmann Publishers*, 235–240.

Santana, E. (14 de diciembre de 2014). *Bagging para mejorar un modelo predictivo*. Obtenido de Bagging para mejorar un modelo predictivo: http://apuntes-r.blogspot.com/2014/12/bagging-para-mejorar-un-modelo.html

Scheller, A., & Silva, S. (2017). La corrupción en la contratación pública: operatividad, tipificación, percepción,costos y beneficios. *Revista VIA IURIS*, 1-36.

Segal, M. R. (2014). Machine Learning Benchmarks and Random Forest Regression. *Center for Bioinformatics & Molecular Biostatistics.*, 1-4.

Serrano, A. (2014). corrupción en la contratación pública en colombia. obtenido de corrupción en la contratación pública en colombia: https://repository.unimilitar.edu.co/bitstream/handle/10654/12906/CORRUPCI%C3%93N%20E

N%20LA%20CONTRATACI%C3%93N%20P%C3%9ABLICA%20EN%20COLOMBIA.%20Aldemar%20Serrano.%20Oct.%208..pdf?sequence=1&isAllowed=y

Singh, A. (18). A Practical Introduction to K-Nearest Neighbors Algorithm for Regression. *August 22*.

Toledo, A. (2016). *Métodos de selección de atributos para clasificación supervisada basados en teoría de información.* La habana: Instituto Superior Politécnico "José Antonio Echeverría".

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 1986-1994.

Transparency International. (2021). *CPI 2020: Resumen global.* Berlín: Transparency International Global.

Yanminsun, Wong, A., & Kamel, M. S. (2011). Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*.

Anexos

Tabla 8. Lista de variables del conjunto de datos.

ID	Nombre Campo	Tipo	Descripción
1	Nombre Entidad	Categórica	Nombre de la Entidad con la que se desarrolló el proceso de compra
2	Departamento	Categórica	Departamento en el cual se registró la entidad del estado que publica el contrato
3	Ciudad	Categórica	Ciudad de Colombia en el cual se registró la entidad del estado que publica el contrato
4	Localización	Categórica	Localización Geográfica de la entidad del estado que publica el contrato
5	Orden	Categórica	Orden entidad del estado que publica el contrato (Nacional, Territorial)
6	ITEP	Numérica	El valor del índice de transparencia entidades publica
7	Sector	Categórica	Sector de la economía de la entidad del estado que publica el contrato
8	Rama	Categórica	Rama del Poder a la que corresponde la entidad del estado que publica el contrato
9	Entidad Centralizada	Categórica	Define si la entidad es descentralizada o centralizada
10	Estado Contrato	Categórica	Estado del contrato, frente a su ejecución, firma o liquidación
11	Código de Categoría Principal	Categórica	Código UNSPSC de la categoría principal para el contrato
12	Tipo de Contrato	Categórica	Tipo de contrato de acuerdo a su marco jurídico

13	Modalidad de Contratación	Categórica	Modalidad de contratación de acuerdo al modelo de selección
14	Fecha de Firma	Fecha	Fecha en que fue firmado digitalmente el contrato
15	Fecha de Inicio del Contrato	Fecha	Fecha de inicio de las responsabilidades contractuales
16	Fecha de Fin del Contrato	Fecha	Fecha de fin de las responsabilidades contractuales
17	Fecha de Inicio de Ejecución	Fecha	Fecha de inicio de la ejecución de las actividades del contrato
18	Fecha de Fin de Ejecución	Fecha	Fecha de finalización de la ejecución de las actividades del contrato
19	Plazo contrato	Numérica	Tiempo transcurrido entre la fecha de firma del contrato y la fecha de finalización de la ejecución
20	Plazo de ejecución	Numérica	Tiempo transcurrido entre la fecha de inicio del contrato y la fecha de finalización de la ejecución
21	Tiempo demora	Numérica	Tiempo transcurrido entre la fecha de firma del contrato y la fecha de inicio de la ejecución
22	Condiciones de Entrega	Categórica	Condiciones bajo las cuales se entrega el producto o servicio
23	Documento Proveedor	Categórica	Número de documento del proveedor adjudicado
24	Proveedor Adjudicado	Categórica	Nombre del proveedor adjudicado
25	Tipo de empresa	Categórica	Tipo de empresa que declara el proveedor al inscribirse
26	Fecha de creación	Fecha	Fecha en la que hizo el primer registro

27	Departamento2	Categórica	en caso de ser colombiano, indica el departamento al que corresponde la ubicación principal del proveedor
28	Municipio2	Categórica	en caso de ser colombiano, indica el municipio al que corresponde la ubicación principal del proveedor
29	Es Grupo	Categórica	Determina el proveedor es un grupo de entidades, existe un conjunto de datos de CCE que contiene la conformación de los grupos
30	Es Pyme	Categórica	Determina si la empresa es una Pyme
31	Habilita Pago Adelantado	Categórica	Determina si el contrato tiene habilitada la opción de pago de adelantos
32	Liquidación	Categórica	Determina si el contrato ha sido liquidado
33	Obligación Ambiental	Categórica	Determina si el contrato tiene compromisos de cumplimiento a obligaciones ambientales
34	Obligaciones Postconsumo	Categórica	Determina si el contrato tiene compromisos de cumplimiento a obligaciones posteriores a la entrega del producto o prestación del servicio
35	Reversión	Categórica	Determina si el contrato ha sido reversado
36	Valor del Contrato	Numérica	Valor total del contrato
37	Valor de pago adelantado	Numérica	Valor del pago por adelantado
38	Valor Facturado	Numérica	Valor Facturado a la fecha
39	Valor Pendiente de Pago	Numérica	Valor Pendiente de Pago a la fecha
40	Valor Pagado	Numérica	Valor Pagado a la fecha

41	Valor Amortizado	Numérica	Valor Amortizado
42	Valor Pendiente de Amortización	Numérica	Valor Pendiente de Amortización a la fecha
43	Valor Pendiente de Ejecución	Numérica	Valor Pendiente de Ejecución a la fecha
44	Saldo CDP	Numérica	Saldo del CDP asignado al proceso y al contrato
45	Saldo Vigencia	Numérica	Saldo actual para la vigencia del CDP asignado al proceso y al contrato
46	Es PostConflicto	Categórica	Determina si el proceso está asociado a algún evento de acuerdo de paz
47	Destino Gasto	Categórica	Destino del gasto, a nivel presupuestal
48	Origen de los Recursos	Categórica	Origen de los Recursos, a nivel presupuestal
49	Días Adicionados	Numérica	Número de días en que el contrato ha sido adicionado
50	Puntos del Acuerdo	Categórica	En caso de ser un proceso que da cumplimiento a compromisos en el acuerdo de paz, determina a qué puntos da conformidad
51	Pilares del Acuerdo	Categórica	En caso de ser un proceso derivado de compromisos del acuerdo de paz, define el pilar de acuerdo de paz al que corresponde
52	Presupuesto General de la Nación – PGN	Categórica	En un escenario de Origen de recursos múltiple, se registra el Valor correspondiente a Presupuesto General de la Nación
53	Sistema General de Participaciones	Categórica	En un escenario de Origen de recursos múltiple, se registra el Valor correspondiente a Sistema General de Participaciones

54	Sistema General de Regalías	Categórica	En un escenario de Origen de recursos múltiple, se registra el Valor correspondiente a Sistema General de Regalías
55	Recursos Propios (Alcaldías, Gobernaciones y Resguardos Indígenas)	Categórica	En un escenario de Origen de recursos múltiple, se registra el Valor correspondiente a Recursos Propios de la entidad, en Alcaldías, Gobernaciones o Resguardos Indígenas
56	Recursos de Crédito	Categórica	En un escenario de Origen de recursos múltiple, se registra el Valor correspondiente a Recursos de Crédito
57	Recursos Propios	Categórica	En un escenario de Origen de recursos múltiple, se registra el Valor correspondiente a Recursos Propios de la entidad.
58	Numero de categoría entidad	Numérica	Número de categorías diferentes de bienes y servicios contratados por la entidad.
59	contratos por proveedor	Numérica	Número de contratos que tiene asignado el proveedor
60	Class	Categórica	Variable binaria que determina 1 si el contrato tiene adición de valor o 0 en caso contrario

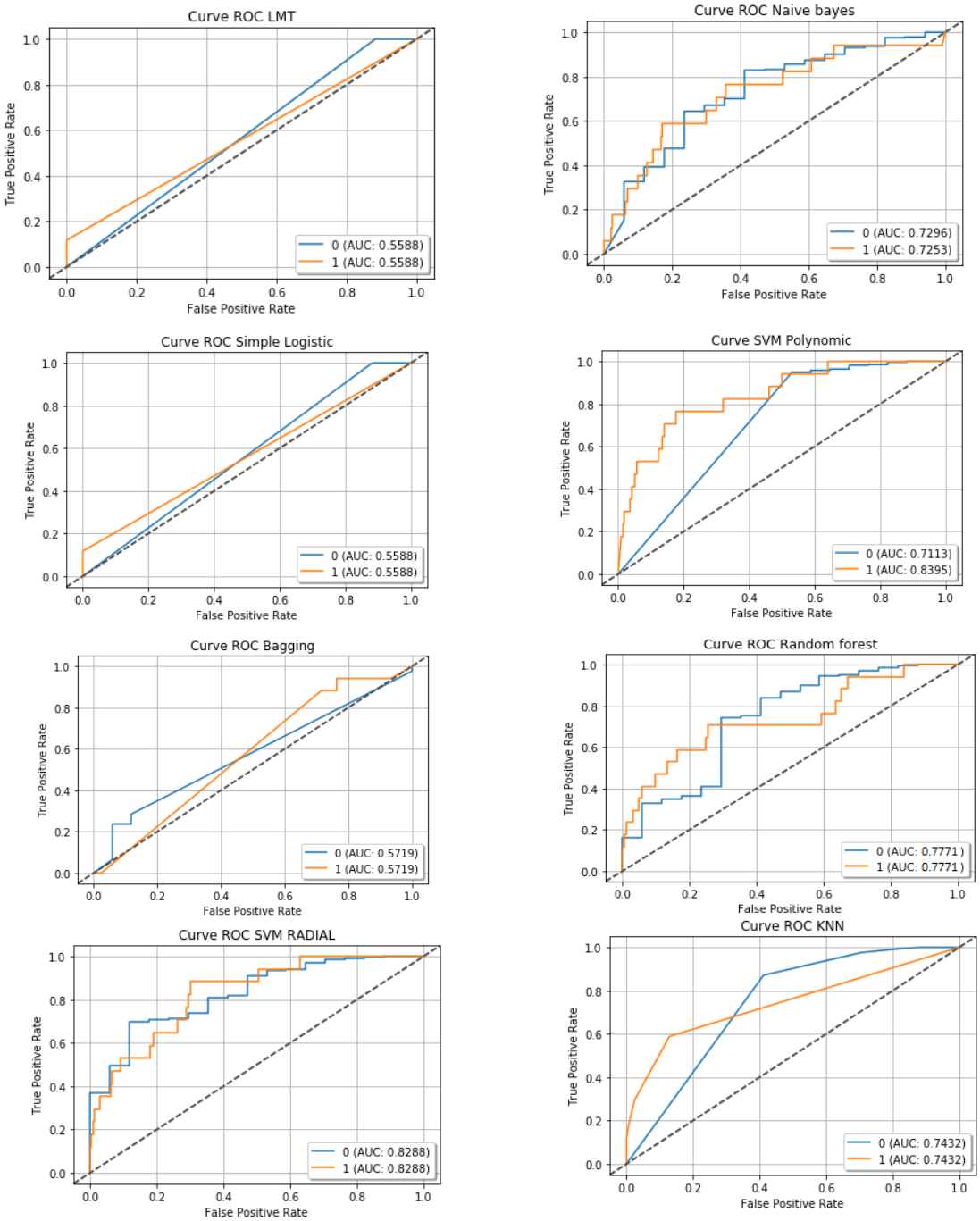


Figura 11. Curva ROC para diferentes modelos con datos originales. Fuente: elaboración propia.

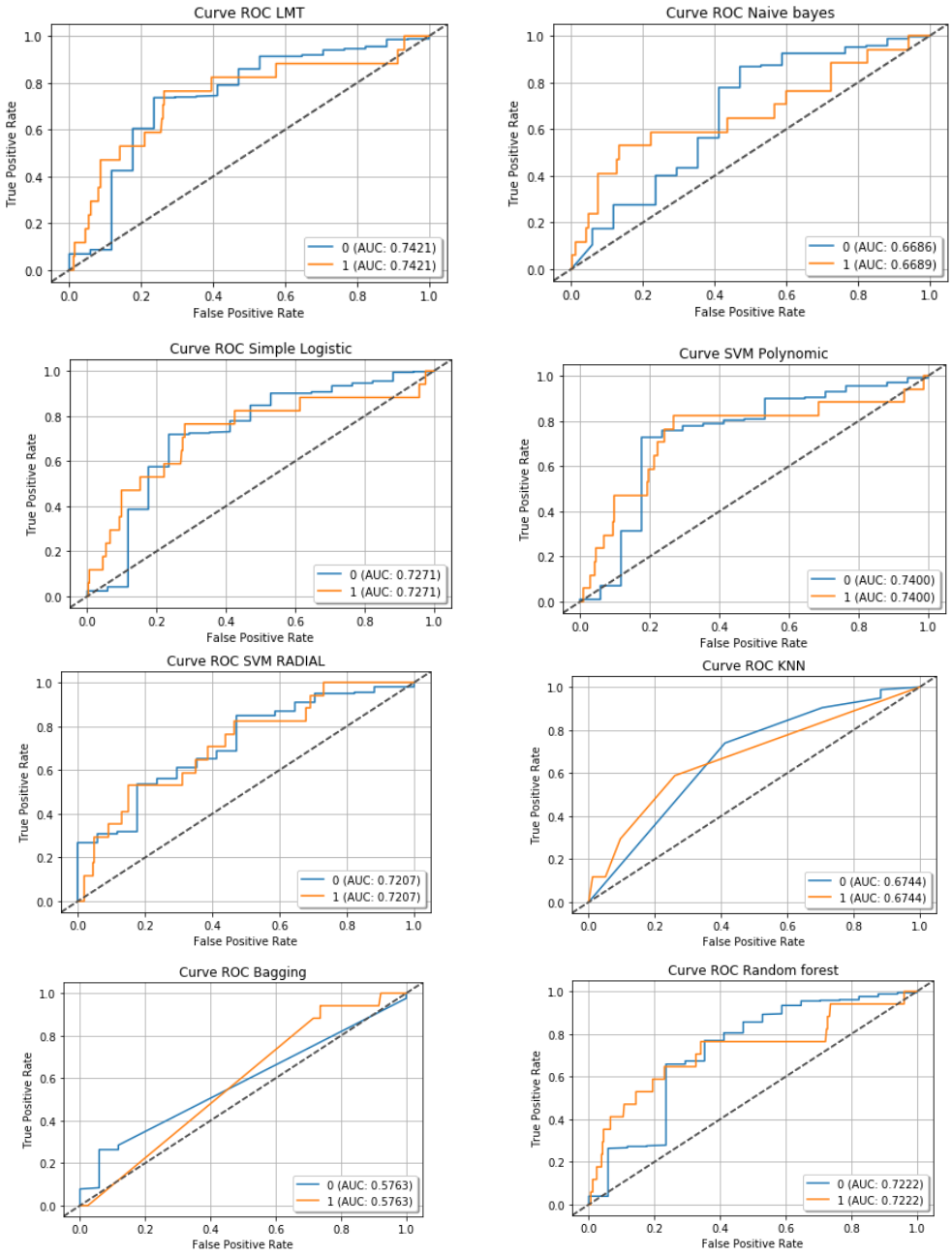


Figura 12. Curva ROC para diferentes modelos con datos CFS. Fuente: elaboración propia.

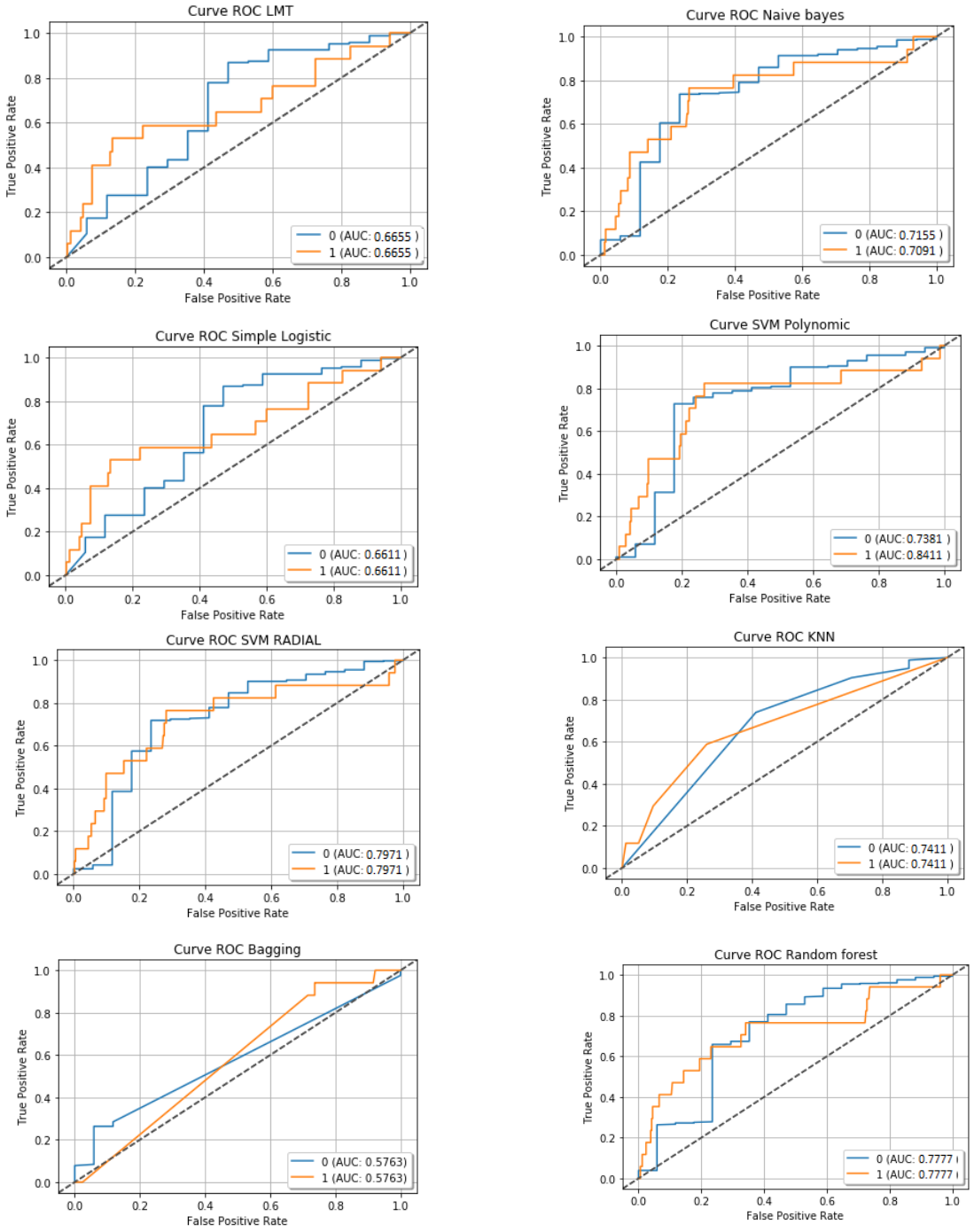


Figura 13. Curva ROC para diferentes modelos con datos SMOTE. Fuente: elaboración propia.