# Modelo de analítica de datos para predecir posibles sanciones disciplinarias hacia funcionarios que ocupen cargos de elección popular en Colombia

**Héctor Andrés Pulido[1], Fredys A. Simanca H.[2] Pablo E. Carreño H.[3], William Insignares[4], Emeldo Caballero[5]**

[1]Universidad Libre, Bogotá, Colombia, hectora-pulidom@unilibre.edu.co

[2]Universidad Cooperativa de Colombia, Bogotá, Colombia, fredysa.simancah@campusucc.edu.co

[3]Universidad Libre, Bogotá, Colombia, pabloe.carrenoh@unilibre.edu.co

[4]Universidad Libre, Barranquilla, Colombia, william.insignares@unilibre.edu.co

[5]Universidad Libre, Barranquilla, Colombia, emeldo.caballerob@unilibre.edu.co

Received: date; Accepted: date; Published: date

**Abstract:** El presente artículo describe el diseño de un modelo de analítica de datos para predecir posibles sanciones disciplinarias hacia funcionarios que ocupan cargos de elección popular en Colombia, se llevó a cabo un proceso de exploración de datos y se aplicaron distintos modelos de Machine Learning con el fin de determinar cuál de ellos es el más adecuado para este propósito. Lo que a su vez permitió identificar situaciones de riesgo y anticiparse a ellas mediante la implementación de medidas preventivas. Para ello, se aplicó un enfoque de aprendizaje supervisado en la construcción del modelo, en el que se utilizaron modelos de clasificación para predecir si un funcionario determinado podría ser objeto de sanciones disciplinarias en el futuro. Uno de los aspectos clave de este artículo fue la optimización de los hiperparámetros del modelo, ya que se logró una buena precisión y un desempeño óptimo. Se exploraron diferentes valores de los hiperparámetros y se seleccionaron aquellos que permitieron obtener los mejores resultados. Por último, se definieron las métricas de medición del modelo, con el fin de evaluar su precisión y capacidad predictiva. El modelo diseñado en este artículo puedo proporcionar una herramienta valiosa para la toma de decisiones en el ámbito del gobierno digital y contribuir a mejorar la eficiencia y transparencia en el desempeño de los funcionarios públicos.

## 1. Introducción

La administración pública en Colombia es una de las actividades más importantes para el desarrollo del país, y es crucial que los funcionarios que ocupan cargos de elección popular mantengan los más altos estándares éticos y de desempeño en sus funciones. Sin embargo, en ocasiones se presentan situaciones de incumplimiento o violación de las normas que pueden dar lugar a sanciones disciplinarias. La corrupción y el mal uso de los recursos públicos son problemas que afectan directamente el bienestar de la población y socavan la confianza en las instituciones.

En este sentido, el presente artículo resultado de investigación tiene como objetivo diseñar un modelo de analítica de datos [1] para predecir posibles sanciones disciplinarias hacia funcionarios que ocupan cargos de elección popular en Colombia. La implementación de este modelo permitirá

identificar patrones y tendencias en la información disponible sobre las características de los funcionarios, lo que a su vez permitirá identificar posibles situaciones de riesgo y anticiparse a ellas mediante la implementación de medidas preventivas.

El proceso de exploración de datos y la aplicación de modelos de Machine Learning permitirán identificar patrones y tendencias en la información disponible. La optimización de los hiperparámetros del modelo será clave para lograr una precisión y un desempeño óptimos, y se definirán las métricas de medición del modelo para evaluar su capacidad predictiva.

La importancia de este artículo radica en su potencial para proporcionar una herramienta valiosa para la toma de decisiones en el ámbito del gobierno digital y contribuir a mejorar la eficiencia y transparencia en el desempeño de los funcionarios públicos. Además, el articulo tiene una relevancia académica en el ámbito de la ingeniería, ya que combina conceptos de análisis de datos, Machine Learning y gobierno digital.

La metodología utilizada en el artículo fue de tipo descriptiva y explicativa, y se llevó a cabo un proceso de investigación documental y de campo. Se describió el problema y se formuló el objetivo general y los objetivos específicos, y se definieron la delimitación y el alcance del artículo de investigación. Se presenta un marco normativo y legal, un marco referencial, un marco teórico y un marco conceptual que permitieron contextualizar el artículo en su entorno.

Finalmente, se presenta el desarrollo del modelo, que incluirá la descripción del proceso de exploración de datos, la aplicación de los modelos de Machine Learning, la optimización de los hiperparámetros, la definición de las métricas de medición y el análisis de resultados. Se concluye con una discusión sobre los hallazgos del articulo y las implicaciones de los resultados obtenidos.

## 2. Materiales y Métodos

### Etapa de revisión

Las entidades gubernamentales suelen tener grandes cantidades de datos de distintas fuentes, que en muchos casos son bases de datos aisladas. La integración de estos datos y el uso de herramientas de Machine Learning pueden permitir la detección de patrones y tendencias que a priori podrían pasar desapercibidos.

En Colombia, uno de los principales problemas es la corrupción que se presenta en distintos municipios y departamentos, lo que se debe en parte a la falta de transparencia y ética de algunos servidores públicos, especialmente aquellos que son de elección popular, como alcaldes, gobernadores, ediles, entre otros.

La corrupción es un problema grave en Colombia y en el mundo. Según el Índice de Percepción de la Corrupción (IPC) 2021 de Transparencia Internacional, Colombia obtuvo 39 puntos sobre 100, siendo 0 corrupción muy elevada y 100 ausencia de corrupción. El país se ubica en el puesto 87 entre 180 países evaluados [2]. Entre 2016 y 2020 se reportó la pérdida de $13,67 billones en 284 hechos de corrupción [3].

Para combatir este problema, las entidades de control gubernamentales invierten grandes sumas de dinero en la lucha contra la corrupción. Sin embargo, debido al gran número de funcionarios de elección popular en el país, no siempre es fácil llevar a cabo una vigilancia efectiva sobre todos ellos. Para hacerse una idea, actualmente en las elecciones regionales se eligen los siguientes cargos.

32 gobernadores.

418 diputados que conformarán las asambleas departamentales.

1.102 alcaldes.

12.072 concejales de todos los municipios y ciudades del país.

6.513 ediles que conformarán las Juntas Administradoras Locales (JAL).

A pesar de los esfuerzos por parte del gobierno y las entidades de control, la corrupción sigue siendo un obstáculo importante para el desarrollo y la prosperidad del país.

Por esta razón, es esencial que se implementen medidas efectivas para prevenir, detectar y sancionar la corrupción en todas sus formas. En particular, la detección temprana y la sanción de los funcionarios de elección popular que incurren en conductas indebidas pueden tener un impacto significativo en la prevención de la corrupción.

Es importante tener en cuenta que los cargos de elección popular son altamente demandados y competitivos, por lo que es fundamental contar con herramientas que permitan a las entidades de control detectar posibles irregularidades en el proceso electoral y en el desempeño de los funcionarios electos. De esta manera, se pueden tomar medidas oportunas y efectivas para garantizar la transparencia y la ética en la administración pública.

Es así como al utilizar técnicas de Machine Learning, se obtiene un mayor grado de precisión y eficacia en la detección temprana de posibles conductas indebidas y en la identificación de los funcionarios que requieren mayor atención por parte de las entidades de control con la implementación de varios modelos de predicción de Machine Learning, específicamente modelos de clasificación (aprendizaje supervisado), los cuales serán alimentados por diversas fuentes de información o bases de datos, como los antecedentes disciplinarios, los ingresos de los candidatos, las inhabilidades de los candidatos, la información de Territata de los municipios y departamentos, entre otros.

**Etapa final**

Las fuentes de datos que se utilizaron para recopilar los datos incluyen el uso de APIs como la de Socrata de datos abiertos del gobierno colombiano, archivos de Excel y bases de datos. Se garantizo que todas las fuentes utilizadas sean confiables y estén disponibles públicamente para asegurar la transparencia del modelo.

Además, se realizó una revisión exhaustiva de los datos para asegurarse de su calidad y consistencia, eliminando datos incompletos o inconsistentes que puedan afectar la precisión y confiabilidad del modelo.

Para procesar, analizar e interpretar los datos se utilizó Python debido a la facilidad que tiene este lenguaje de programación para analizar grandes cantidades de datos y en general por las múltiples librerías que tiene para analítica de datos y Machine Learning como pandas, numpy, SkyLearn, entre otros.

La población o universo de estudio son los funcionarios de elección popular electos en los periodos 2011, 2015 y 2019, dispuestos por el Consejo Nacional Electoral [4].

## 3. Resultados

El proceso de análisis de datos en un proyecto de investigación es fundamental para garantizar la calidad y fiabilidad de los resultados obtenidos. Para ello, es necesario seguir una metodología rigurosa que permita abordar de manera sistemática cada una de las fases del proceso. En este capítulo, se aplicará la metodología CRISP-DM, que consta de seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue [1].

**Comprensión del negocio**

La Procuraduría General de la Nación (PGN) es una entidad autónoma del Estado colombiano encargada de garantizar la transparencia y la ética en la función pública [2], así como la defensa del patrimonio público y la legalidad en general. Fue creada en 1830 como una dependencia del Ministerio de Gobierno, y posteriormente fue reformada y transformada en su forma actual mediante la Constitución de 1991.

**Las funciones misionales de la PGN son tres:**

*La función preventiva:* Es una de las principales responsabilidades de la Procuraduría General de la Nación, ya que su objetivo es evitar que los servidores públicos cometan faltas disciplinarias.
*La función de intervención:* La Procuraduría interviene ante las diferentes jurisdicciones para defender los derechos y las garantías fundamentales.

*La función disciplinaria:* La entidad es la encargada de investigar y fallar las sanciones correspondientes en caso de faltas disciplinarias cometidas por servidores públicos y particulares que ejercen funciones públicas o manejan dineros del Estado (Ley 734 de 2002).

Es importante destacar que la Procuraduría no solo se enfoca en los delitos electorales cometidos por servidores públicos, sino que también investiga y sanciona a los particulares que infrinjan la normativa electoral o que colaboren en la comisión de estos delitos.

En Colombia, cada 4 años se realizan las elecciones regionales, en las que se escogen gobernadores de los 32 departamentos que tiene el país, 418 diputados de asamblea departamental, alcaldes de 1.102 municipios, 12.072 concejales y 6.513 ediles locales, estos funcionarios son de gran importancia para la Procuraduría, ya que son los encargados de tomar decisiones y gestionar recursos en sus respectivas regiones. El modelo de analítica de este proyecto de investigación utilizó los datos recopilados en distintas bases de datos, para predecir posibles sanciones disciplinarias hacia estos funcionarios electos.

**Comprensión de los datos**

Para llevar a cabo el análisis y la predicción de posibles sanciones disciplinarias en cargos de elección popular en Colombia, fue necesario contar con información actualizada y precisa así que se utilizaron diversas fuentes de información que permitieron obtener una visión completa y detallada de los antecedentes disciplinarios, inhabilidades, ingresos y financiadores de los candidatos, así como los datos de los partidos políticos y la información estadística sobre los territorios del país [3]. A continuación, en la tabla No. 1, se presentan las 5 fuentes de información o bases de datos que se unirán para este análisis.

| Base de datos | Fuente | Sistema de información | Contenido |
|---|---|---|---|
| Antecedentes disciplinarios | datos.gov.co | API Web | Sanciones disciplinarias certificables proferidas contra servidores, exservidores públicos y particulares que desempeñen funciones públicas. |
| Ingresos de los candidatos | cnecuentasclaras.gov.co | Archivo Excel | Información referente a los ingresos de los candidatos de las elecciones regionales en Colombia, así como también datos de sus financiadores. |

| | | | |
|---|---|---|---|
| Inhabilidades de los candidatos | Sistema de información de registro de sanciones y causas de inhabilidad | Archivo Excel | Inhabilidades de los candidatos a elecciones regionales en Colombia. |
| Candidatos inscritos en elecciones territoriales | cnecuentasclaras.gov.co | Archivo Excel | Listado con los datos de los candidatos inscritos a las elecciones regionales en Colombia y los partidos políticos a los que pertenecen. |
| Terridata | Departamento Nacional de Planeación DNP | Archivo plano | Conjunto de datos que provee información territorial y estadística de Colombia, que se encuentra organizada en diferentes capas geográficas. |

*Tabla No. 1: Bases de datos usadas en el modelo analítico*

La información recopilada para el caso de las bases de datos de "Ingresos de los candidatos", "Inhabilidades de los candidatos" y "Candidatos inscritos en elecciones territoriales" corresponde a tres periodos electorales distintos [4]: 2011, 2015 y 2019.

1. Ingresos de los candidatos: 27 columnas

2011: 103.177 registros
2015: 130.862 registros
2019: 185.682 registros

| Nombre columna | Descripción |
|---|---|
| Corporación o Cargo | El cargo o la posición dentro de la corporación a la que el candidato está postulando |
| Circunscripción Electoral | La región electoral a la que pertenece el candidato |
| Departamento | El departamento en el que el candidato está postulando |
| Municipio | El municipio en el que el candidato está postulando |
| Localidad | La localidad en la que el candidato está postulando |
| Organización Política | El partido político o coalición a la que el candidato pertenece |
| Identificación Candidato | El número de identificación del candidato |
| Nombre Candidato | El nombre del candidato |
| Código | Un código asignado para identificar el ingreso |
| Nombre del ingreso | Nombre asignado al tipo de ingreso |
| Tipo Persona | Indica si la donación fue realizada por una persona natural o jurídica |
| Nombre de la Persona | El nombre de la persona o entidad que realizó la donación |
| Valor | La cantidad de dinero donada |
| Departamento Ingreso | El departamento en el que se realizó la donación |
| Ciudad Ingreso | La ciudad en la que se realizó la donación |

| | |
|---|---|
| Tipo de Identificación | Indica el tipo de identificación utilizada por la persona o entidad que realizó la donación |
| Número de Identificación | El número de identificación de la persona o entidad que realizó la donación |
| Acta No. | El número del acta correspondiente al registro del ingreso |
| Valor Pignoración | Indica si la donación fue pignorada o no |
| Descripción del ingreso | Una descripción detallada del origen del ingreso |
| Observaciones | Cualquier observación adicional relacionada con la donación |
| Parentesco | Indica si la persona o entidad que realizó la donación tiene algún parentesco con el candidato |
| Tipo Donación | Indica el tipo de donación (en efectivo, en especie, etc.) |
| Partido Coalición | El partido político o coalición a la que está afiliada la persona o entidad que realizó la donación |
| Fecha Registro Movimiento | La fecha en que se registró el ingreso |
| Número Comprobante Interno | El número de comprobante interno correspondiente al registro del ingreso |
| Concepto del Ingreso | Una descripción breve del origen del ingreso |

*Tabla No. 2: Descripción de las columnas de la tabla de Ingresos de los candidatos*

2. Inhabilidades de los candidatos: 35 columnas
   2011: 236 registros
   2015: 1.782 registros
   2019: 3.702 registros
3. Candidatos inscritos en elecciones territoriales: 28 columnas
   2011: 100.159 registros
   2015: 112.244 registros
   2019: 117.160 registros
4. Antecedentes disciplinarios: 24 columnas, 52.527 registros
5. Territata: 71 columnas, 1.102 registros

**Preparación de los datos**

Etapa crítica en el proceso de minería de datos, ya que es aquí donde se lleva a cabo la limpieza, la integración, la selección y la transformación de los datos que se van a utilizar en el análisis y en los modelos de Machine Learning.

**Integración de los datos**

Para integrar los datos se llevó a cabo la unificación de los tres periodos de elecciones (2011, 2015 y 2019) en un único periodo, en el caso de las bases de datos que estaban separadas por estos periodos específicos, como "Ingresos de los candidatos", "Inhabilidades de los candidatos" y

"Candidatos inscritos en elecciones territoriales". De esta manera, se obtuvo una sola tabla por cada una de las bases de datos que contenía todos los datos correspondientes a los periodos mencionados. Posteriormente, se procedió a realizar un cruce de estas tres bases de datos con la base de datos "Antecedentes disciplinarios". Para ello, se utilizó la columna de cédula del candidato como llave natural, ya que esta columna se encontraba presente en todas las tablas y permitía establecer una relación entre ellas.

Como resultado de este proceso, se obtuvo una tabla denominada "funcionarios base" que contenía 31 columnas y 713.961 registros, Figura 1. Posteriormente se integró esta tabla con la de Territata, dando como resultado una tabla llamada "funcionarios electos", Figura 2.

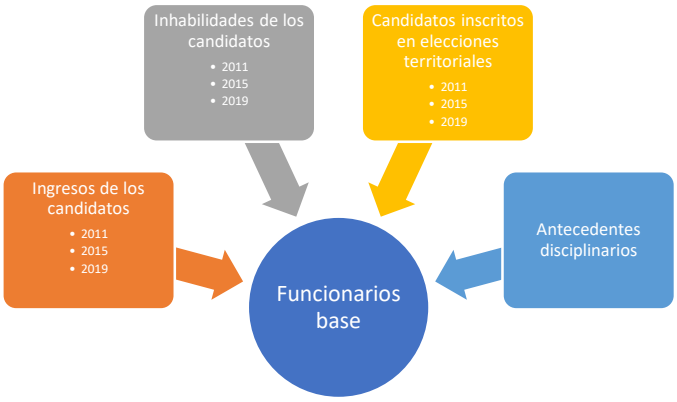


*Figura 1- Integración de las bases de datos de funcionarios*

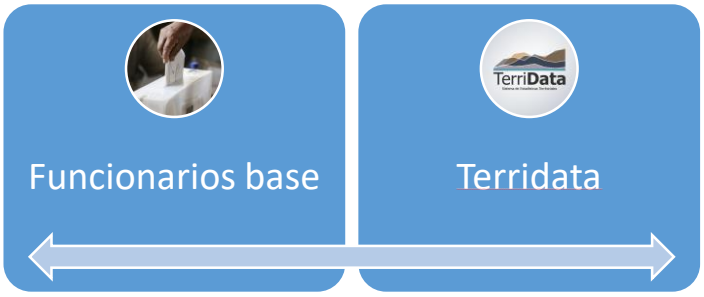


*Figura 2- Integración de la base de datos de funcionarios con Terridata*

**Eliminar variables irrelevantes y redundantes**

En este paso, se eliminaron aquellas variables que no tienen relevancia para el análisis o que presentaban una correlación alta con otras variables en la base de datos.

**Limpieza de datos**

Para limpiar los datos, se llevó a cabo un filtrado de la base de datos de "funcionarios base", con el objetivo de dejar solamente a los funcionarios que resultaron electos para su respectivo cargo. Este proceso se realizó mediante la filtración de la columna "Electo", dejando únicamente aquellos registros cuyo valor fuera distinto a "NO", Figura 3. Además, se unificaron los nombres de los

registros en dicha columna, ya que, al unir distintas bases de datos, se presentaban inconsistencias en la escritura de los mismos, algunos en minúsculas y otros en mayúsculas, pero con el mismo valor.

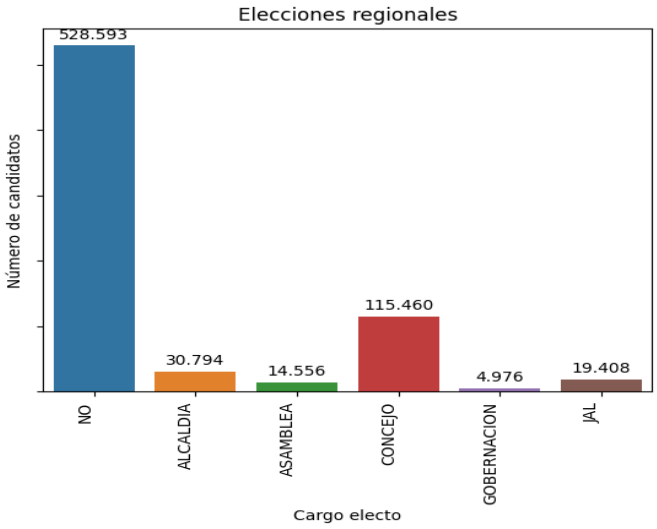


*Figura 3- Clasificación de los candidatos según el cargo electo*

Debido a este primer paso en el proceso de limpieza, se logró una reducción en el número de registros, dejando únicamente a los funcionarios electos para sus respectivos cargos. De esta manera, se obtuvo una tabla enfocada en los objetivos del proyecto, lo que permitirá un análisis más eficiente y preciso de los datos.

**Creación de nuevas variables**

Con la finalidad de mejorar la calidad de los datos, se procedió a crear una nueva columna que contenía el nombre completo de los funcionarios electos en cargos de elección popular. De esta manera, se pudieron eliminar cuatro columnas que contenían los nombres y apellidos del funcionario por separado, Figura 4. Lo que simplificó la estructura de la tabla y evitó la duplicidad en los registros.

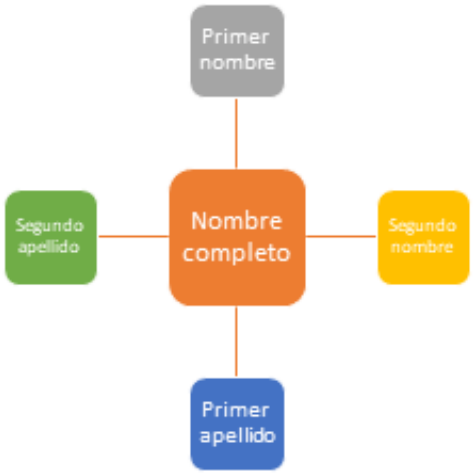


*Figura 4- Estructura columna "Nombre completo"*

De igual manera, se procedió a crear una variable llamada "DIVIPOLA" con el objetivo de mejorar la estructura de la tabla. Esta nueva variable permitió reemplazar las columnas de "Código Departamento" y "Código Municipio", las cuales fueron eliminadas para simplificar la estructura de la tabla, Figura 5. De esta manera, se logró una mayor eficiencia en el análisis de los datos, ya que se redujo el número de columnas innecesarias y se generó una tabla más limpia y clara.

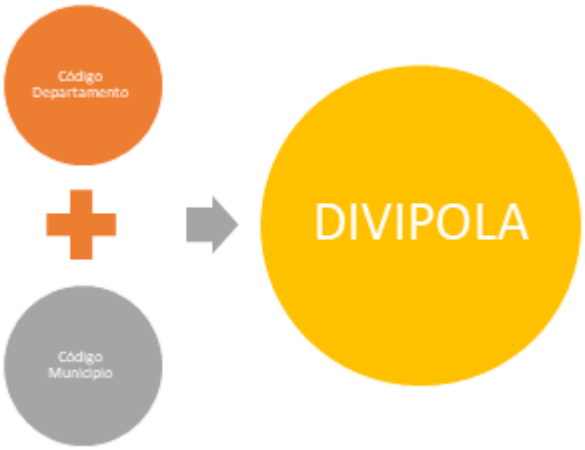


*Figura 5- Estructura columna "DIVIPOLA"*

**Análisis exploratorio de los datos**

Todos los pasos necesarios para preparar los datos de la base de datos "funcionarios electos" han sido completados previamente. Esta base de datos consta de 84 columnas y 41.139 registros.

La columna "sanciones numérica" de la base de datos "funcionarios electos" se clasifica en dos categorías:

- 0 - No sancionado
- 1 - Sancionado

En la Figura 6, se puede observar la distribución de las edades de los funcionarios electos en relación a la variable de salida "sanciones numérica".

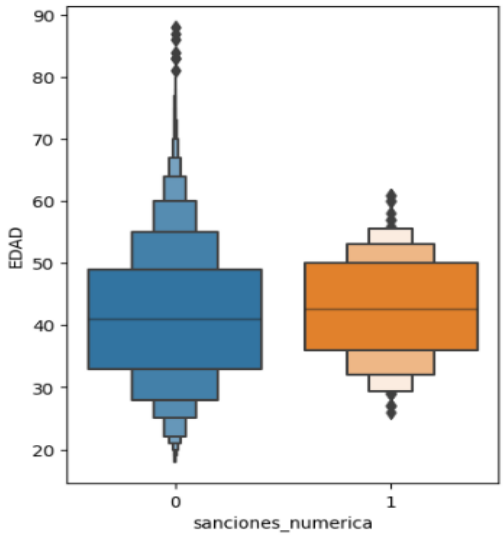


*Figura 6- Edad de los funcionarios electos VS variable de salida "sanciones numérica"*

De acuerdo con la Figura 6, la mayoría de los funcionarios electos se encuentran en el rango de edad entre 32 y 50 años. Además, se puede observar que, en general, los funcionarios electos que han sido sancionados tienen una edad ligeramente mayor que aquellos que no han sido sancionados.

En la Figura 7, se presenta la relación entre el indicador de pobreza multidimensional rural y la variable de salida "sanciones numérica". Se puede observar que los funcionarios electos sancionados suelen provenir de municipios con un índice de pobreza multidimensional rural más alto.

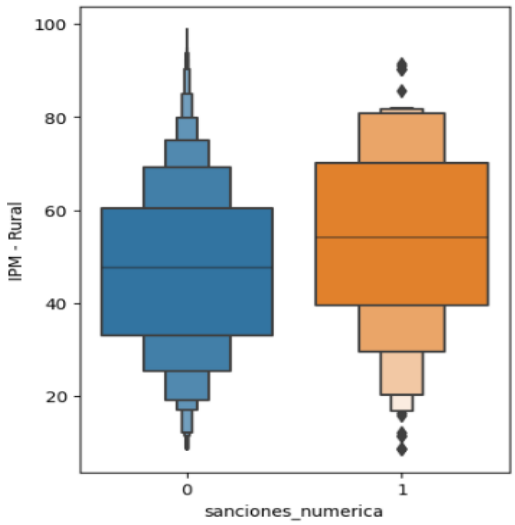


*Figura 7- IPM - Rural VS variable de salida "sanciones numérica"*

En la Figura 8, se presenta una gráfica que muestra el número de funcionarios electos que fueron sancionados y el número de funcionarios electos que no fueron sancionados, con respecto a su cargo dentro del gobierno. Se observa que, de los 27.876 concejales, 148 han sido sancionados, mientras que, entre los alcaldes, 91 han recibido sanciones, lo que representa un 3.05% del total de alcaldes electos.

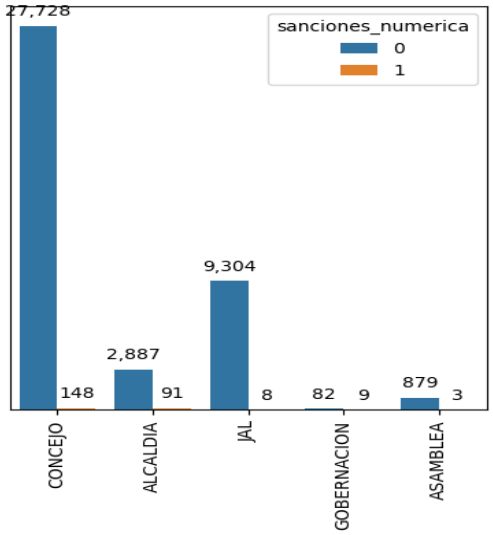


*Figura 8-Número de funcionarios electos sancionados y no sancionados por cargo*

**Reducción de variables**

Una de las técnicas de reducción de variables utilizadas en este proyecto fue el Análisis de Componentes Principales (PCA), que puede utilizarse para reducir la dimensionalidad de los datos y retener las características más importantes.

**Balanceo de datos**

Se utilizó el parámetro de class_weight='balanced' de la librería scikit-learn de Python en los modelos. Este parámetro ajusta automáticamente los pesos de las clases en función de la frecuencia de aparición de cada una de ellas en los datos, de modo que las muestras de la clase minoritaria tienen un mayor peso que las muestras de la clase mayoritaria. Esto ayuda a los modelos a tener en cuenta adecuadamente las muestras de la clase minoritaria durante el entrenamiento.

**Transformación de tipo de datos**

Se utilizaron técnicas de codificación de variables categóricas para poder utilizarlas en los modelos de Machine Learning. Para ello, se empleó la función ColumnTransformer de la librería sklearn.compose de Python.

Se creó un ColumnTransformer con dos transformadores diferentes. El primero utilizó la técnica de codificación one-hot con el parámetro handle_unknown establecido en 'ignore', para codificar las variables categóricas: Corporación cargo, Circunscripción, Nombre región, Tipo partido, Nombre partido político y Electo.

El segundo transformador también utilizó la técnica de codificación one-hot, pero con el parámetro drop establecido en 'if_binary', para codificar las variables binarias: Genero e Inhabilitado. De esta forma, se transformaron las variables categóricas y binarias en variables numéricas para ser usadas en los modelos de aprendizaje automático más adelante.

**Modelado**

Se utilizaron modelos de clasificación para predecir qué funcionarios de elección popular serían sancionados disciplinariamente después de su elección. El problema se planteó como una clasificación entre funcionarios sancionados y no sancionados.

Además de los modelos de clasificación, también se utilizaron técnicas de clustering con el mismo propósito de identificar patrones y características comunes entre los funcionarios sancionados. A través del análisis de agrupamiento, se buscó identificar los grupos de funcionarios sancionados y entender las características distintivas de cada uno de ellos.

De esta manera, la combinación de técnicas de clasificación y clustering permitió una visión más completa y detallada de los funcionarios sancionados. Mientras que la clasificación permitió predecir quiénes serían sancionados disciplinariamente, el clustering permitió entender mejor las razones detrás de esas sanciones y agrupar a los funcionarios según patrones de comportamiento y características comunes.

Los pasos que se siguieron para el modelado fueron:

**Importe de las librerías**

Se utilizó la librería scikit-learn (también conocida como sklearn) en Python. scikit-learn es una librería de aprendizaje automático de código abierto que proporciona una amplia variedad de herramientas para construir modelos de aprendizaje automático.

Se importó la función train_test_split de la librería sklearn.model_selection, que se utilizó para dividir los datos en conjuntos de entrenamiento y prueba.

La función ColumnTransformer de la librería sklearn.compose se utilizó para aplicar diferentes transformaciones a diferentes columnas del conjunto de datos. Asimismo se utilizaron transformaciones de datos en los modelos, como es el caso de OneHotEncoder de la librería sklearn.preprocessing, para transformar variables categóricas en variables numéricas.

Se aplicaron métodos de escalamiento para normalizar los datos y mejorar la precisión del modelo, utilizando MaxAbsScaler de la librería sklearn.preprocessing. Se utilizó SimpleImputer de la librería sklearn.impute, para imputar valores faltantes en los datos. Se probaron varios modelos de clasificación, como SVC, KNeighborsClassifier, DecisionTreeClassifier, LogisticRegression y RandomForestClassifier de la librería sklearn. Para optimizar la selección de hiperparámetros, se utilizó la función GridSearchCV y RandomizedSearchCV de la librería sklearn.model_selection.

En cuanto a la librería sklearn.metrics, se utilizó la función classification_report para obtener un informe detallado de las métricas de rendimiento del modelo de clasificación. Esto incluye la precisión, el recall, la puntuación F1 y el soporte para cada clase.

Para reducir la dimensionalidad de los datos se utilizaron dos técnicas: UMAP y PCA de las librerías umap y sklearn.decomposition, respectivamente.

También se aplicó el algoritmo KMeans y DBSCAN de la librería sklearn.cluster, para realizar la clusterización de los datos. El rendimiento de la clusterización se evaluó utilizando silhouette_score de la librería sklearn.metrics.cluster.

Finalmente, se utilizó Pipeline de la librería sklearn.pipeline para crear flujos de trabajo de modelado más complejos y reducir la cantidad de código escrito.

**Separación de los datos**

Se dividieron los datos de la base de datos de "funcionarios electos" en dos DataFrames "X" y "y", en los cuales "X" son todos los datos de entrada a los distintos modelos de Machine Learning y "y" es la variable de salida, es decir la variable a predecir "sanciones numérica".
Se separaron los datos de entrenamiento y prueba, de tal forma que en el entrenamiento se escogieron el 70% de los datos y en la prueba el 30% restante, esto se hizo utilizando la función 'train_test_split' para dividir el conjunto de datos en dos partes.

**Transformador**

Se creo el objeto "transfomer" de la clase ColumnTransformer, que se utiliza para transformar diferentes columnas de un DataFrame en características numéricas para su uso en modelos de Machine Learning. En este caso, se utilizan tres transformadores diferentes: **oh_encoder**, **oh_encoder_binary** y **pass**.

**Escalador**

La variable "escalar" que se encuentra en el código es un objeto de la clase MaxAbsScaler, que se utiliza para escalar los datos en un rango de [-1, 1] dividiéndolos por el valor máximo absoluto de cada característica. Este tipo de escalamiento es útil para datos dispersos y es invariante ante cualquier desplazamiento en la distribución original de los datos.

**Imputador**

Se crea un objeto de la clase SimpleImputer llamado "imputador" que se utiliza para imputar valores faltantes en los datos. La estrategia de imputación seleccionada es "most_frequent", lo que significa que se reemplazarán los valores faltantes con el valor más frecuente de la columna correspondiente. En otras palabras, se utilizará el valor más común de cada columna para reemplazar los valores faltantes en esa columna.

**Métrica de medición seleccionada**

En este caso, se ha seleccionado la métrica de recall para evaluar los modelos.

**Entrenamiento de los modelos de clasificación**

Se utilizaron varios algoritmos de aprendizaje supervisado. Cada uno de estos modelos se entrenó utilizando el conjunto de datos de entrenamiento previamente preprocesado y se evaluó su rendimiento utilizando métricas de evaluación como la precisión y el recall. Modelos utilizados: SVC, KNeighborsClassifier, DecisionTreeClassifier, LogisticRegression y RandomForestClassifier.

**Análisis de resultados**

En la comparación de los resultados para SVC antes y después de la optimización de los hiperparámetros, se puede observar que el recall de la clase 0 disminuyó ligeramente de 0.84 a 0.80, mientras que el recall de la clase 1 aumentó de 0.59 a 0.63, Tabla 3.

| Métrica | Antes de la optimización | Después de la optimización |
|---|---|---|
| Precisión | 1.00 / 0.02 | 1.00 / 0.02 |
| Recall | 0.84 / 0.59 | 0.80 / 0.63 |

| | | |
|---|---|---|
| F1-score | 0.91 / 0.04 | 0.89 / 0.04 |
| Accuracy | 0.84 | 0.80 |
| Macro avg | 0.51 / 0.72 | 0.51 / 0.71 |
| Weighted avg | 0.99 / 0.84 | 0.99 / 0.80 |

*Tabla 3- Comparación métricas de SVC antes y después de la optimización de hiperparámetros*

Al entrenar el modelo con el algoritmo DecisionTreeClassifier, se obtuvo un recall de 0.99 para la clase 0 (funcionarios no sancionados) y un recall de 0.10 para la clase 1 (funcionarios sancionados). Sin embargo, al optimizar los hiperparámetros, no se logró mejorar el rendimiento en términos de recall de la clase 1. En cambio, el recall de la clase 0 subió a 1.00 después de la optimización, Tabla 4.

| **Métrica** | **Antes de la optimización** | **Después de la optimización** |
|---|---|---|
| Precisión | 0.99 / 0.07 | 0.99 / 0.24 |
| Recall | 0.99 / 0.10 | 1.00 / 0.06 |
| F1-score | 0.99 / 0.09 | 1.00 / 0.10 |
| Accuracy | 0.99 | 0.99 |
| Macro avg | 0.53 / 0.55 | 0.62 / 0.53 |
| Weighted avg | 0.99 / 0.99 | 0.99 / 0.99 |

*Tabla 4- Comparación métricas de DecisionTreeClassifier antes y después de la optimización de hiperparámetros*

Utilizando el algoritmo LogisticRegression, se obtuvo un recall de 0.81 para la clase 0 (funcionarios no sancionados) y un recall de 0.60 para la clase 1 (funcionarios sancionados). Después de la optimización de los hiperparámetros, se logró mantener el mismo rendimiento en términos de recall para ambas clases, es decir, el recall de la clase 0 se mantuvo en 0.81 y el recall de la clase 1 se mantuvo en 0.60, Tabla 5.

| **Métrica** | **Antes de la optimización** | **Después de la optimización** |
|---|---|---|
| Precisión | 1.00 / 0.02 | 1.00 / 0.02 |
| Recall | 0.81 / 0.60 | 0.81 / 0.60 |
| F1-score | 0.89 / 0.04 | 0.89 / 0.04 |
| Accuracy | 0.81 | 0.81 |
| Macro avg | 0.51 / 0.71 | 0.51 / 0.71 |

| | | |
|---|---|---|
| Weighted avg | 0.99 / 0.81 | 0.99 / 0.81 |

*Tabla 5- Comparación métricas de LogisticRegression antes y después de la optimización de hiperparámetros*

## 5. Conclusiones

Contar con un conjunto de datos robusto y equilibrado es primordial para obtener predicciones, clasificaciones y agrupamientos precisos. Mientras más balanceados estén los datos que se están analizando, el modelo será capaz de aprender con mayor eficiencia, lo que se traduce en predicciones más precisas.

Las etapas de comprensión del negocio y de los datos desempeñan un papel vital en el aprovechamiento de los datos. No basta con realizar el preprocesamiento de estos y aplicar algoritmos de Machine Learning. Si no se cuenta con una comprensión adecuada de las necesidades funcionales de los usuarios, por un lado, se corre el riesgo de llegar a conclusiones erróneas y, por otro lado, se podrían obtener conclusiones que carecen de valor para el negocio.

Sin duda, la etapa que demanda mayor esfuerzo y trabajo en la minería de datos es la preparación de los datos. En el mundo real, los datos suelen ser mayormente no estructurados y se encuentran dispersos en diferentes bases de datos y archivos de almacenamiento. Esta etapa requiere una cuidadosa extracción, transformación y carga (ETL) de los datos para lograr una estructura coherente y lista para su análisis posterior.

Los datos abiertos del gobierno colombiano se revelaron como una valiosa fuente de información para este proyecto de investigación. Al combinarlos con diversas bases de datos, se logró establecer una sólida estructura inicial para el análisis de datos. Esta integración de fuentes de datos diversas permitió obtener una visión más completa y enriquecedora de los patrones y tendencias presentes en los datos.

Aunque la optimización de hiperparámetros desempeñó un papel importante en mejorar el rendimiento de los modelos, se observó que el factor más determinante para evaluar los modelos con mayor rendimiento fue el balanceo de los datos. El equilibrio entre las clases en el conjunto de datos resultó ser una variable crucial para obtener modelos más precisos y confiables. Esto subraya la importancia de abordar adecuadamente el desafío del desequilibrio de clases al realizar análisis y predicciones.

Es esencial destacar que los algoritmos de Machine Learning, aunque son poderosos y eficaces, requieren de una revisión posterior por parte de expertos funcionales. Estos expertos tienen la tarea de validar las salidas del modelo y verificar la ausencia de sesgos o información inexacta. Su experiencia y conocimiento permiten identificar posibles problemas y asegurar que los resultados del modelo sean coherentes y confiables. Esta validación por parte de expertos es una práctica recomendada para garantizar la calidad y utilidad de los resultados obtenidos.

# Referencias

[1] C. K. F. &. G. J. M. Schröer, «A Systematic Literature Review on Applying CRISP-DM Process Model.,» *Procedia Computer Science,* vol. 181, nº 1, pp. 526-534, 2021.

[2] P. G. d. l. Nación, «Objetivos y funciones.,» 2023. [En línea]. Available: https://www.procuraduria.gov.co/procuraduria/conozca-entidad/Pages/objetivos-funciones.aspx. [Último acceso: 2023].

[3] R. Martínez, «La analítica de la corrupción,» *R.I.T.I.,* vol. 7, pp. 1-5, 2018.

[4] R. N. d. E. Civil, «¿Cuáles son las clases de elecciones?.,» 2023. [En línea]. Available: https://www.registraduria.gov.co/Cuales-son-las-clases-de-elecciones.html. [Último acceso: 2023].

[5] L. S. Pazos, Las Evaluaciones de Políticas Públicas en el Estado Liberal, Cali: Universidad del Valle, 2004.

© 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).