

Reporte de selección y parametrización de modelos

Índice

1 Problema de negocio.....	3
1.1 Contexto.....	3
1.2 Planteamiento del problema de negocio.....	3
1.3 Requerimientos.....	3
2 Tratamiento previo de los datos.....	4
2.1 Preprocesamiento.....	4
2.1.1 Valores atípicos.....	4
2.1.2 Imputación de valores nulos.....	7
2.1.3 Extracción de variables no explicativas.....	8
2.2 Datos para predicción de adiciones y prórrogas.....	9
2.3 Datos para predicción de estados finales.....	12
2.4 Preselección de variables explicativas.....	16
3 Introducción a los modelos.....	16
3.1 Definición de métricas adecuadas de desempeño.....	18
3.1.1 Clustering.....	18
3.1.2 Clasificación.....	18
4 Modelo descriptivo.....	19
4.1 Procesos de calibración o entrenamiento.....	19
4.1.1 Selección de variables.....	19
4.1.2 Pruebas de verificación de los supuestos requeridos.....	19
4.1.3 Parametrización.....	20
4.1.4 Definición de métricas adecuadas de desempeño.....	22
4.2 Análisis de resultados.....	23
5 Modelo estados finales.....	25
5.1 Procesos de calibración o entrenamiento.....	25
5.1.1 Selección de variables.....	26
5.1.2 Pruebas de verificación de los supuestos requeridos.....	27
5.1.3 Parametrización.....	32
5.2 Análisis de resultados.....	34
6 Modelo adiciones.....	37
6.1 Procesos de calibración o entrenamiento.....	37
6.1.1 Selección de variables.....	37
6.1.2 Pruebas de verificación de los supuestos requeridos.....	37
6.1.3 Parametrización.....	41
6.2 Análisis de resultados.....	43
7 Modelo prórrogas.....	46

7.1 Procesos de calibración o entrenamiento.....	46
7.1.1 Selección de variables.....	46
7.1.2 Pruebas de verificación de los supuestos requeridos.....	46
7.1.3 Parametrización.....	50
7.2 Análisis de resultados.....	52
8 Implementación de ajustes.....	55
8.1 Modelos.....	55
8.2 Esquema general de solución.....	56
8.3 Problema de negocio.....	57
9 Plan de implementación del prototipo.....	58
10 Referencias.....	60
11 Anexos.....	61
11.1 Tratamiento previo de los datos.....	61
11.1.1 Valores nulos.....	61
11.1.1.1 Valores nulos SECOP I.....	61
11.1.2 Datos para predicción de adiciones y prórrogas.....	64
11.1.3 Datos para predicción de estados finales.....	67
11.2 Introducción a los modelos.....	73
11.2.1 Modelos de clasificación lineal.....	73
11.2.2 Modelos lineales generalizados.....	75
11.2.3 Modelos basados en particiones.....	75
11.2.4 Modelos no paramétricos.....	79
11.2.5 Métricas de desempeño.....	79
11.3 Modelo estados finales.....	81
11.3.1 Análisis de resultados.....	81
11.4 Modelo adiciones.....	83
11.4.1 Análisis de resultados.....	83
11.5 Modelo prórrogas.....	84
11.5.1 Análisis de resultados.....	84

1 Problema de negocio

1.1 Contexto

Las inversiones en infraestructura se realizan con recursos públicos del estado entregados a particulares para que desarrollen obras de infraestructura, por lo cual se pretende identificar diferencias entre los cronogramas y presupuestos de la etapa de planeación y ejecución, que en conjunto con datos históricos de los contratos ejecutados permitan generar alertas que le faciliten a los actores la mitigación de riesgos potenciales en la formulación de los contratos, lo anterior debido a que históricamente en la contratación pública se contratan las ofertas más económicas pero en el momento de su ejecución, por problemas de planeación o desconocimiento, se generan adiciones en recursos y prórrogas en tiempo, provocando que al final de la ejecución del contrato éste termine siendo mucho más costoso que las otras ofertas presentadas, con los riesgos de corrupción e incumplimiento relacionados.

1.2 Planteamiento del problema de negocio

Las Entidades Estatales realizaron durante los últimos años inversiones por más de 30 billones de pesos anuales con recursos públicos entregados a particulares para que desarrollen, entre otras, obras de infraestructura; pero desafortunadamente entre el 2016 y 2020 se han perdido 13 billones de pesos en hechos de corrupción (Transparencia por Colombia), por lo que se pretende identificar características en los contratos que permitan anticipar sobrecostos como prórrogas y adiciones o pérdidas por suspensiones en la contratación, generando alertas desde las etapas previas o iniciales de contratación que le faciliten a los actores (Entidades estatales, Contratistas, Entes de Control y Sociedad Civil) la prevención y mitigación de riesgos potenciales en los contratos.

1.3 Requerimientos

El proyecto pretende entregar una herramienta que al aplicarla sobre los contratos que están en etapa de borrador, de aprobación o de inicio de ejecución, permita identificar alertas sobre problemas en sus características actuales y predicciones sobre si terminarían con dificultades en su cierre, conllevando a generar sobrecostos como adiciones y prórrogas.

Estas alertas se generarán a partir del entendimiento de la eficiencia en la contratación, mediante la descripción de datos históricos, donde se realizan comparaciones entre lo presupuestado vs lo realmente ejecutado y se identifican características en estos que generan riesgo como adiciones presupuestales y prórrogas de tiempo.

Estas alertas pueden constituirse en banderas rojas y ser utilizadas para la toma de decisiones de los encargados de la ejecución y la auditoría, y también se constituyen como un indicador para los entes de control permitiéndoles enfocar sus esfuerzos.

Tabla: Requerimientos de negocio - Etapa inicial - Prototipo

Aspecto	Nombre	Requerimiento
Negocio		
R1	Artefacto descriptivo	Desarrollar un artefacto que proporcione a los usuarios información de la contratación, el cual se podría utilizar para mejorar la comprensión y el entendimiento del contexto de la contratación de tipo Obra.
R2	Artefacto de predicción de estados finales	Desarrollar un artefacto que permita predecir los estados en los que terminará un contrato, el cual se podría utilizar como alerta para tomar decisiones para mejorar la eficiencia en la contratación y generar ahorros.
R3	Artefacto de predicción de adiciones	Desarrollar un artefacto que permita predecir el riesgo de adiciones presupuestales durante el proceso de contratación, el cual se podría utilizar para tomar decisiones para mejorar la eficiencia en la contratación y generar ahorros.
R4	Artefacto de predicción de prórrogas	Desarrollar un artefacto que permita predecir el riesgo de ocurrencia de prórrogas en los que se incurría en el contrato, el cual se podría utilizar para tomar decisiones para mejorar la eficiencia en la contratación y generar ahorros.

2 Tratamiento previo de los datos

A partir de los requerimientos planteados y validando las fuentes de datos, se decide tomar como fuente de información la base de datos entregada por SECOP I.

2.1 Preprocesamiento

Se retoma el preprocesamiento de datos realizado en el Anteproyecto (Alfaro Rojas et al., 2022) en cuanto a:

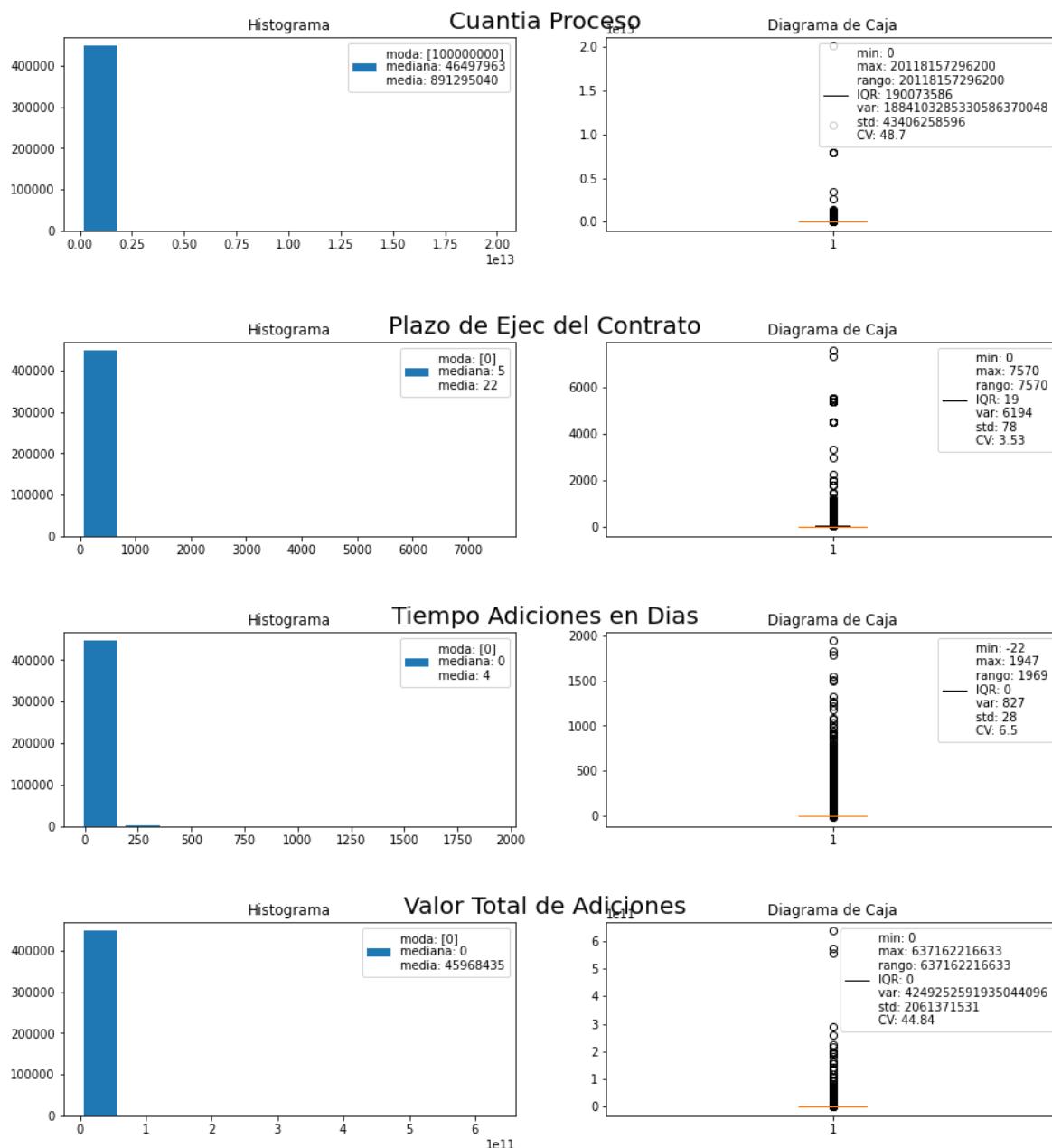
- Técnicas de limpieza de datos (Valores atípicos, Valores nulos)
- Imputación de valores nulos
- Extracción de variables explicativas o features

2.1.1 Valores atípicos

Producto del preprocesamiento realizado, teniendo en cuenta el gran volumen del dataset y con la intención de enfocar el esfuerzo del trabajo para que los resultados puedan ser interpretables, se filtran y seleccionan los contratos de tipo 'Obra' eliminado del análisis otros tipos de contratos como los de prestación de servicios, comodato o los de concesiones que tienen valores y características distintas entre ellos. Este proceso reduce los registros a 448.205 contratos de tipo 'Obra' en SECOP I.

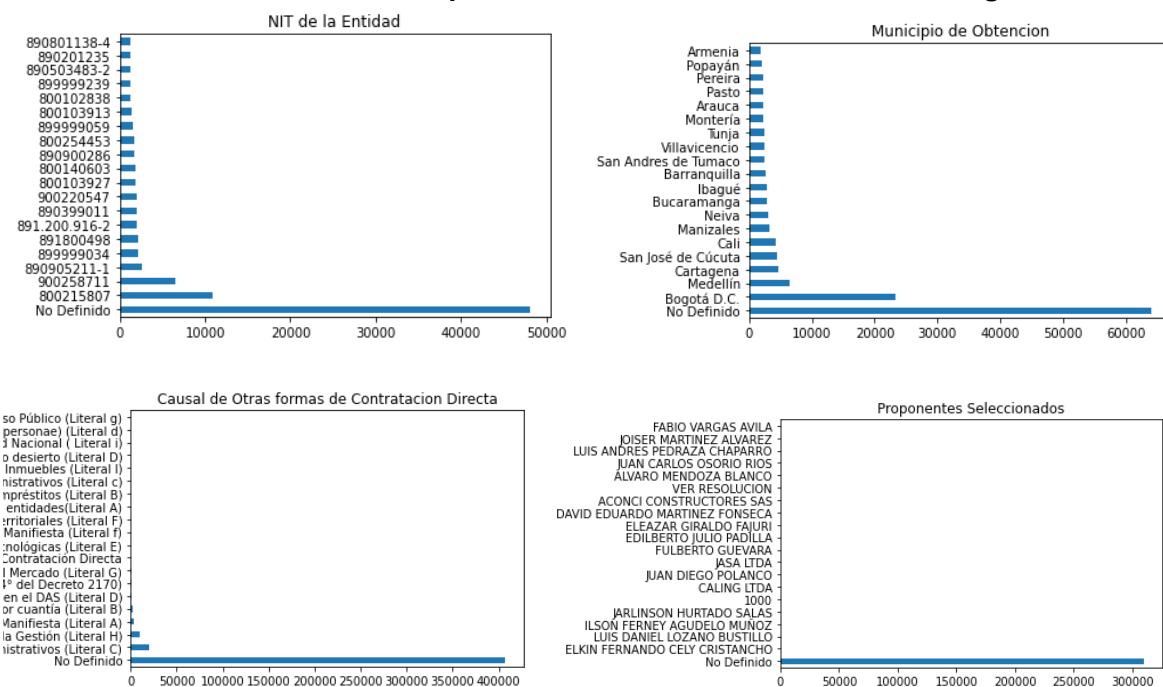
Se buscan valores atípicos en las diferentes columnas.

Gráficas: Valores atípicos en columnas numéricas



Los valores atípicos presentados en las columnas numéricas pertenecen a contratos de mayor cuantía los cuales al ser verificados individualmente se identifican casos como el contrato con identificación 11-1-67586 con objeto de “MEJORAMIENTO Y PAVIMENTACION DE LA VIA ENTRE PUENTE NACIONAL Y GUAVATA” por un valor de \$11.111.111.111.111 y el contrato con identificación 14-1-112756 con objeto “CONSTRUCCION DE DOS CENTROS DE DESARROLLO INFANTIL EN LOS MUNICIPIOS DE ALGECIRAS Y TARQUI” por un valor de \$20.118.157.296.200, los cuales claramente son errores de digitación que afectan significativamente las estadísticas descriptivas de tendencia y dispersión así como el análisis y entendimiento de los datos, razón por la cual se decide eliminar del dataset los registros con Cuantía de Proceso mayores a cien mil millones.

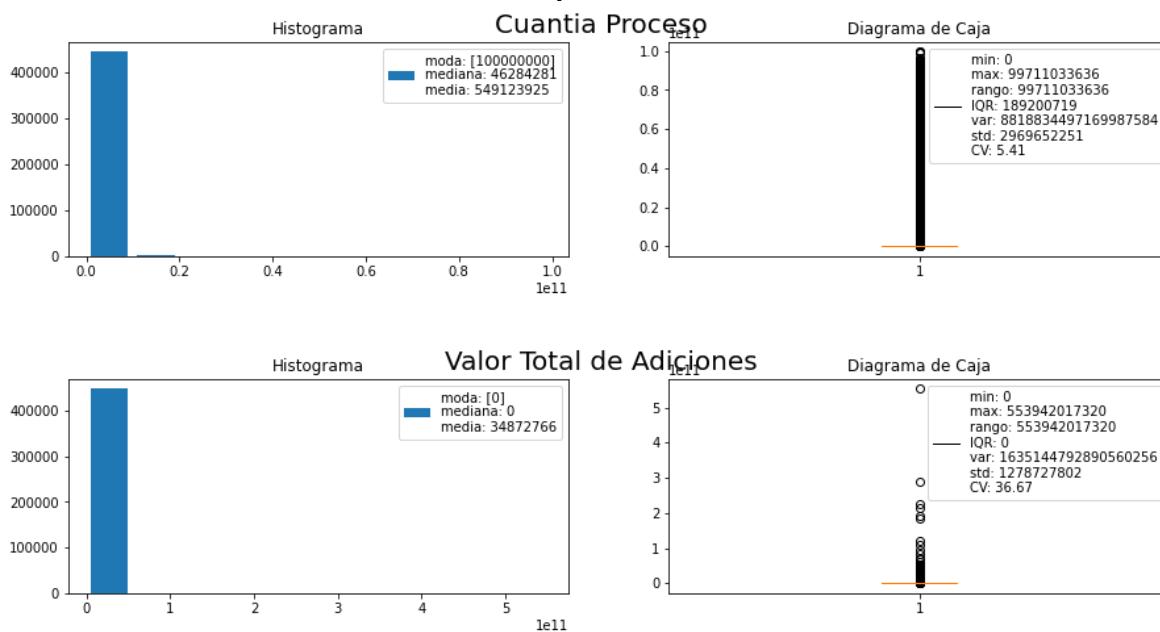
Gráficas: Valores atípicos en cantidades de columnas categóricas



Los valores atípicos presentados en las variables categóricas tienen que ver con la categoría de No Definido, la cual parece ser el valor por defecto, por lo cual estas variables serán revisadas en la parte de entendimiento de los datos para verificar si pueden ser anormales o se pueden complementar para obtener mayor información.

Al aplicar la limpieza de los datos, eliminando los contratos con un valor de Cuantía de Proceso mayor a cien mil millones se reduce el dataset a 447.876 registros.

Gráficas: Valores atípicos en columnas numéricas



El promedio de la Cuantía de Proceso pasa de \$891 millones a \$549 millones y el promedio del Valor Total de Adiciones pasa de \$45 millones a \$34 millones.

2.1.2 Imputación de valores nulos

Producto del preprocesamiento realizado se obtienen las siguientes imputaciones para valores nulos.

En el anexo (Ver [Valores nulos SECOP I](#)) se presenta el proceso de identificación de valores nulos en SECOP I y en la siguiente tabla se presenta el tratamiento realizado a los valores nulos encontrados.

Tabla: Imputación de valores nulos

Columna	Non-Null	Tratamiento
Detalle del Objeto a Contratar	447851	Se reemplaza por el valor “SinInformacion”
Municipios Ejecucion	0	Por la cantidad, se elimina la columna
Número de Proceso	447863	Se reemplaza por el valor “SinInformacion”
Número de Contrato	447734	Se reemplaza por el valor “SinInformacion”
Identificación del Contratista	447852	Se reemplaza por el valor “SinInformacion”
Nom Razon Social Contratista	447868	Se reemplaza por el valor “SinInformacion”
Identific Representante Legal	447385	Se reemplaza por el valor “SinInformacion”
Nombre del Represen Legal	447239	Se reemplaza por el valor “SinInformacion”
Fecha de Firma del Contrato	338272	Se reemplaza por el campo “Fecha de Cargue en el SECOP”
Fecha Ini Ejec Contrato	338272	Se reemplaza por el campo “Fecha de Cargue en el SECOP”
Fecha Fin Ejec Contrato	337498	Es normal cuando el contrato no ha terminado por lo que no se realiza tratamiento
Cuantía Contrato	338272	Se reemplaza por el campo “Cuantía Proceso”
Objeto del Contrato a la Firma	447865	Se reemplaza por el valor “SinInformacion”
Proponentes Seleccionados	446187	Se reemplaza por el valor “SinInformacion”
Calificación Definitiva	445606	Se reemplaza por el valor “SinInformacion”
Fecha Liquidación	149370	Es normal cuando el contrato no ha terminado por lo que no se realiza tratamiento

Todos los campos que son reemplazados por “SinInformacion” se imputan de esta manera porque por el momento no hay forma de identificar o sustituir los datos por los reales y esta imputación no afecta la información que están aportando a las preguntas planteadas en el proyecto.

Las fechas de Firma Del Contrato y Fecha Ini Ejec Contrato se reemplaza por “Fecha de Cargue en el SECOP”, esta decisión se toma porque al analizar y comparar las fechas que sí

contienen información, se logra evidenciar que estas fechas normalmente son las mismas o tiene una varianza entre 2 y 3 días que no afectan su información.

2.1.3 Extracción de variables no explicativas

Producto del preprocesamiento realizado se obtienen las siguientes variables explicativas, después de eliminar las variables que contenían un gran porcentaje de nulos y las que no aportaban información para la pregunta de negocio:

Tabla: Variables explicativas

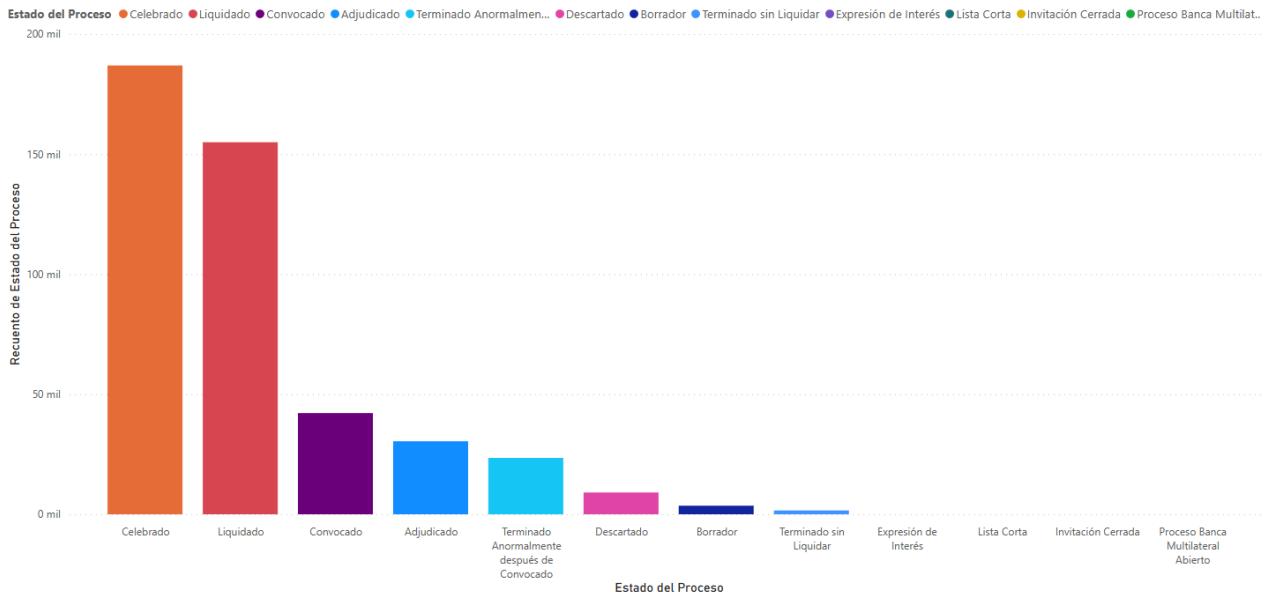
#	Columna	Importa	#	Columna	Importa
1	UID	B	33	ID Adjudicacion	B
2	Anno Cargue SECOP	A	34	Tipo Identifi del Contratista	A
3	Anno Firma Contrato	A	35	Identificacion del Contratista	A
4	Nivel Entidad	B	36	Nom Razon Social Contratista	A
5	Orden Entidad	A	37	Dpto y Muni Contratista	B
6	Nombre Entidad	B	39	Identific Representante Legal	B
7	NIT de la Entidad	B	40	Nombre del Represen Legal	B
8	Código de la Entidad	B	41	Fecha de Firma del Contrato	A
9	ID Modalidad	A	42	Fecha Ini Ejec Contrato	A
10	Modalidad de Contratacion	A	43	Plazo de Ejec del Contrato	A
11	Estado del Proceso	A	44	Rango de Ejec del Contrato	A
15	ID Objeto a Contratar	A	45	Tiempo Adiciones en Dias	A
16	Objeto a Contratar	A	46	Tiempo Adiciones en Meses	B
17	Detalle del Objeto a Contratar	A	47	Fecha Fin Ejec Contrato	A
19	Municipio de Obtencion	A	49	Cuantia Contrato	A
20	Municipio de Entrega	A	50	Valor Total de Adiciones	A
23	Numero de Constancia	B	51	Valor Contrato con Adiciones	A
24	Numero de Proceso	B	52	Objeto del Contrato a la Firma	A
25	Numero de Contrato	B	53	Proponentes Seleccionados	B
26	Cuantia Proceso	A	54	Calificacion Definitiva	B
27	ID Grupo	A	59	Es PostConflict	B
28	Nombre Grupo	A	67	Municipio Entidad	A
29	ID Familia	A	68	Departamento Entidad	A
30	Nombre Familia	A	70	Fecha Liquidacion	A
31	ID Clase	A	71	Cumple Decreto 248	B

#	Columna	Importa	#	Columna	Importa
32	Nombre Clase	A	72	IncluyeBienesDecreto248	C

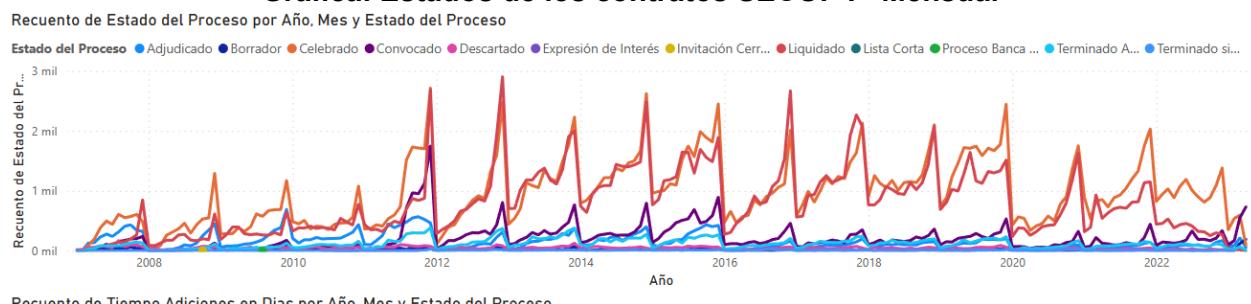
2.2 Datos para predicción de adiciones y prórrogas

Se generan las siguientes gráficas (ver más gráficas en los anexos [Datos para predicción de adiciones y prórrogas](#)) para el entendimiento de los datos:

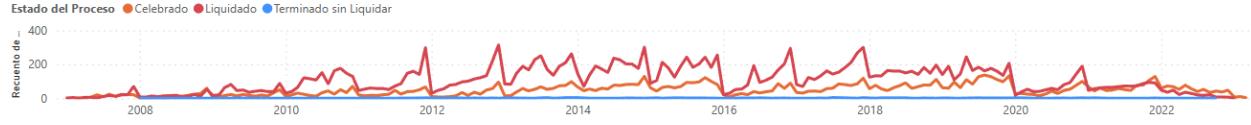
Gráfica Estados de los contratos SECOP I



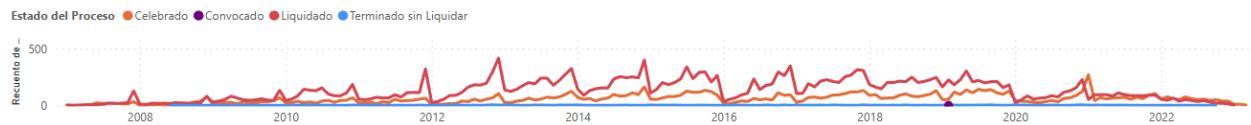
Gráfica: Estados de los contratos SECOP I - Mensual



Recuento de Tiempo Adiciones en Días por Año, Mes y Estado del Proceso

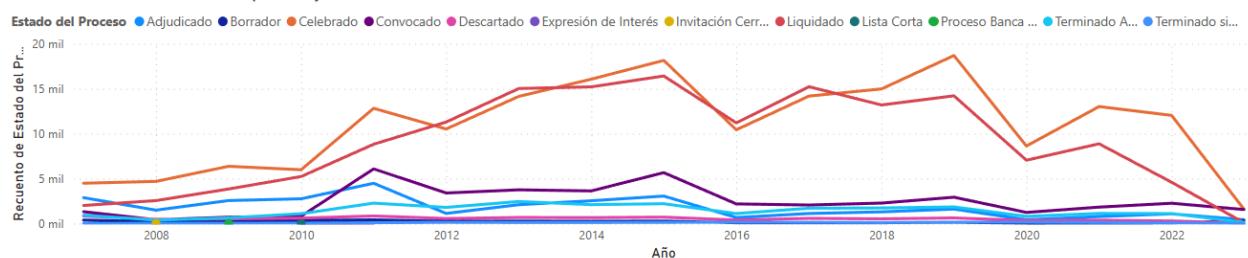


Recuento de Valor Total de Adiciones por Año, Mes y Estado del Proceso

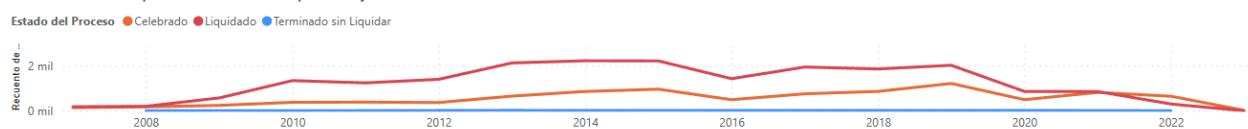


Gráfica: Estados de los contratos SECOP I - Anual

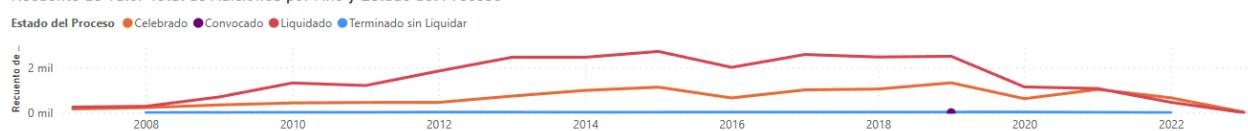
Recuento de Estado del Proceso por Año y Estado del Proceso



Recuento de Tiempo Adiciones en Días por Año y Estado del Proceso



Recuento de Valor Total de Adiciones por Año y Estado del Proceso

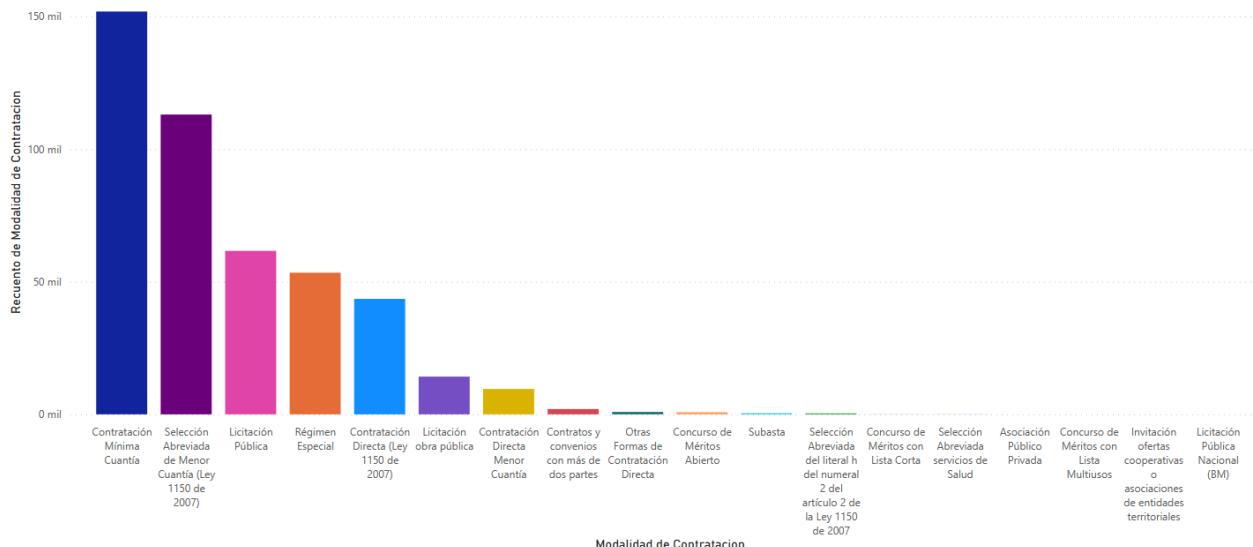


A partir de las gráficas se puede inferir que:

- Las prórrogas sólo aparecen en los estados Celebrado, Liquidado y Terminado sin Liquidar
- Las adiciones sólo aparecen en los estados Celebrado, Convocado, Liquidado y Terminado sin Liquidar
- Durante los años 2011, 2015 y 2019 se evidencia una disminución en la liquidación de contratos, probablemente afectado por la ley de garantías de las elecciones de alcaldía.

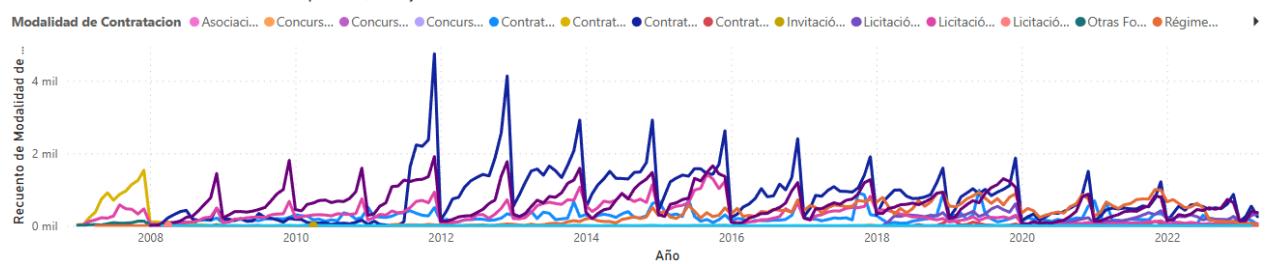
Gráfica: Modalidad de los contratos SECOP I

Modalidad de Contratación

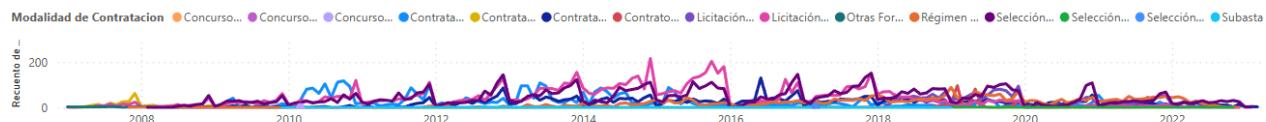


Gráfica: Modalidad de los contratos SECOP I - Mensual

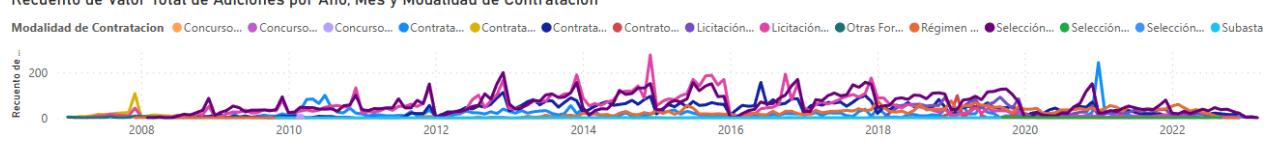
Recuento de Modalidad de Contratación por Año, Mes y Modalidad de Contratación



Recuento de Tiempo Adiciones en Días por Año, Mes y Modalidad de Contratación

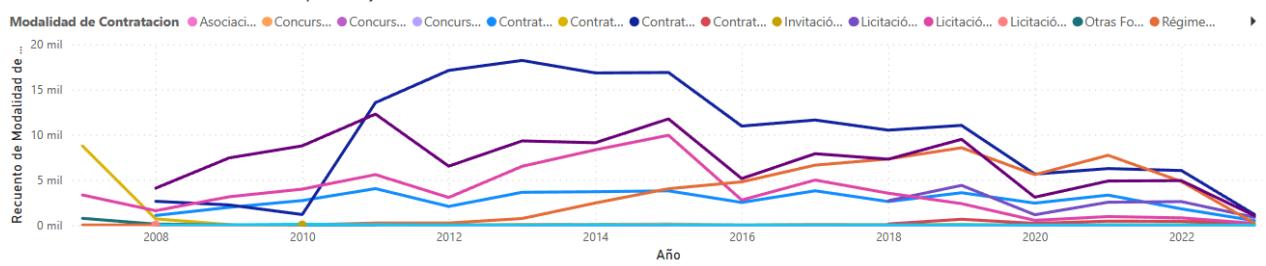


Recuento de Valor Total de Adiciones por Año, Mes y Modalidad de Contratación



Gráfica: Modalidad de los contratos SECOP I - Anual

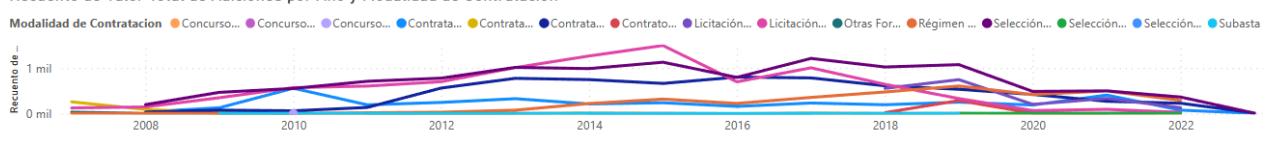
Recuento de Modalidad de Contratación por Año y Modalidad de Contratación



Recuento de Tiempo Adiciones en Días por Año y Modalidad de Contratación



Recuento de Valor Total de Adiciones por Año y Modalidad de Contratación



A partir de las gráficas se puede inferir que:

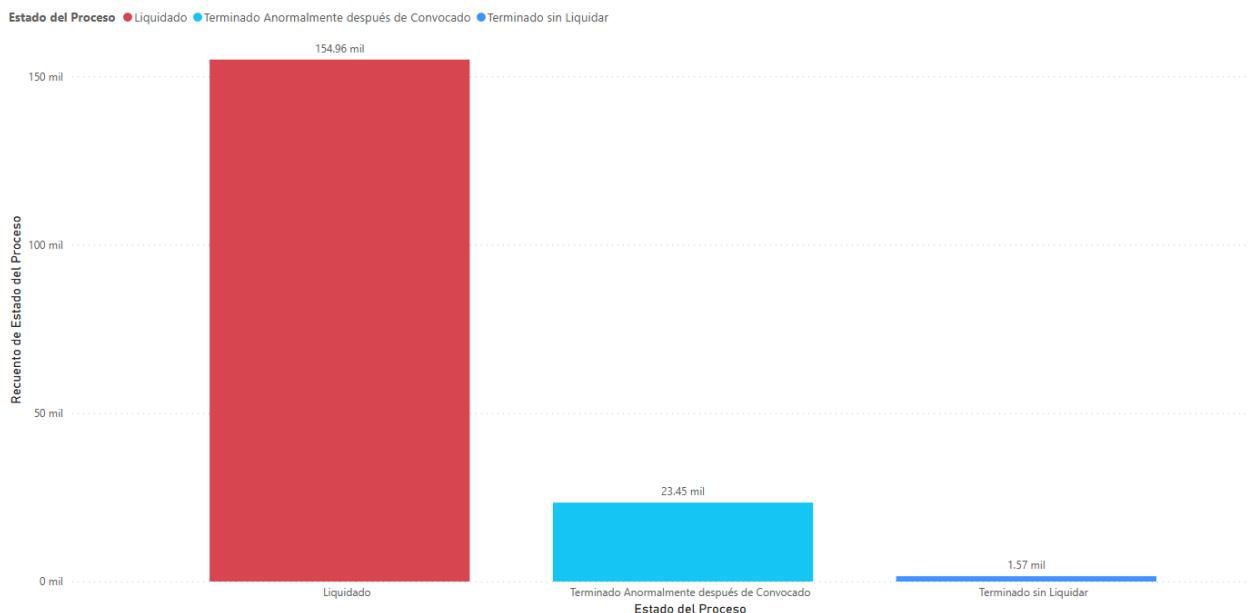
- De 2011 a 2015 se evidencia una fuerte influencia de la Modalidad de contratación de Contratación de mínima cuantía.
- En 2020 se evidencia una alta variación en los patrones de los datos, quizás por los efectos de la pandemia COVID19.

Teniendo en cuenta lo anterior para los modelos de predicciones se debería utilizar los años 2016 a 2019.

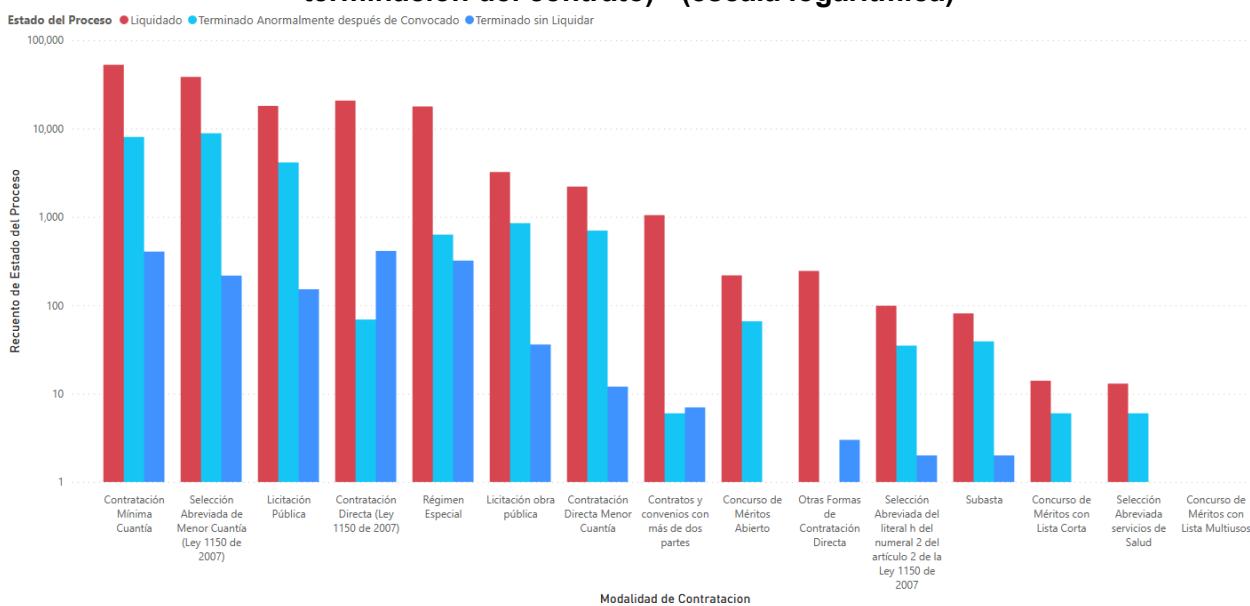
2.3 Datos para predicción de estados finales

Se generan las siguientes gráficas (ver más gráficas en los anexos [Datos para predicción de estados finales](#)) para el entendimiento de los datos:

Gráfica: Estados de finalización de los contratos SECOP I



Gráfica: Modalidades de contratación SECOP I - Anual - (filtrado por Estado de terminación del contrato) - (escala logarítmica)

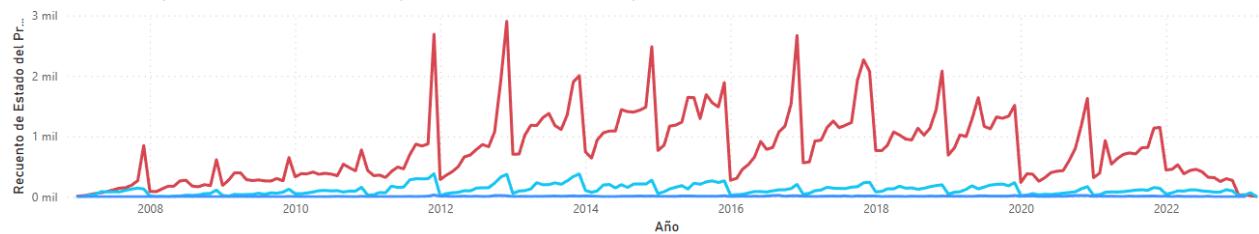


En la gráfica se evidencia que la mayoría de Modalidades de contratación presentan todos los Estados de terminación del contrato.

Gráfica: Estados de finalización de los contratos SECOP I - Mensual

Recuento de Estado del Proceso por Año, Mes y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado Anormalmente después de Convocado ● Terminado sin Liquidar



Recuento de Tiempo Adiciones en Días por Año, Mes y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



Recuento de Valor Total de Adiciones por Año, Mes y Estado del Proceso

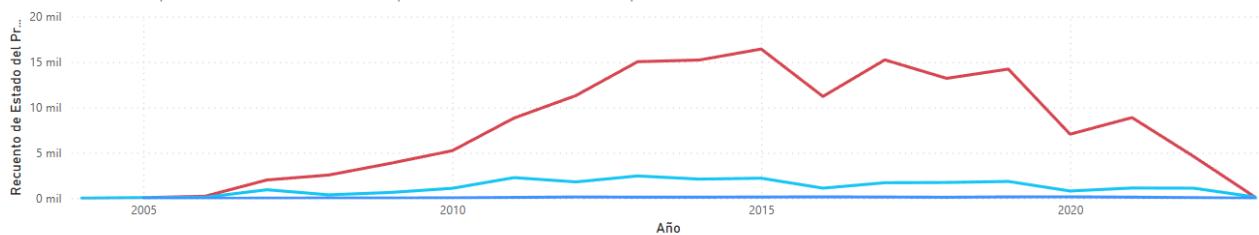
Estado del Proceso ● Liquidado ● Terminado sin Liquidar



Gráfica: Estados de finalización de los contratos SECOP I - Anual

Recuento de Estado del Proceso por Año y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado Anormalmente después de Convocado ● Terminado sin Liquidar



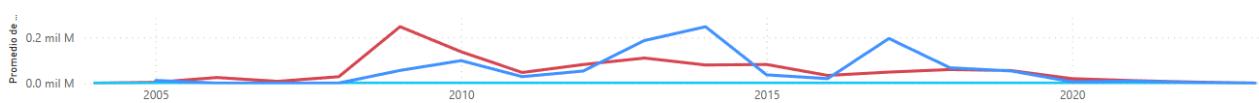
Promedio de Tiempo Adiciones en Días por Año y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado Anormalmente después de Convocado ● Terminado sin Liquidar



Promedio de Valor Total de Adiciones por Año y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado Anormalmente después de Convocado ● Terminado sin Liquidar



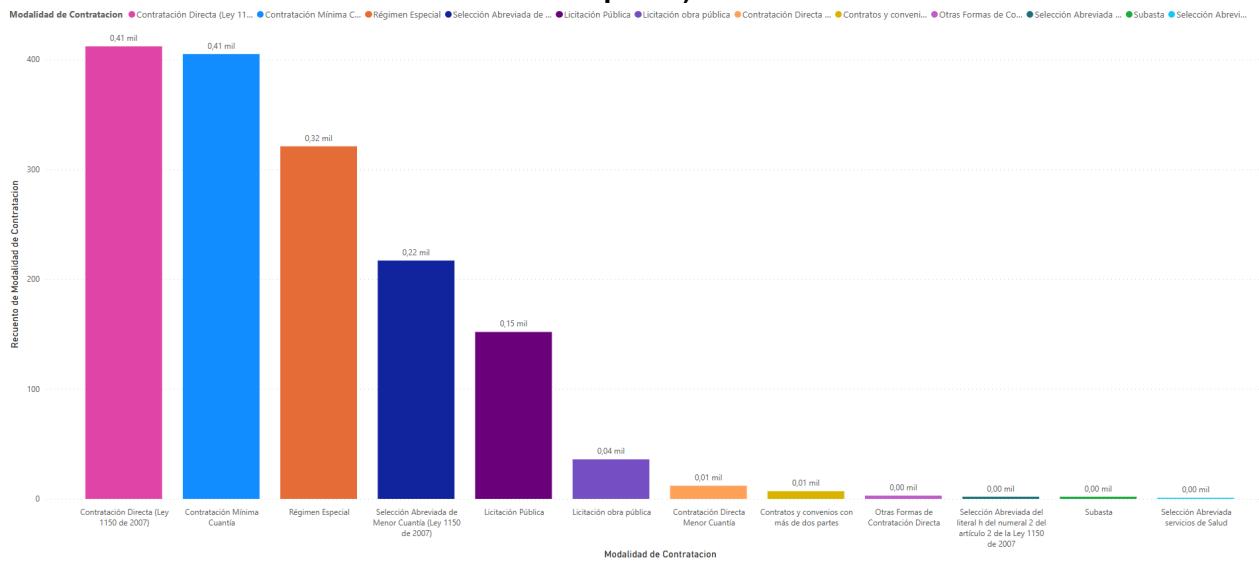
En el año 2020 se evidencian cambios en los patrones de las gráficas, probablemente debido a la pandemia COVID19, por lo que quizás no debería utilizarse ese año para los modelos de predicción.

Gráfica: Estados de finalización de los contratos SECOP I - Anual - (Filtrando con Multas y Sanciones)



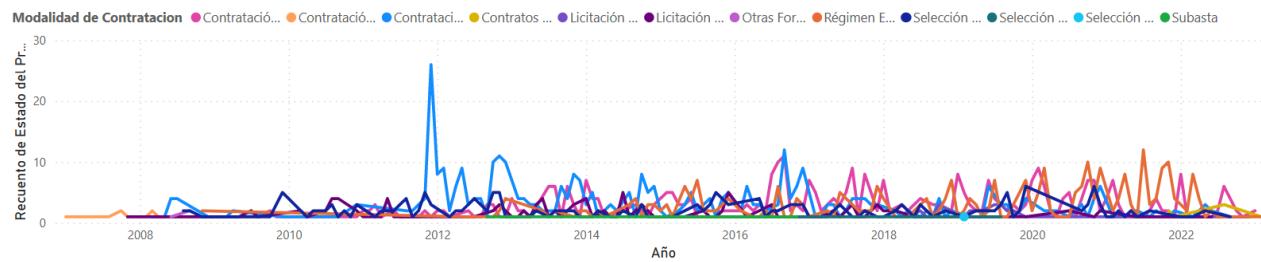
Se filtran los contratos con la base de datos de Multas y Sanciones, pero el resultado del cruce es mínimo para el estado de Terminado sin Liquidar por lo que esta base de datos no se incluye en el análisis de los datos.

Gráfica: Modalidades de contratación SECOP I - (filtrado por Estado Terminado sin Liquidar)



Gráfica: Modalidades de contratación SECOP I - Mensual - (filtrado por Estado Terminado sin Liquidar)

Recuento de Estado del Proceso por Año, Mes y Modalidad de Contratacion



Recuento de Tiempo Adiciones en Dias por Año, Mes y Modalidad de Contratacion

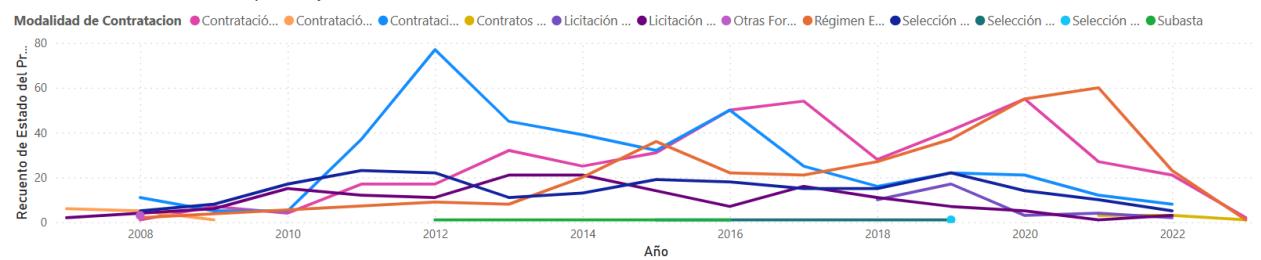


Recuento de Valor Total de Adiciones por Año, Mes y Modalidad de Contratacion



Gráfica: Modalidades de contratación SECOP I - Anual - (filtrado por Estado Terminado sin Liquidar)

Recuento de Estado del Proceso por Año y Modalidad de Contratacion



Recuento de Tiempo Adiciones en Dias por Año y Modalidad de Contratacion



Recuento de Valor Total de Adiciones por Año y Modalidad de Contratacion



Para predecir los Estados de terminación del contrato, según la teoría, no se deberían tener en cuenta las modalidades de contratación que no requieren liquidación del contrato, pero en las gráficas se evidencia que todas las modalidades de contratación están relacionadas con el estado Terminado sin liquidar, por lo que, en la predicción de este estado, el usuario debe decidir si es algo normal o es una alerta de acuerdo a los lineamientos de la modalidad de contratación en su entidad.

2.4 Preselección de variables explicativas

Para seleccionar las variables más importantes para el proyecto de análisis de eficiencia en la contratación y desarrollar una herramienta que permita realizar las predicciones planteadas con los datos de SECOP I, específicamente se seleccionaron los años del 2016 al 2019 debido a que estos años contaban con información típica. El año 2020 fue descartado debido a que por motivos del COVID19 el comportamiento de los contratos no tuvo un comportamiento normal, lo que podría afectar la precisión de los modelos al incluir datos atípicos.

Inicialmente, con estos filtros, la base de datos contenía 135.710 filas y 72 columnas, se eliminaron columnas duplicadas que podrían generar correlaciones al correr los modelos o que no aportaban información al problema de negocio, como por ejemplo Nombre Grupo vs ID Grupo. Las columnas eliminadas fueron: NIT de la Entidad, ID Modalidad, ID Regimen de Contratacion, ID Objeto a Contratar, Municipios Ejecucion, ID Grupo, ID Familia, ID Clase, ID Adjudicacion, Objeto del Contrato a la Firma, ID Sub Unidad Ejecutora, Ruta Proceso en SECOP I, Moneda, Última Actualización, Cumple Decreto 248, IncluyeBienesDecreto248, Posición Rubro, Nombre Rubro, Valor Rubro, Sexo RepLegal, Nombre Entidad, Objeto a Contratar, Numero de Constancia, Identificación del Contratista, Nombre Sub Unidad Ejecutora, Calificación Definitiva, Pilar Acuerdo Paz, Detalle del Objeto a Contratar, Municipio Entidad, Compromiso Presupuestal, Modalidad de Contratacion, Tipo De Contrato, Código de la Entidad y Año Cargue Secop.

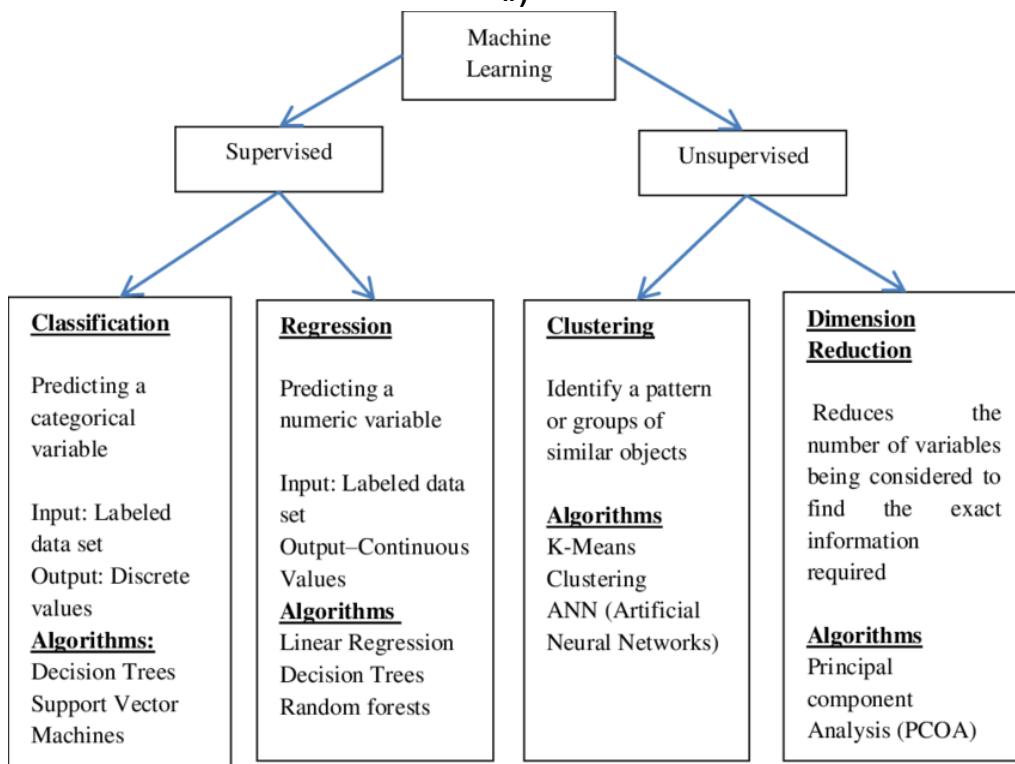
Después, se realizó una selección variable por variable y se relacionó con las demás variables para determinar si aportaba información importante no redundante con respecto a otras variables existentes. Esto llevó a la selección de las siguientes variables en un principio:

Nivel Entidad, Orden Entidad, Estado del Proceso, Nombre Regimen de Contratacion, Municipio de Obtencion, Fecha de Cargue en el SECOP, Numero de Proceso, Numero de Contrato, Objeto a Contratar, Cuantia Proceso, Nombre Grupo, Tipo Identifi del Contratista, Dpto y Muni Contratista, Tipo Doc Representante Legal, Nombre del Represen Legal, Fecha de Firma del Contrato, Fecha Ini Ejec Contrato, Plazo de Ejec del Contrato, Rango de Ejec del Contrato, Tiempo Adiciones en Dias, Tiempo Adiciones en Meses, Fecha Fin Ejec Contrato, Cuantia Contrato, Valor Total de Adiciones, Valor Contrato con Adiciones, Proponentes Seleccionados, Es PostConflict, Marcacion Adiciones, Punto Acuerdo Paz, Departamento Entidad y Fecha Liquidacion.

3 Introducción a los modelos

Machine Learning (ML) es un campo de estudio que involucra el desarrollo de algoritmos y modelos que permiten a las computadoras aprender de los datos para mejorar su rendimiento y hacer predicciones que permitirán tomar decisiones basadas en ese aprendizaje. Dentro del campo de machine learning hay diferentes áreas de especialidad entre las que se encuentran el aprendizaje supervisado y el aprendizaje no supervisado (Kuhn & Johnson, 2013, 41).

Imagen: Clasificación amplia de técnicas de aprendizaje automático (Suryakanthi, 2020, #)



El aprendizaje supervisado implica entrenar un modelo en un conjunto de datos etiquetados, donde se conocen las etiquetas. El objetivo es aprender una función que pueda predecir las etiquetas para datos nuevos de los que no se tiene información. El aprendizaje supervisado se basa en encontrar una función que pueda asignar características de entrada a etiquetas de salida, dado un conjunto de ejemplos de entrenamiento. En este caso el algoritmo aprende a predecir una variable objetivo (también conocida como variable dependiente) en función de los datos de entrada (también conocidos como características o variables independientes) tomando como base un conjunto de ejemplos para definir una función o modelo que permita completar los datos faltantes en la relación entre la variable dependiente y las variables independientes permitiendo realizar predicciones (James et al., 2013, 26).

En el aprendizaje no supervisado el algoritmo intenta identificar patrones en los datos sin etiquetas ni resultados predefinidos en los que no hay una variable dependiente de la que se tengan ejemplos conocidos. El algoritmo se entrena en un conjunto de datos sin etiquetar sin ningún conocimiento previo de la estructura de datos. El objetivo del aprendizaje no supervisado es identificar patrones o agrupaciones en los datos que se pueden usar para descubrir información y tomar decisiones informadas. A diferencia del aprendizaje supervisado, donde los datos de entrenamiento están etiquetados, los algoritmos de aprendizaje no supervisado tienen que encontrar patrones y agrupaciones ocultos en los datos por sí mismos. La agrupación en clusters ('clustering') es una tarea común en el aprendizaje no supervisado, donde el algoritmo se usa para agrupar puntos de datos similares (James et al., 2013, 26).

En el presente proyecto se utilizaran modelos de clustering y clasificación:

- El agrupamiento o ‘clustering’ es un tipo de aprendizaje no supervisado en el que el objetivo es agrupar observaciones similares en función de sus características, sin ningún conocimiento previo de los grupos o categorías. En los problemas de agrupamiento, el objetivo es encontrar conglomerados o grupos de observaciones que sean similares entre sí y diferentes entre los grupos en función de un conjunto de características.
- La clasificación es un tipo de aprendizaje supervisado donde el objetivo es predecir una variable de salida categórica o discreta, en función de un conjunto de características de entrada. En los problemas de clasificación, el objetivo es asignar una etiqueta a cada observación en función de un conjunto de características.

3.1 Definición de métricas adecuadas de desempeño

3.1.1 Clustering

Para el proceso de Clustering la definición de métricas de desempeño es de vital importancia ya que son estas las que permiten evaluar y comparar diferentes algoritmos. Estas, proporcionan una medida cuantitativa de qué tan bien se agruparon los datos y cuán similares son los puntos dentro de cada grupo. Adicionalmente también pueden ayudar a identificar posibles problemas con el clustering, como la presencia de grupos superpuestos o la falta de homogeneidad dentro de los grupos (Elizabeth León Guzmán, n.d.). Además, las métricas de desempeño pueden ser útiles para ajustar los parámetros de clustering y mejorar los resultados.

Para poder obtener los mejores resultados , se usaron las siguientes métricas:

Silhouette score: medida de la similitud entre un punto y los otros puntos de datos en su propio grupo, en comparación con otros grupos (*Algoritmo Meta-Heurístico Para Clustering Particional De Datos Basado En Global-Best Harmony Search, K-Means Y Restricted Growth Strings*, n.d.).

Índice de Davies-Bouldin: una medida de la similitud promedio entre cada grupo y su grupo más similar (*ALGORITMO DE SELECCIÓN Y VALIDACIÓN DEL MÉTODO DE CLUSTERIZACIÓN ÓPTIMO PARA DATOS NO SUPERVISADOS Universidad Tecnológica*, n.d.).

3.1.2 Clasificación

Para evaluar el desempeño de los diferentes modelos utilizados en la predicción de los modelos, se utilizaron tres métricas diferentes. La primera de ellas fue el MCE (Mínimo Error de Clasificación), que es una medida que muestra qué tan preciso es un modelo al ejecutar la tarea de clasificación.

Además, se utilizó el área bajo la curva ROC (AUC) para evaluar la eficiencia del modelo en la separación de las clases positivas y negativas. La curva ROC es una representación gráfica de la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) a diferentes umbrales de probabilidad. Cuanto mayor sea el AUC, mejor será la capacidad del modelo para distinguir entre las clases positivas y negativas.

Por último, se utilizó la matriz de confusión para evaluar el desempeño del modelo. La matriz de confusión muestra la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, lo que permite calcular la precisión, la sensibilidad, la especificidad y otras medidas de desempeño.

La matriz de confusión, de la que se generan las siguientes métricas (Aprende Machine Learning, 2019):

- La Accuracy del modelo es básicamente el número total de predicciones correctas dividido por el número total de predicciones.
- La Precisión de una clase define cuán confiable es un modelo en responder si un punto pertenece a esa clase.
- El Recall de una clase expresa cuán bien puede el modelo detectar a esa clase.
- El F1 Score de una clase combina precisión y recall en una sola métrica.

El uso de estas métricas permitió evaluar el desempeño de los modelos en diferentes aspectos y acorde a las características de los datos, que presentan clases muy desbalanceadas, y el problema de negocio, nos enfocamos principalmente en los resultados de Recall.

4 Modelo descriptivo

Con este modelo se pretende dar respuesta al Artefacto descriptivo (Desarrollar un artefacto que proporcione a los usuarios información de la contratación, el cual se podría utilizar para mejorar la comprensión y el entendimiento del contexto de la contratación de tipo Obra.), por lo que se busca, por medio de un modelo de clustering, entender las relaciones entre las variables y sus agrupaciones, de tal manera que se puedan presentar de una forma de fácil interpretación para el usuario final.

4.1 Procesos de calibración o entrenamiento

4.1.1 Selección de variables

Para esta base de datos se tomaron aproximadamente 14 variables, en las cuales tenemos variables cualitativas y cuantitativas, por lo que es necesario para las variables cualitativas generar un proceso de encoder, con el fin de poder correr el algoritmo. La definición de ellas se basó en la preselección de variables explicativas y el análisis de independencia, el cuál está soportado con la matriz de correlación.

Cuadro: Variables Seleccionadas para K-Means

```
contactos1=contactos1[['Orden Entidad', 'Modalidad de Contratacion', 'Estado del Proceso',
    'Plazo de Ejec del Contrato', 'Rango de Ejec del Contrato',
    'Tiempo Adiciones en Dias', 'Tiempo Adiciones en Meses',
    'Cuantia Contrato', 'Nombre Grupo',
    'Tipo Identifi del Contratista', 'Valor Total de Adiciones',
    'Valor Contrato con Adiciones', 'Departamento Entidad', 'YearIniEje','YearFinEje' ]]
contactos1
```

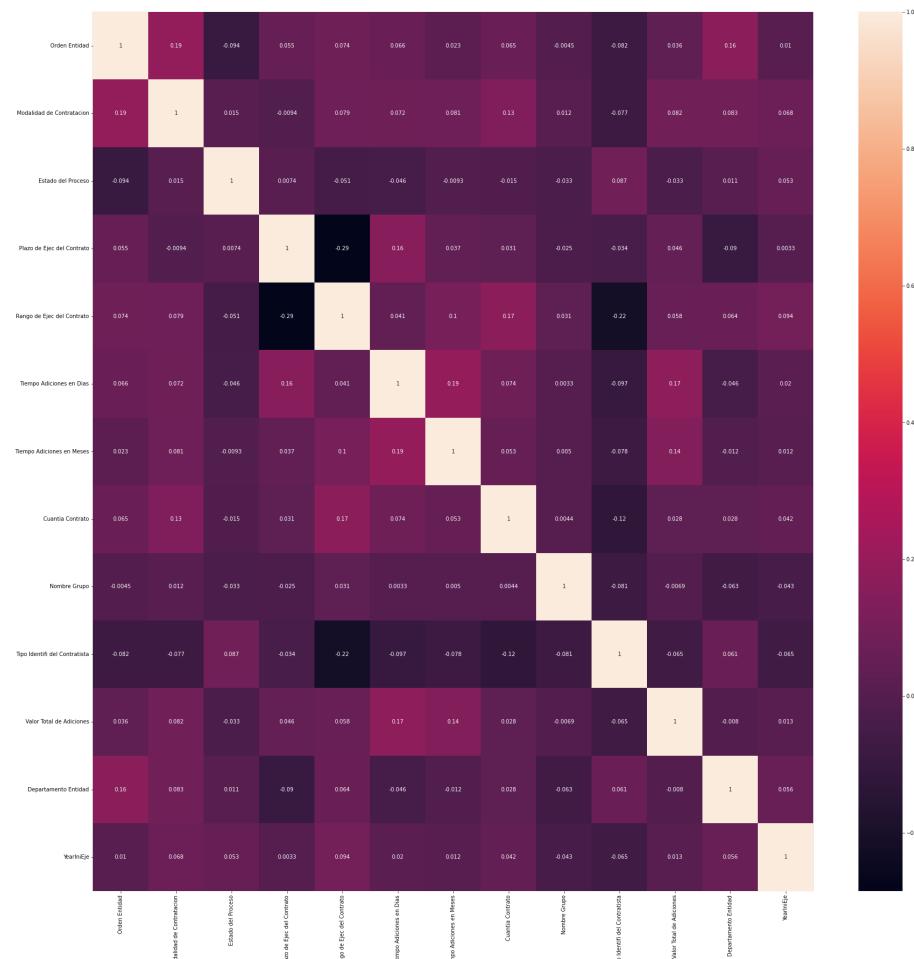
4.1.2 Pruebas de verificación de los supuestos requeridos

Para poder aplicar el modelo K-Means Clustering, después de los resultados de los procesos de limpieza, filtrado de variables e imputación de valores nulos; es necesario verificar los supuestos necesarios para determinar si se podría hacer una interpretación adecuada de los modelos. Para este caso, es necesario aplicar 2 supuestos, el primero es la Independencia de las variables y seguido de este la estandarización de las variables.

Tabla: Verificación de supuestos

	Supuestos	K-Means	PCA
Independencia	cumple		
Estandarizados	cumple	cumple	

Gráfica: Matriz de correlación variables seleccionadas para K-Means



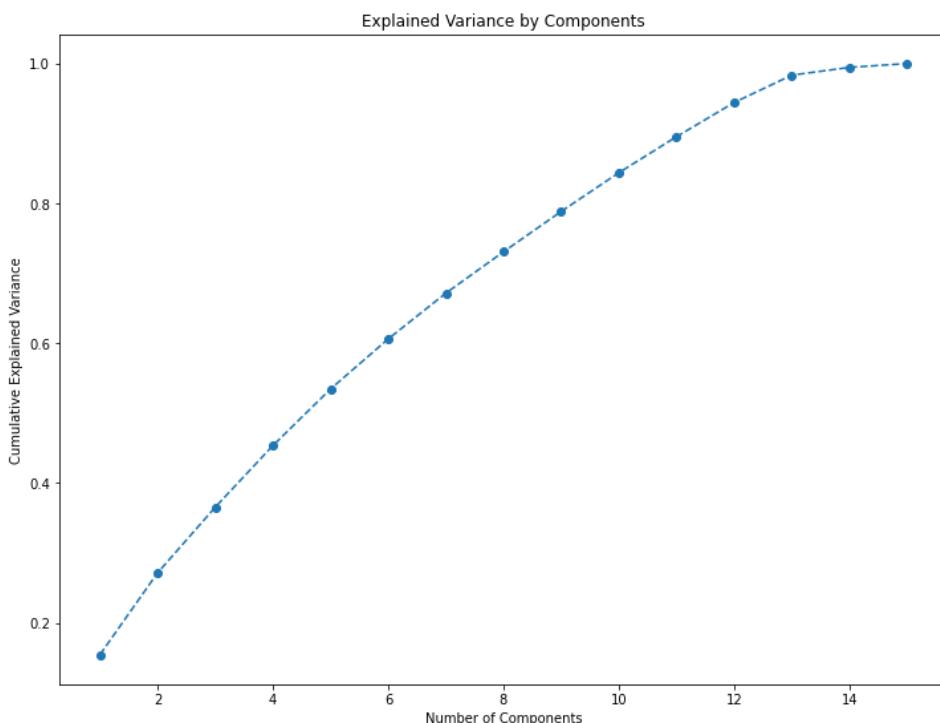
Al estimar la matriz de correlación, se puede observar que no se evidencia correlación lineal fuerte entre las variables seleccionadas por lo que el supuesto de independencia se cumpliría y se puede realizar una interpretación adecuada del modelo al momento de analizar los resultados.

4.1.3 Parametrización

Después de seleccionar las variables pertinentes, es fundamental definir como objetivo el reducir la dimensionalidad de la base de datos manteniendo la mayor cantidad de información posible. En este sentido, se utilizó la técnica de Análisis de Componentes Principales (PCA, por sus siglas en inglés) en conjunto con el proceso de Clusterización.

Para aplicar PCA, fue necesario estandarizar las variables seleccionadas para compararlas adecuadamente. Una vez hecho esto, se calculó la matriz de covarianza, que representa la relación entre las variables (Trefethen & Bau, 2009). Los autovalores y vectores de la matriz de covarianza permitieron determinar la cantidad de varianza explicada por cada componente principal. Se seleccionaron 11 componentes que explican un poco más del 0.85 de la varianza en una escala de 0 a 1. Este proceso permitió reducir la complejidad del conjunto de datos sin perder información relevante para el posterior proceso de clusterización.

Gráfica: Explicación de la varianza por componentes



En la gráfica se observa que después de este valor no hay un aumento notorio y es por esta razón que se seleccionan 11 componentes.

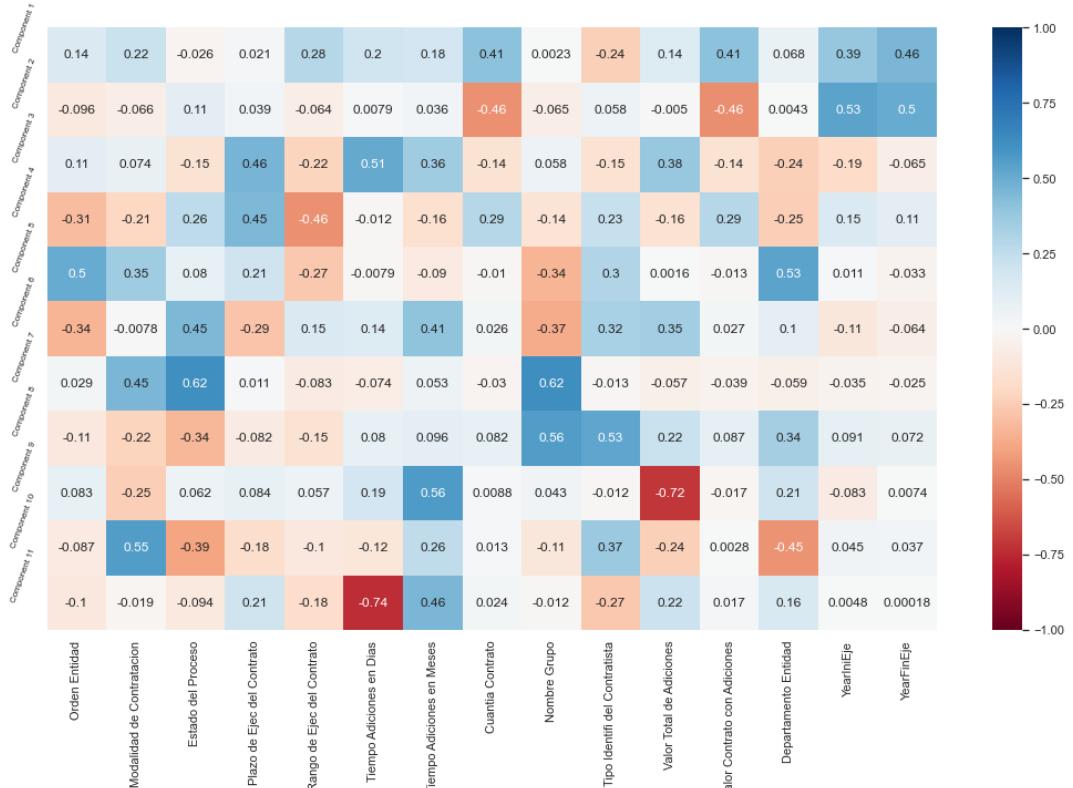
Estimados los componentes, es necesario entender con qué variables se correlacionan más. Por esta razón por medio de un correlograma se puede determinar estas relaciones y así generar un mayor entendimiento de los Clusters que se generan.

Algunas correlaciones de los componentes con las variables son :

- Componente 1 Cuantía Contrato - Valor Contrato con Adiciones - Año Ejecución Y Año Fin Ejecución
- Componente 2 Año de Ini Ejecución y Año Fin de Ejecución
- Componente 3 Plazo ejecución contrato y Tiempo Adiciones en días
- Componente 4 Plazo Ejecución contrato
- Componente 5 Orden Entidad y Departamento Entidad
- Componente 6 Tiempo Adición en Meses
- Componente 7 Estado del Proceso y Nombre Grupo
- Componente 8 Nombre del Grupo Y Tipo Identificación de Contratista
- Componente 9 Valor Total Adiciones
- Componente 10 Modalidad Contratación

- Componente 11 Tiempo en Adiciones en Meses y Tiempo Adiciones en días (-)

Gráfica: Matriz de correlación de los componentes



4.1.4 Definición de métricas adecuadas de desempeño

Generados los componentes y comprendiendo las relaciones que se presentan entre las variables y los componentes estimados, se procede a utilizar K-Means. Para ello, lo primero que se debe identificar es la cantidad de clusters necesarios que logren representar los datos en base a las características de los mismos. Para ello se utilizaron 2 métodos, el método de codo y el método de silhouette, encontrando que la cantidad de grupos sugeridos deben de ser 4.

Gráfica: Método de Codo e Índice de Silhouette



4.2 Análisis de resultados

Conocida la cantidad de clusters se procedió a generar una matriz que contenga todas las variables seleccionadas inicialmente junto con la cantidad de componentes estimados. Al realizar la agrupación por los clusters se hace uso de la media para todas las variables.

Tabla: Segmentación por K-Means y PCA

Orden Entidad	Modalidad de Contratacion	Estado del Proceso	Plazo de Ejec del Contrato	Rango de Ejec del Contrato	Tiempo Adiciones en Dias	Tiempo Adiciones en Meses	Cuantia Contrato	Nombre Grupo	Tipo Identifi del Contratista	... Component 4	Compor
Segment K-means PCA											
PrimeraCategoria	5.499087	3.506999	0.458917	146.626293	0.573341	123.393792	4.600730	3.680758e+07	0.513086	0.107730	...
SegundaCategoria	4.301256	1.768081	0.550357	23.977372	0.398471	2.543299	0.084554	1.495131e+07	0.481589	0.628777	...
TerceraCategoria	4.344961	1.862765	0.497691	23.078089	0.328743	2.085985	0.049250	5.859782e+07	0.571140	0.693413	...
Cuartacategoría	5.086907	4.235654	0.550211	26.221534	0.835624	11.462776	0.512081	9.006600e+08	0.486031	0.231954	...

4 rows × 28 columns

Si bien para un entendimiento de los datos con esta matriz no se pueden generar deducciones fáciles de los resultados por el tamaño del dataset, si se puede dar una idea de las características que tienen cada uno de los clusters generados.

Por ejemplo, el Cluster 3 describe los siguientes tipos de contratos:

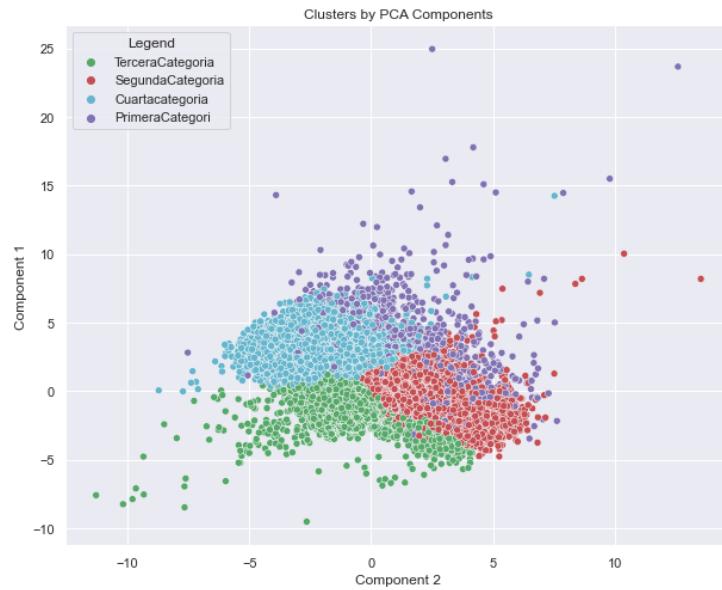
De **Orden entidad** 'NACIONAL DESCENTRALIZADO', 'DISTRITO CAPITAL', 'TERRITORIAL DEPARTAMENTAL DESCENTRALIZADO', 'TERRITORIAL DEPARTAMENTAL CENTRALIZADO', en donde la modalidad de contrato se centraron en Selección Abreviada de Menor Cuantía (Ley 1150 de 2007)', 'Régimen Especial', 'Contratación Directa (Ley 1150 de 2007)' , en donde el Estado del Proceso son 'Liquidado', 'Celebrado', 'Terminado sin Liquidar', en donde en promedio el plazo de Ejecución del Contrato es de 23 días y el Rango de Ejec la mayoría es por días. En estos se presentaron la menor cantidad de adiciones en días (3). El nombre del grupo la mayoría son Componentes y Suministros,Terrenos, Edificios, Estructuras y vías', Materias Primas'. El tipo de identificación la mayoría es 'Cédula de Ciudadanía', 'Nit de Persona Natural. Por parte del departamento Entidad , estuvo un poco centrado en 'Magdalena', 'Cundinamarca', 'Boyacá', y los contratos fueron iniciados y finalizados en el 2016.

Comprendidos estos escenarios y el tipo de variables que tiene la base de datos, se procede a generar gráficos de dispersión entre los componentes estimados.

Componente 1 vs Componente 2

Relación entre los años de ejecución y fin del contrato con respecto a la cuantía del contrato.

Gráfica: Clusters por PCA - Componente 1 vs Componente 2



En el gráfico se observa que la mayoría de los contratos iniciados en un año determinado finalizaron en el mismo año. Por lo tanto, estos contratos concentran la mayor cantidad de registros en el gráfico. Sin embargo, independientemente del grupo al que pertenezcan, los contratos que iniciaron en un año y finalizaron en un año posterior se caracterizan por tener una cuantía mayor. Esta tendencia se refleja en las observaciones dispersas que aparecen en el gráfico.

Componente 7 vs Componente 1

Relación entre el estado en el que terminó el contrato, el nombre del grupo y las adiciones de contrato en valor.

Gráfica: Clusters por PCA - Componente 7 vs Componente 1



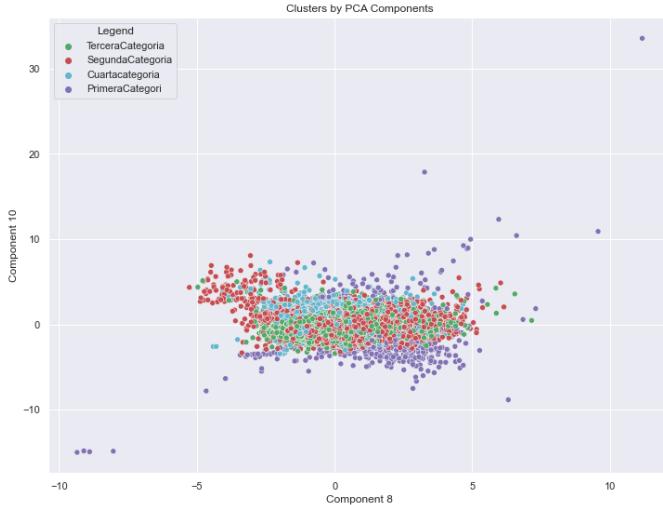
En los Clusters 4 y 1, se puede observar que la mayoría de los contratos presentan una gran dispersión. Esto se debe a que en estos grupos es donde se encuentran la mayor cantidad de

contratos que difieren de lo esperado y que, además, han requerido adiciones de valor en el transcurso del proyecto.

Componente 8 vs Componente 10

Relación entre Nombre Grupo y Modalidad de contratación.

Gráfica: Clusters por PCA - Componente 8 vs Componente 10



Para este caso, se ha observado que únicamente el Cluster 1 presenta una diversidad en cuanto a la modalidad de contratación. Esto se debe a que dentro de este grupo se encuentran diferentes objetos de contratos.

5 Modelo estados finales

Con este modelo se pretende dar respuesta al Artefacto de predicción de estados finales (Desarrollar un artefacto que permita predecir los estados en los que terminará un contrato, el cual se podría utilizar como alerta para tomar decisiones para mejorar la eficiencia en la contratación y generar ahorros.), por lo que se busca, por medio de un modelo de clasificación, predecir el estado final en el que terminará un contrato.

5.1 Procesos de calibración o entrenamiento

Se realizó un proceso de calibración o entrenamiento enfocado en desarrollar modelos de clasificación para predecir si un proceso de contratación culminó en estado "Liquidado" o "Terminado sin Liquidar". Para ello, se factorizaron algunas columnas y se eliminaron las filas con observaciones nulas. A continuación, se graficó el comportamiento de los datos y se definió la variable de interés (Estado del Proceso) como la variable a predecir. Se seleccionaron las variables predictoras y se separaron del conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba utilizando la función `train_test_split`, donde se asignó el 67% para el conjunto de entrenamiento y el 33% restante para el conjunto de prueba. Se inició el modelo con una muestra aleatoria de 30000 registros.

Se realizaron diferentes pruebas con los modelos, tanto con los parámetros por defecto como con diferentes valores de hiperparámetros para determinar cuáles eran los mejores para

obtener el mejor desempeño. Se implementaron diferentes estrategias de desbalance y el desempeño del modelo se midió utilizando las diferentes métricas descritas en el capítulo de [Definición de métricas adecuadas de desempeño](#), con énfasis en el Recall.

5.1.1 Selección de variables

Se realizó una verificación inicial para conocer los estados en los que podían estar los contratos. Los estados posibles eran celebrado, liquidado, convocado, terminado anormalmente después de convocado, adjudicado, descartado, terminado sin liquidar y borrador. Luego de realizar un análisis de los estados en los que podría terminar un contrato, se determinó que los estados candidatos para el proyecto eran: liquidado, terminado anormalmente después de convocado, descartado y terminado sin liquidar, ya que estos estados determinan el fin del contrato y cómo terminó este, lo que es esencial para los modelos de predicción.

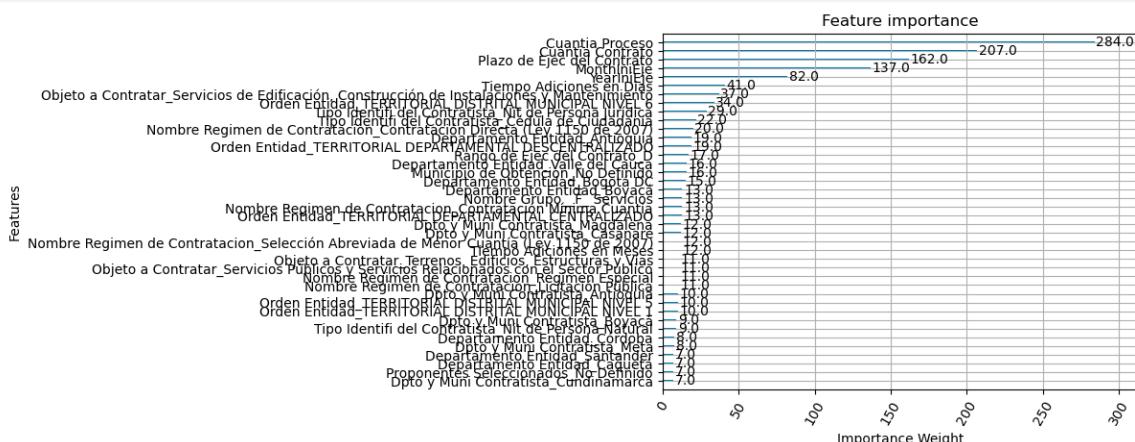
Además de realizar los procesos previos de limpieza y filtrado de datos, se llevó a cabo un análisis de variables con el objetivo de identificar aquellas que fueran más relevantes para el modelo. Se aplicaron técnicas como la generación de clusters y la categorización de las variables cualitativas, y se utilizó la variable "Estado del proceso" como variable de respuesta. Para transformar las variables categóricas en numéricas, se empleó un proceso de encoding y se convirtió la columna "Estado del proceso" a valores enteros (1 y 0), ya que solo quedaron dos tipos de estado del proceso tras la aplicación de los filtros: "Terminado" y "Terminado sin liquidar".

A continuación, se llevó a cabo un submuestreo para equilibrar los dos estados y se obtuvieron 1064 registros para cada uno de ellos. Luego, se realizó un proceso de Feature Engineering para determinar las variables más importantes que debían ser utilizadas para el modelo. Aunque se observó que todas las variables presentaban un nivel de importancia significativo, se mantuvieron todas ellas.

Para verificar la robustez de las variables ante variaciones, se utilizó la técnica Permutation Based Feature Importance y se midió su nivel de importancia utilizando el conjunto de prueba. Esta técnica se puede considerar como una "validación cruzada de k-fold" para la importancia de las características.

Finalmente, se seleccionaron 20 variables significativas que se utilizarán para correr el modelo y predecir el estado de los contratos.

Gráfica: Importancia de las variables con Permutation Based Feature Importance



Además del proceso de Feature Engineering, se llevó a cabo el análisis de SHAP (SHapley Additive exPlanations), el cual proporciona una comprensión interpretativa de las relaciones involucradas en la creación de un modelo y nos muestra el impacto de las características en la salida deseada para los modelos de aprendizaje automático. El objetivo de este análisis es explicar cómo funciona un modelo de aprendizaje automático a través de la comprensión de las variables más significativas, lo que permite la depuración del modelo al seleccionar las variables más importantes y descartar aquellas que no mejoran o aportan al rendimiento del mismo. Esto a su vez, optimiza la eficiencia del modelo y mejora las predicciones del mismo.

Es importante destacar que el análisis de SHAP complementa el proceso de Feature Engineering al corroborar las variables sugeridas en dicho proceso. Al seleccionar estas variables, se espera obtener un modelo óptimo al momento de realizar las predicciones.

Gráfica: Importancia de las variables con SHAP



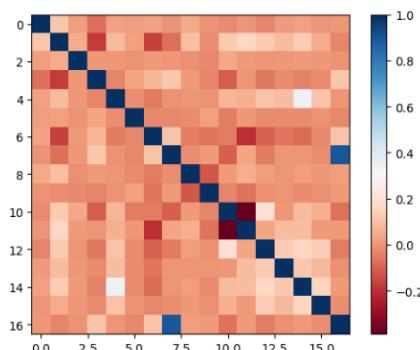
5.1.2 Pruebas de verificación de los supuestos requeridos

Para implementar los diferentes modelos, primero se utilizaron los datos resultantes de los pasos de limpieza, filtrado de variables e imputación de valores nulos. Posteriormente, se llevó a cabo un proceso de verificación de los supuestos necesarios para determinar si se podría hacer una interpretación adecuada de los modelos.

Supuesto para modelos de clasificación lineal

- Análisis de independencia de observaciones:

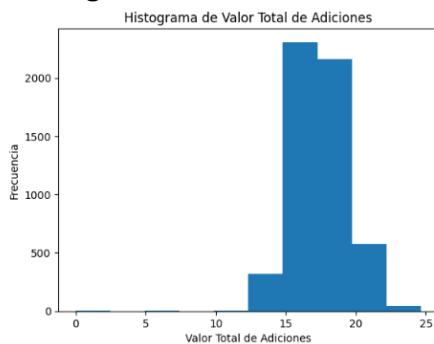
Gráfica: Matriz de correlación de las observaciones



Al observar los valores de la matriz de correlación, se puede notar que la mayoría de estos están cercanos a cero. Este hecho sugiere que no hay una relación lineal fuerte entre las variables en el conjunto de datos. En otras palabras, si el valor de una variable aumenta, no necesariamente implica que el valor de la otra variable aumente o disminuya en la misma proporción. Es posible que las variables sean independientes entre sí. Por lo tanto, la correlación cercana a cero indica que las variables no están relacionadas de manera significativa o que la relación entre ellas no es lineal.

- Supuesto de normalidad de datos numéricos
 - Valor total de adicciones

Gráfica: Histograma del Valor total de adicciones



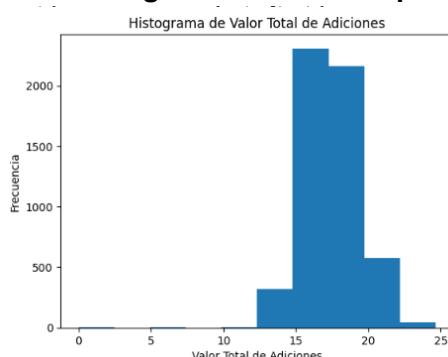
- Valor Contrato con adicciones

Gráfica: Histograma del Valor contrato con adicciones



- Cuantía proceso

Gráfica: Histograma de Cuantía proceso



Al observar el comportamiento de los histogramas de las variables numéricas se puede apreciar que tienen una distribución parecida a la normal, por lo que se puede inferir que cumplen con el supuesto de normalidad.

- Linealidad:

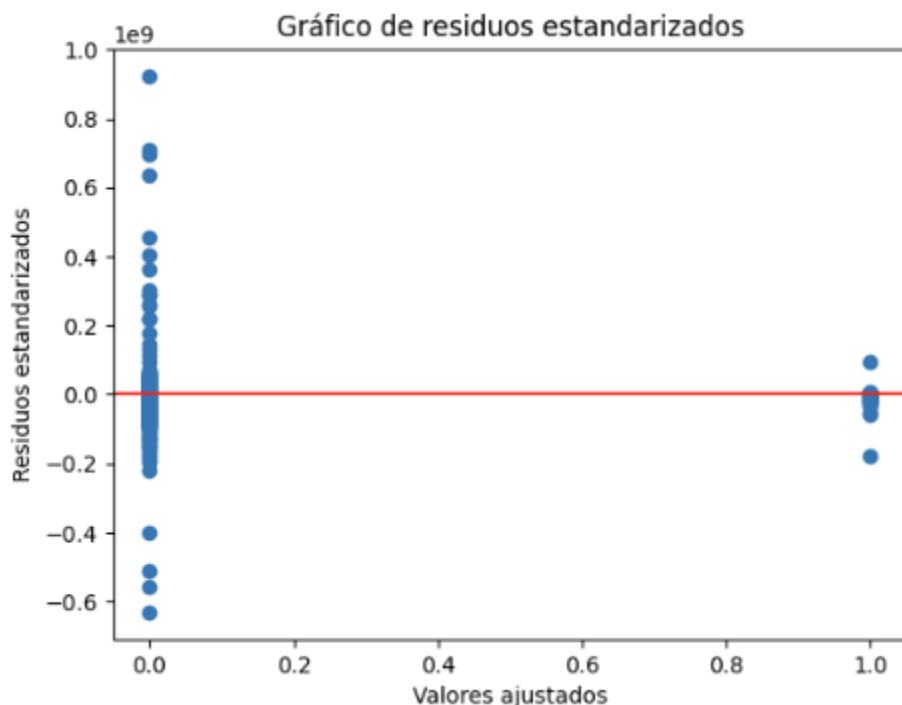
Se realiza una prueba de linealidad Rainbow, obteniendo los siguientes resultados:

- Rainbow statistic: 1.2964988568819171
- Rainbow p-value: 4.825511641860705e-57

Se cumple el supuesto. El rainbow statistic es cercano a 1, sugiriendo que se ajusta al supuesto de linealidad y el p-value es menor que 0.05, lo que significa que no se puede rechazar la hipótesis nula.

- Homocedasticidad

Gráfica: Residuos estandarizados



Breusch-Pagan test:

Estadístico de prueba: 17012.095960373987

Valor p: 0.0

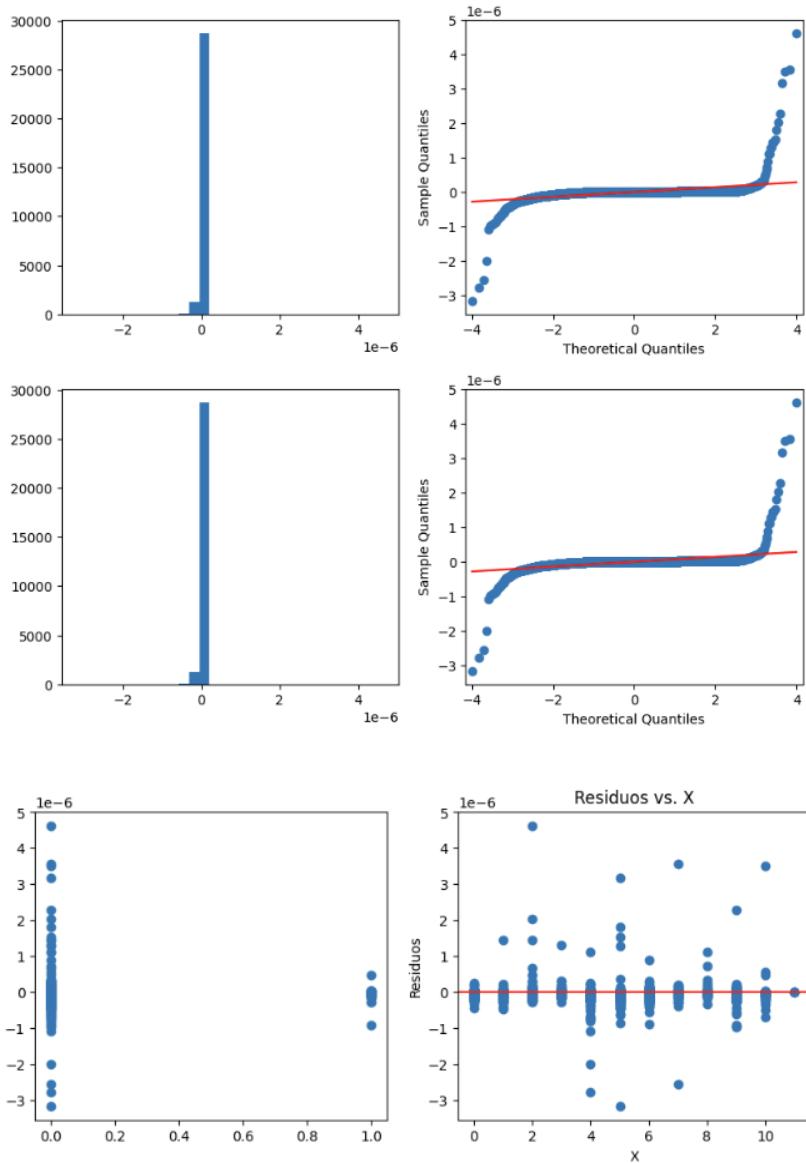
Estadístico de prueba alternativo: 2310.0981431181845

Valor p alternativo: 0.0

Hay evidencia estadística para rechazar la hipótesis nula de homocedasticidad, y que la varianza del error no es constante. En otras palabras, existe heterocedasticidad en el modelo de regresión por lo que no se cumple con el supuesto.

- Análisis de residuos

Gráfica: Análisis de residuos



Al realizar el análisis de residuos, se puede observar la distribución de los residuos alrededor de cero, lo que sugiere una aleatoriedad en los datos. Esto significa que los residuos no están siguiendo ningún patrón sistemático y no hay ninguna relación evidente entre ellos y las variables independientes o dependientes en el modelo, por lo que se cumple el supuesto.

- Supuesto de independencia de los errores

Estadística de Durbin-Watson: 1.995006884950339
No hay autocorrelación

El resultado de la estadística sugiere que el supuesto de independencia de errores se cumple, ya que los errores no están correlacionados entre sí y, por lo tanto, son independientes.

Resultados:

Después de realizar las validaciones de los supuestos necesarios para el análisis de datos, se llevaron a cabo diversas técnicas para verificar si se cumplían o no. En concreto, se validó la normalidad mediante la visualización de un gráfico de histograma, la homocedasticidad mediante un gráfico de residuos estandarizados, y la independencia entre las observaciones mediante un gráfico de autocorrelación. Se verificó el tamaño de muestra adecuado y la existencia de clases equilibradas mediante el balanceo de las muestras. Para la linealidad se utilizó la prueba Rainbow statistic, mientras que la ausencia de multicolinealidad se verificó con un gráfico de autocorrelación. Se llevó a cabo un diagnóstico para comprobar la presencia o ausencia de valores atípicos, y se convirtieron las variables categóricas predictoras en 1 y 0 para la variable dependiente binaria. La imputación de datos faltantes se realizó con ceros, mientras que la factorización y estandarización de las entradas numéricas se hizo para mejorar la calidad de los datos.

En la siguiente tabla se muestra una síntesis del cumplimiento de los supuestos en el análisis de datos:

Gráfica: Validación de supuestos del modelo de estados finales

Supuestos	LDA	QDA	Naive Bayes	Regresión Logística	Decision Tree Classifier	Random Forest	AdaBoost	Gradient Boosting	XGBoost	K Vecinos
Normalidad	cumple	cumple	cumple	cumple						
Homocedasticidad	no cumple	no cumple								
Independencia	cumple	cumple		cumple			cumple	cumple	cumple	cumple
Matrices de covarianza iguales	cumple									
Tamaño de muestra adecuado		cumple	cumple	cumple						
Independencia condicional de clase			cumple							
Clases equilibradas			cumple					cumple		
Linealidad				cumple					cumple	
Ausencia de multicolinealidad				cumple	cumple					
Sin valores atípicos				cumple		cumple		cumple	cumple	
Variable dependiente binaria				cumple	cumple					
Sin datos faltantes						cumple			cumple	
Entradas numéricas								cumple	cumple	
Estandarizados										cumple

Después de realizar las validaciones correspondientes, se lograron verificar la mayoría de los supuestos necesarios para el análisis de datos. En particular, se pudo validar la normalidad, la independencia entre las observaciones, el tamaño de muestra adecuado, las clases equilibradas, la linealidad, la ausencia de multicolinealidad, la ausencia de valores atípicos, la variable dependiente binaria y la ausencia de datos faltantes. Sin embargo, se encontró que no se cumplió el supuesto de homocedasticidad, según los resultados obtenidos del test de Breusch-Pagan, lo cual solo afecta los modelos de LDA y QDA.

5.1.3 Parametrización

La revisión inicial de los modelos se realizó con los parámetros por defecto, aplicando las siguientes técnicas de balanceo según el modelo:

1. Ninguna: no se implementa ninguna técnica de balanceo de clases.
2. Penalización para compensar
3. Subsampling en la clase mayoritaria: se implementa la técnica de submuestreo 'NearMiss' en el conjunto de datos de entrenamiento seleccionando las muestras de la clase mayoritaria más cercanas a las de la clase minoritaria.
4. Oversampling de la clase minoritaria: se implementa la técnica de sobremuestreo 'RandomOverSampler' en el conjunto de datos de entrenamiento generando muestras sintéticas de la clase minoritaria para equilibrar la distribución de clases.
5. Combinamos resampling con Smote-Tomek: se implementa la técnica de sobremuestreo y submuestreo combinados SMOTETomek generando muestras sintéticas de la clase minoritaria y eliminando las muestras de la clase mayoritaria que están cerca de las muestras de la clase minoritaria.
6. Ensamble de Modelos con Balanceo: se implementa el uso de 'BalancedBaggingClassifier()' para abordar el problema de desequilibrio de clases en el conjunto de datos de entrenamiento mediante el muestreo de la clase minoritaria para equilibrar la distribución de clases.

Los resultados se pueden ver en el siguiente numeral de [Análisis de resultados](#).

A partir de estos resultados se identificó el modelo con mejor desempeño general según las métricas planteadas, que en este caso fue el modelo XGBoost.

La calibración del modelo XGBoost es un proceso de ajustar los hiperparámetros del modelo para mejorar su desempeño y reducir el sobreajuste. Se decide utilizar una calibración por pasos de los diferentes hiperparámetros hasta obtener la mejor combinación posible de acuerdo a la métrica definida.

Se toma como base el modelo XGBClassifier().

El proceso de calibración se realizó de la siguiente manera:

Paso 1 - Preparación de los Datos

Tomando como base el resultado de la selección de variables se continúa el trabajo de preparación de los datos para aplicarlos al modelo.

Las variables categóricas son convertidas en numéricas utilizando 'OneHotEncoder()'.

Las variables numéricas se estandarizan utilizando 'StandardScaler'.

Se hace una separación de variables predictoras (X) y variable de interés (y) y se divide la muestra en un set de entrenamiento y un set de pruebas test usando la función 'train_test_split'.

Paso 2 - Definición de los parámetros a calibrar

Se definen en un valor inicial los hiperparámetros a calibrar dentro del modelo XGBoost con los valores preestablecidos por defecto por el modelo:

- objective = 'binary:logistic'

- eval_metric = "logloss"
- learning_rate=None
- max_depth = 6
- n_estimators=100
- colsample_bytree = 1
- gamma = 0
- min_child_weight = 1
- reg_alpha=None
- reg_lambda=None

Paso 3 - Calibración Iterativa de los hiperparámetros

La estrategia es implementar una búsqueda en cuadrícula con el método 'GridSearchCV()' para ajustar los hiperparámetros utilizando la métrica de puntuación 'recall' para evaluar el rendimiento del modelo, y una validación cruzada de 5 veces (cv=5) para evitar el sobreajuste.

La búsqueda se desarrolla de la siguiente manera.

7. Selecciona uno o dos hiperparámetros para calibrar y los demás permanecen constantes.
8. Prueba valores dentro de un rango predefinido para cada hiperparámetro.
9. Evalúa cada combinación de hiperparámetros contra la métrica seleccionada.
10. Guarda temporalmente la combinación de mejor desempeño.
11. Prueba nuevamente con una grilla más detallada alrededor de los parámetros seleccionados como temporales.
12. Guarda el resultado de esta última prueba como el valor definitivo para los hiperparámetros.

Tabla: Calibración hiperparámetros

Hiperparámetro	Rango de prueba inicial	Óptimo Temporal	Rango de prueba detallado	Óptimo final
'max_depth'	range(3,10,2)	'max_depth': 5	[max_depth-1, max_depth, max_depth+1]	'max_depth': 5
'min_child_weight'	[1,3,5]	'min_child_weight': 3	[min_child_weight-1, min_child_weight-0.5, min_child_weight, min_child_weight+0.5, min_child_weight+1]	'min_child_weight': 2.5
'gamma'	[i/10.0 for i in range(1,10,2)]	'gamma': 0.7	['gamma'-1, 'gamma'-0.5, 'gamma', 'gamma'+0.5, 'gamma'+1]	'gamma': 0.7
'subsample'	'subsample': [i/10.0 for i in range(6,11)]	'subsample': 0.9	[i/100.0 for i in range(int((subsample-0.1)*100.0), min(int((subsample+0.1)*100),105), 5)]	'subsample': 0.9
'colsample_bytree':	'colsample_bytree': [i/10.0 for i in range(6,11)]	'colsample_bytree': 0.6	[i/100.0 for i in range(int((colsample_bytree-0.1)*100.0), min(int((subsample+0.1)*100),105), 5)]	'colsample_bytree': 0.75
'reg_alpha':	'reg_alpha': [1e-5, 1e-2, 0.1, 1, 100]	'reg_alpha': 0.1	['reg_alpha'*0.2, 'reg_alpha'*0.5, 'reg_alpha', 'reg_alpha'*2, 'reg_alpha'*5]	'reg_alpha': 0.01

Hiperparametro	Rango de prueba inicial	Óptimo Temporal	Rango de prueba detallado	Óptimo final
'reg_lambda'	'reg_lambda': [1e-5, 1e-2, 0.1, 1, 100]	'reg_lambda': 0.01	[reg_lambda*0.2, reg_lambda*0.5, reg_lambda, reg_lambda*2, reg_lambda*5]	'reg_lambda': 0.05
'learning_rate':	'learning_rate': [0.0001, 0.001, 0.01, 0.1, 0.2, 0.3]	'learning_rate': 0.3	[learning_rate*0.2, learning_rate*0.5, learning_rate, learning_rate*2, learning_rate*5]	'learning_rate': 0.3
'n_estimators':	'n_estimators': [50, 100, 200]	'n_estimators': 100	[(n_estimators*3/4), n_estimators, n_estimators*5/4]	'n_estimators': 100

Paso 4 - Aplicación de estrategia de balanceo

Se aplican las 5 técnicas de balanceo de clases que permite XGBoost para probar mejoras en los resultados del modelo y los resultados comparativos se pueden ver en el siguiente numeral de [Análisis de resultados](#).

5.2 Análisis de resultados

Para analizar los resultados, es importante tener en cuenta que se realizaron dos transformaciones de las variables durante el procesamiento de los datos: la conversión de las variables en dummies y la transformación por factorización. Estas transformaciones dieron lugar a diferentes resultados en los modelos entrenados.

Factorize:

A continuación se presentan los resultados de los modelos implementados con el uso de diferentes estrategias de calibración de datos.

Tabla: Resultados de las métricas para parámetros por defecto

Modelo	Estrategia	0			1			accuracy	prom
		precision	recall	f1	precision	recall	f1		
Decision Tree Classifier	Ninguna	0,99	0,99	0,99	0,14	0,19	0,16	0,98	0,58
	Penalización para compensar	0,99	0,99	0,99	0,12	0,14	0,13	0,98	0,56
	Subsampling en la clase mayoritaria	0,99	0,02	0,04	0,01	0,98	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	0,99	0,99	0,13	0,16	0,14	0,98	0,57
	Combinamos resampling con Smote-Tomek	0,99	0,96	0,97	0,05	0,23	0,09	0,95	0,55
	Ensamble de Modelos con Balanceo	1,00	0,89	0,94	0,05	0,59	0,10	0,88	0,60
Random Forest	Ninguna	0,99	1,00	0,99	0,78	0,07	0,12	0,99	0,66
	Penalización para compensar	0,99	1,00	0,99	0,88	0,07	0,12	0,99	0,68
	Subsampling en la clase mayoritaria	0,99	0,02	0,04	0,01	0,97	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	1,00	0,99	0,30	0,12	0,17	0,99	0,60
	Combinamos resampling con Smote-Tomek	0,99	0,99	0,99	0,15	0,16	0,16	0,98	0,57
	Ensamble de Modelos con Balanceo	0,99	0,86	0,92	0,04	0,54	0,08	0,86	0,57
AdaBoost	Ninguna	0,99	1,00	0,99	0,00	0,00	0,00	0,99	0,50

Modelo	Estrategia	0			1			accuracy	prom
		precision	recall	f1	precision	recall	f1		
	Subsampling en la clase mayoritaria	0,99	0,02	0,04	0,01	0,98	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	0,79	0,88	0,03	0,57	0,05	0,79	0,55
	Combinamos resampling con Smote-Tomek	0,99	0,80	0,89	0,02	0,32	0,03	0,80	0,51
	Ensamble de Modelos con Balanceo	0,99	0,75	0,86	0,03	0,64	0,05	0,75	0,55
Gradient Boosting	Ninguna	0,99	1,00	0,99	0,22	0,04	0,07	0,99	0,55
	Subsampling en la clase mayoritaria	0,99	0,02	0,03	0,01	0,98	0,02	0,03	0,34
	Oversampling de la clase minoritaria	1,00	0,89	0,94	0,05	0,61	0,10	0,88	0,60
	Combinamos resampling con Smote-Tomek	0,99	0,87	0,93	0,03	0,34	0,05	0,87	0,54
	Ensamble de Modelos con Balanceo	0,99	0,83	0,91	0,04	0,60	0,07	0,83	0,57
XGBoost	Ninguna	0,99	1,00	1,00	0,72	0,12	0,21	0,99	0,67
	Subsampling en la clase mayoritaria	0,99	0,02	0,05	0,01	0,97	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	0,99	0,99	0,31	0,30	0,30	0,99	0,65
	Combinamos resampling con Smote-Tomek	0,99	0,98	0,99	0,09	0,16	0,12	0,97	0,56
	Ensamble de Modelos con Balanceo	0,99	0,84	0,91	0,04	0,60	0,07	0,84	0,58

Cada modelo ha sido entrenado con una de las siguientes estrategias: ninguna, penalización para compensar, subsampling en la clase mayoritaria, oversampling de la clase minoritaria, combinación de resampling con Smote-Tomek y ensamblaje de modelos con balanceo.

Para cada estrategia, se proporcionan medidas de rendimiento, como la precisión, el recall y la puntuación f1 para ambas clases, así como la precisión general.

Para el modelo Decision Tree Classifier, la técnica de ensamble de modelos con balanceo mostró los mejores resultados en términos de precisión, recall y F1-score para ambas clases, mientras que la técnica de subsampling en la clase mayoritaria mostró el peor rendimiento en términos de precisión, recall y F1-score para ambas clases.

Para el modelo Random Forest, la técnica de oversampling de la clase minoritaria mostró los mejores resultados en términos de precisión, recall y F1-score para la clase minoritaria, mientras que la técnica de ensamble de modelos con balanceo mostró el mejor rendimiento en términos de precisión y F1-score para la clase mayoritaria.

En el caso de AdaBoost, la técnica de oversampling de la clase minoritaria mostró los mejores resultados en términos de precisión, recall y F1-score para la clase minoritaria, mientras que la técnica de ensamble de modelos con balanceo mostró el mejor rendimiento en términos de precisión y F1-score para la clase mayoritaria.

Para el modelo Gradient Boosting, la técnica de oversampling de la clase minoritaria mostró los mejores resultados en términos de precisión, recall y F1-score para la clase minoritaria, mientras que la técnica de ensamble de modelos con balanceo mostró el mejor rendimiento en términos de precisión y F1-score para la clase mayoritaria.

Dummies:

Los resultados obtenidos a través de estas variables mostraron un rendimiento inferior al de la factorización, por lo tanto, se presentarán en los anexos (ver [Análisis de resultados](#)) para centrarnos en los resultados de la factorización.

El paso final, fue implementar una búsqueda en cuadrícula con el método 'GridSearchCV()' para ajustar los hiperparámetros utilizando la métrica de puntuación 'recall' para evaluar el rendimiento del modelo, y una validación cruzada de 5 veces (cv=5) para evitar el sobreajuste.

Este proceso se aplica para las 5 estrategias de tratamiento de muestra desbalanceadas y se compara su desempeño:

Tabla: Comparativos de resultados por defecto vs calibrados

Comparación de resultados									
Clase		0			1			modelo	
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
Parámetros por Defecto									
XGBoost	Ninguna	0,99	1,00	1,00	0,72	0,12	0,21	0,99	0,67
	Subsampling en la clase mayoritaria	0,99	0,02	0,05	0,01	0,97	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	0,99	0,99	0,31	0,30	0,30	0,99	0,65
	Combinamos resampling con Smote-Tomek	0,99	0,98	0,99	0,09	0,16	0,12	0,97	0,56
	Ensamble de Modelos con Balanceo	0,99	0,84	0,91	0,04	0,60	0,07	0,84	0,58
Parámetros Calibrados									
XGBoost	DEFAULT	0,99	1,00	1,00	0,91	0,10	0,18	0,99	0,74
	Subsampling en la clase mayoritaria	0,99	0,05	0,09	0,01	0,94	0,02	0,50	0,37
	Oversampling de la clase minoritaria	0,99	0,95	0,97	0,08	0,43	0,14	0,95	0,64
	Combinamos resampling con Smote-Tomek	0,99	1,00	1,00	0,57	0,16	0,25	0,99	0,71
	Ensamble de Modelos con Balanceo	1,00	0,79	0,88	0,03	0,62	0,05	0,79	0,59

La calibración ha mejorado el desempeño de los modelos al compararlo con los hiperparámetros por defecto y se evidencia que la estrategia de oversampling de la clase minoritaria y el ensamble de Modelos con balanceo son las estrategias que tienen mejor desempeño al comparar los valores de la métrica de recall siendo estos los candidatos para utilizar en la predicción de la variable 'Estado del Proceso'.

6 Modelo adiciones

Con este modelo se pretende dar respuesta al Artefacto de predicción de adiciones (Desarrollar un artefacto que permita predecir el riesgo de adiciones presupuestales durante el proceso de contratación, el cual se podría utilizar para tomar decisiones para mejorar la eficiencia en la contratación y generar ahorros.), por lo que se busca, por medio de un modelo de clasificación, predecir si un contrato tendrá adiciones presupuestales.

6.1 Procesos de calibración o entrenamiento

Se realizó un proceso de calibración o entrenamiento enfocado en desarrollar modelos de clasificación para predecir si un proceso de contratación tendrá adiciones presupuestales. Para ello, se factorizaron algunas columnas y se eliminaron las filas con observaciones nulas. A continuación, se graficó el comportamiento de los datos y se creó como la variable a predecir la variable de Adición, la cual se define como igual a 1 si el valor total de adiciones es distinto de cero, y 0 en caso contrario. Se seleccionaron las variables predictoras y se separaron del conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba utilizando la función `train_test_split`, donde se asignó el 67% para el conjunto de entrenamiento y el 33% restante para el conjunto de prueba. Se inició el modelo con una muestra aleatoria de 30000 registros.

Se realizaron diferentes pruebas con los modelos, tanto con los parámetros por defecto como con diferentes valores de hiperparámetros para determinar cuáles eran los mejores para obtener el mejor desempeño. Se implementaron diferentes estrategias de desbalance y el desempeño del modelo se midió utilizando las diferentes métricas descritas en el capítulo de [Definición de métricas adecuadas de desempeño](#), con énfasis en el Recall.

6.1.1 Selección de variables

A partir de las variables seleccionadas en el numeral de [Preselección de variables explicativas](#), se realizó una verificación inicial de las variables que pueden tener correlación con la variable creada de Adición, por lo que se eliminaron las variables de: Prorroga, Tiempo Adiciones en Dias, Tiempo Adiciones en Meses, Adicion, Valor Total de Adiciones y Valor Contrato con Adiciones.

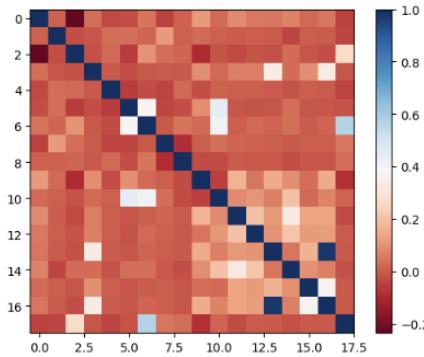
Las variables utilizadas para el modelo son: Orden Entidad, Modalidad de Contratacion, Estado del Proceso, Municipio de Obtencion, Cuantia Proceso, Nombre Grupo, Tipo Identifi del Contratista, Dpto y Muni Contratista, Anno Ini Ejec Contrato, Mes Ini Ejec Contrato, Plazo de Ejec del Contrato, Rango de Ejec del Contrato, Cuantia Contrato, Departamento Entidad.

6.1.2 Pruebas de verificación de los supuestos requeridos

Supuesto para modelos de clasificación lineal

- Análisis de independencia de observaciones:

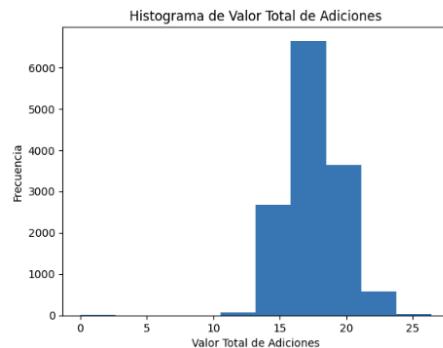
Gráfica: Matriz de correlación de las observaciones



Al observar los valores de la matriz de correlación, se puede notar que la mayoría de estos están cercanos a cero. Este hecho sugiere que no hay una relación lineal fuerte entre las variables en el conjunto de datos. En otras palabras, si el valor de una variable aumenta, no necesariamente implica que el valor de la otra variable aumente o disminuya en la misma proporción. Es posible que las variables sean independientes entre sí. Por lo tanto, la correlación cercana a cero indica que las variables no están relacionadas de manera significativa o que la relación entre ellas no es lineal. Cabe destacar que esta correlación no lineal puede no ser capturada por la matriz de correlación lineal.

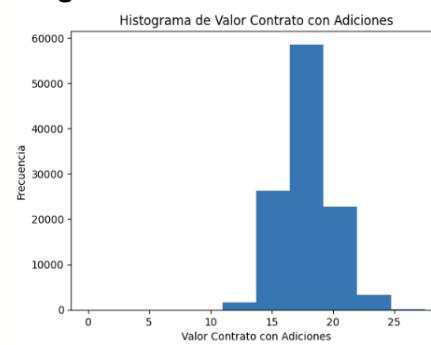
- Supuesto de normalidad de datos numéricos
 - Valor total de adicciones

Gráfica: Histograma del Valor total de adicciones



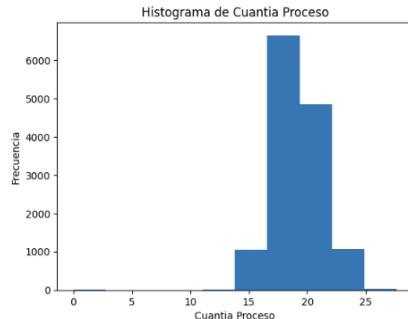
- Valor contrato con adicciones

Gráfica: Histograma del Valor contrato con adicciones



- Cuantia proceso

Gráfica: Histograma de Cuantía proceso



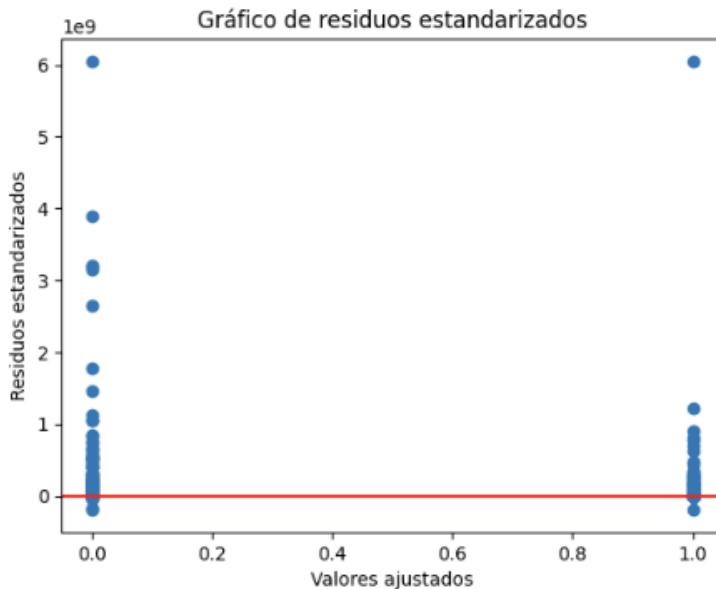
Al observar el comportamiento de los histogramas de las variables numéricas se puede apreciar que tienen una distribución parecida a la normal, por lo que se puede inferir que cumplen con el supuesto de normalidad.

- Linealidad:
 - Rainbow statistic: 1.278593435662614
 - Rainbow p-value: 2.4918266722368118e-51

Se cumple el supuesto. El rainbow statistic es cercano a 1, sugiriendo que se ajusta al supuesto de linealidad y el p-value es menor que 0.05, lo que significa que no se puede rechazar la hipótesis nula.

- Homocedasticidad

Gráfica: Residuos estandarizados



Breusch-Pagan test:
 Estadístico de prueba: 25877.033295927067
 Valor p: 0.0
 Estadístico de prueba alternativo: 11069.202576473921

Hay evidencia estadística para rechazar la hipótesis nula de homocedasticidad, y que la varianza del error no es constante. En otras palabras, existe heterocedasticidad en el modelo de regresión por lo que no se cumple con el supuesto.

- Supuesto de independencia de los errores

Estadística de Durbin-Watson: 1.994668836535482
No hay autocorrelación

El resultado de la estadística sugiere que el supuesto de independencia de errores se cumple, ya que los errores no están correlacionados entre sí y, por lo tanto, son independientes.

Resultados:

Después de realizar las validaciones de los supuestos necesarios para el análisis de datos, se llevaron a cabo diversas técnicas para verificar si se cumplían o no. En concreto, se validó la normalidad mediante la visualización de un gráfico de histograma, la homocedasticidad mediante un gráfico de residuos estandarizados, y la independencia entre las observaciones mediante un gráfico de autocorrelación. Se verificó el tamaño de muestra adecuado y la existencia de clases equilibradas mediante el balanceo de las muestras. Para la linealidad se utilizó la prueba Rainbow statistic, mientras que la ausencia de multicolinealidad se verificó con un gráfico de autocorrelación. Se llevó a cabo un diagnóstico para comprobar la presencia o ausencia de valores atípicos, y se convirtieron las variables categóricas predictoras en 1 y 0 para la variable dependiente binaria. La imputación de datos faltantes se realizó con ceros, mientras que la factorización y estandarización de las entradas numéricas se hizo para mejorar la calidad de los datos.

En la siguiente tabla se muestra una síntesis de los supuestos que se cumplieron y los que no en el análisis de datos:

Gráfica: Validación de supuestos del modelo de adiciones

Supuestos	LDA	QDA	Naive Bayes	Regresión Logística	Decision Tree Classifier	Random Forest	AdaBoost	Gradient Boosting	XGBoost	K Vecinos
Normalidad	cumple	cumple	cumple	cumple						
Homocedasticidad	no cumple	no cumple								
Independencia	cumple	cumple		cumple			cumple	cumple	cumple	cumple
Matrices de covarianza iguales	cumple									
Tamaño de muestra adecuado		cumple	cumple	cumple						
Independencia condicional de clase			cumple							
Clases equilibradas			cumple				cumple			
Linealidad				cumple					cumple	
Ausencia de multicolinealidad				cumple	cumple					
Sin valores atípicos				cumple		cumple		cumple	cumple	
Variable dependiente binaria				cumple	cumple					

Supuestos	LDA	QDA	Naive Bayes	Regresión Logística	Decision Tree Classifier	Random Forest	AdaBoost	Gradient Boosting	XGBoost	K Vecinos
Sin datos faltantes						cumple			cumple	
Entradas numéricas								cumple	cumple	
Estandarizados										cumple

Después de realizar las validaciones correspondientes, se lograron verificar la mayoría de los supuestos necesarios para el análisis de datos. En particular, se pudo validar la normalidad, la independencia entre las observaciones, el tamaño de muestra adecuado, las clases equilibradas, la linealidad, la ausencia de multicolinealidad, la ausencia de valores atípicos, la variable dependiente binaria y la ausencia de datos faltantes. Sin embargo, se encontró que no se cumplió el supuesto de homocedasticidad, según los resultados obtenidos del test de Breusch-Pagan, lo cual solo afecta los modelos de LDA y QDA.

6.1.3 Parametrización

La revisión inicial de los modelos se realizó con los parámetros por defecto, aplicando las siguientes técnicas de balanceo según el modelo:

1. Ninguna: no se implementa ninguna técnica de balanceo de clases.
2. Penalización para compensar
3. Subsampling en la clase mayoritaria: se implementa la técnica de submuestreo 'NearMiss' en el conjunto de datos de entrenamiento seleccionando las muestras de la clase mayoritaria más cercanas a las de la clase minoritaria.
4. Oversampling de la clase minoritaria: se implementa la técnica de sobremuestreo 'RandomOverSampler' en el conjunto de datos de entrenamiento generando muestras sintéticas de la clase minoritaria para equilibrar la distribución de clases.
5. Combinamos resampling con Smote-Tomek: se implementa la técnica de sobremuestreo y submuestreo combinados SMOTETomek generando muestras sintéticas de la clase minoritaria y eliminando las muestras de la clase mayoritaria que están cerca de las muestras de la clase minoritaria.
6. Ensamble de Modelos con Balanceo: se implementa el uso de 'BalancedBaggingClassifier()' para abordar el problema de desequilibrio de clases en el conjunto de datos de entrenamiento mediante el muestreo de la clase minoritaria para equilibrar la distribución de clases.

Los resultados se pueden ver en el siguiente numeral de [Análisis de resultados](#).

A partir de estos resultados se identificó el modelo con mejor desempeño general según las métricas planteadas, que en este caso fue el modelo XGBoost.

La calibración del modelo XGBoost es un proceso de ajustar a los hiperparámetros del modelo para mejorar su rendimiento y reducir el sobreajuste. Se decide utilizar una calibración por pasos de los diferentes hiperparámetros hasta obtener la mejor combinación posible de acuerdo a la métrica definida.

Se toma como base el modelo XGBClassifier().

El proceso de calibración se realizó de la siguiente manera:

Paso 1 - Preparación de los Datos

Tomando como base el resultado de la selección de variables se continúa el trabajo de preparación de los datos para aplicarlos al modelo.

Las variables categóricas son convertidas en numéricas utilizando 'OneHotEncoder()'.

Las variables numéricas se estandarizan utilizando 'StandardScaler'.

Se hace una separación de variables predictoras (X) y variable de interés (y) y se divide la muestra en un set de entrenamiento y un set de pruebas test usando la función 'train_test_split'.

Paso 2 - Definición de los parámetros a calibrar

Se definen en un valor inicial los hiperparámetros a calibrar dentro del modelo XGBoost con los valores preestablecidos por defecto por el modelo:

- objective = 'binary:logistic'
- eval_metric = "logloss"
- learning_rate=None
- max_depth = 6
- n_estimators=100
- colsample_bytree = 1
- gamma = 0
- min_child_weight = 1
- reg_alpha=None
- reg_lambda=None

Paso 3 - Calibración Iterativa de los hiperparámetros

La estrategia es implementar una búsqueda en cuadrícula con el método 'GridSearchCV()' para ajustar los hiperparámetros utilizando la métrica de puntuación 'recall' para evaluar el rendimiento del modelo, y una validación cruzada de 5 veces (cv=5) para evitar el sobreajuste.

La búsqueda se desarrolla de la siguiente manera.

1. Selecciona uno o dos hiperparámetros para calibrar los demás permanecen constantes.
2. Prueba valores dentro de un rango predefinido para cada hiperparámetro.
3. Evalúa cada combinación de hiperparámetros contra la métrica seleccionada.
4. Guarda temporalmente la combinación de mejor desempeño.
5. Prueba nuevamente con una grilla más detallada alrededor de los parámetros seleccionados como temporales.
6. Guarda el resultado de esta última prueba como el valor definitivo para los hiperparámetros.

Tabla: Calibración hiperparámetros

Hiperparametro	Rango de prueba inicial	Óptimo Temporal	Rango de prueba detallado	Óptimo final
'max_depth'	range(3,10,2)	max_depth': 5	[max_depth-1, max_depth+1]	max_depth': 5

Hiperparametro	Rango de prueba inicial	Óptimo Temporal	Rango de prueba detallado	Óptimo final
'min_child_weight'	[1,3,5]	'min_child_weight': 3	[min_child_weight-1, min_child_weight-0.5, min_child_weight, min_child_weight+0.5, min_child_weight+1]	'min_child_weight': 2.5
'gamma'	[i/10.0 for i in range(1,10,2)]	'gamma': 0.7	['gamma'-1, 'gamma'-0.5, 'gamma', 'gamma'+0.5, 'gamma'+1]	'gamma': 0.7
'subsample'	'subsample': [i/10.0 for i in range(6,11)]	'subsample': 0.9	[i/100.0 for i in range(int((subsample-0.1)*100.0), min(int((subsample+0.1)*100),105) , 5)]	'subsample': 0.9
'colsample_bytree':	'colsample_bytree': [i/10.0 for i in range(6,11)]	'colsample_bytree': 0.6	[i/100.0 for i in range(int((colsample_bytree-0.1)*100.0), min(int((subsample+0.1)*100),105), 5)]	'colsample_bytree': 0.75
'reg_alpha':	'reg_alpha': [1e-5, 1e-2, 0.1, 1, 100]	'reg_alpha': 0.1	[reg_alpha*0.2, reg_alpha*0.5, reg_alpha, reg_alpha*2, reg_alpha*5]	'reg_alpha': 0.01
'reg_lambda'	'reg_lambda': [1e-5, 1e-2, 0.1, 1, 100]	'reg_lambda': 0.01	[reg_lambda*0.2, reg_lambda*0.5, reg_lambda, reg_lambda*2, reg_lambda*5]	'reg_lambda': 0.05
'learning_rate':	'learning_rate': [0.0001, 0.001, 0.01, 0.1, 0.2, 0.3]	'learning_rate': 0.3	[learning_rate*0.2, learning_rate*0.5, learning_rate, learning_rate*2, learning_rate*5]	'learning_rate': 0.3
'n_estimators':	'n_estimators': [50, 100, 200]	'n_estimators': 100	[(n_estimators*3/4), n_estimators, n_estimators*5/4]	'n_estimators': 100

Paso 4 - Aplicación de estrategia de balanceo

Se aplican las 5 técnicas de balanceo de clases que permite XGBoost para probar mejoras en los resultados del modelo y los resultados comparativos se pueden ver en el siguiente numeral de [Análisis de resultados](#).

6.2 Análisis de resultados

Para un análisis preciso de los resultados, es fundamental considerar que durante el procesamiento de los datos se llevaron a cabo dos transformaciones en las variables: la conversión en variables "dummies" y la factorización. Estas transformaciones tuvieron un impacto significativo en los resultados obtenidos por los modelos entrenados.

Factorize:

A continuación se presentan los resultados de los modelos implementados con el uso de diferentes estrategias de calibración de datos.

Tabla: Resultados de las métricas para parámetros por defecto

Modelo	Estrategia	0			1			accuracy	prom
		precision	recall	f1	precision	recall	f1		
Decision Tree Classifier	Ninguna	0,93	0,92	0,92	0,30	0,32	0,31	0,86	0,62
	Penalización para compensar	0,92	0,92	0,92	0,30	0,30	0,30	0,86	0,61
	Subsampling en la clase mayoritaria	0,84	0,10	0,19	0,09	0,82	0,16	0,17	0,37
	Oversampling de la clase minoritaria	0,92	0,92	0,92	0,30	0,30	0,30	0,86	0,61
	Combinamos resampling con Smote-Tomek	0,93	0,87	0,90	0,26	0,43	0,33	0,82	0,62
	Ensamble de Modelos con Balanceo	0,95	0,82	0,88	0,27	0,63	0,38	0,80	0,66
Random Forest	Ninguna	0,91	0,99	0,95	0,61	0,13	0,22	0,91	0,64
	Penalización para compensar	0,91	0,99	0,95	0,64	0,12	0,20	0,91	0,64
	Subsampling en la clase mayoritaria	0,92	0,28	0,43	0,11	0,78	0,19	0,33	0,45
	Oversampling de la clase minoritaria	0,92	0,97	0,95	0,46	0,22	0,30	0,90	0,64
	Combinamos resampling con Smote-Tomek	0,93	0,91	0,92	0,34	0,41	0,37	0,86	0,65
	Ensamble de Modelos con Balanceo	0,97	0,77	0,86	0,26	0,76	0,39	0,77	0,67
AdaBoost	Ninguna	0,91	0,99	0,95	0,47	0,06	0,10	0,90	0,58
	Subsampling en la clase mayoritaria	0,92	0,28	0,43	0,11	0,78	0,19	0,33	0,45
	Oversampling de la clase minoritaria	0,97	0,72	0,82	0,23	0,78	0,36	0,72	0,65
	Combinamos resampling con Smote-Tomek	0,95	0,81	0,87	0,25	0,58	0,35	0,78	0,64
	Ensamble de Modelos con Balanceo	0,97	0,72	0,83	0,24	0,79	0,37	0,73	0,65
Gradient Boosting	Ninguna	0,91	1,00	0,95	0,62	0,07	0,13	0,90	0,61
	Subsampling en la clase mayoritaria	0,82	0,12	0,21	0,09	0,76	0,16	0,18	0,36
	Oversampling de la clase minoritaria	0,97	0,73	0,83	0,24	0,79	0,37	0,73	0,66
	Combinamos resampling con Smote-Tomek	0,94	0,83	0,89	0,26	0,55	0,36	0,81	0,64
	Ensamble de Modelos con Balanceo	0,97	0,72	0,83	0,24	0,81	0,37	0,73	0,66
XGBoost	Ninguna	0,92	0,98	0,95	0,50	0,18	0,26	0,90	0,63
	Subsampling en la clase mayoritaria	0,85	0,12	0,21	0,09	0,81	0,16	0,18	0,37
	Oversampling de la clase minoritaria	0,95	0,85	0,90	0,31	0,61	0,41	0,83	0,67
	Combinamos resampling con Smote-Tomek	0,93	0,92	0,93	0,34	0,39	0,36	0,87	0,65
	Ensamble de Modelos con Balanceo	0,97	0,78	0,86	0,28	0,77	0,41	0,78	0,68

Se puede observar que las estrategias que utilizan subsampling y oversampling para balancear las clases mejoran el recall y F1-score de la clase minoritaria, mientras que disminuyen el recall de la clase mayoritaria. La estrategia de ensamble de modelos con balanceo también logra mejorar el F1-score de la clase minoritaria, pero sin afectar significativamente el rendimiento de la clase mayoritaria.

En cuanto a las métricas de precisión y exactitud, se puede observar que no hay una estrategia que sea claramente superior a las demás. En general, las estrategias de subsampling y oversampling logran una precisión comparable a la de la estrategia sin muestreo, mientras que las estrategias de penalización y ensamble de modelos con balanceo obtienen una precisión

ligeramente inferior. La estrategia de combinación de resampling con Smote-Tomek logra una precisión intermedia.

Dummies:

Los resultados obtenidos a través de estas variables mostraron un rendimiento inferior al de la factorización, por lo tanto, se presentarán en los anexos (ver [Análisis de resultados](#)) para centrarnos en los resultados de la factorización.

El paso final, fue implementar una búsqueda en cuadrícula con el método 'GridSearchCV()' para ajustar los hiperparámetros utilizando la métrica de puntuación 'recall' para evaluar el rendimiento del modelo, y una validación cruzada de 5 veces (cv=5) para evitar el sobreajuste.

Este proceso se aplica para las 5 estrategias de tratamiento de muestra desbalanceadas y se compara su desempeño:

Tabla: Comparativos de resultados por defecto vs calibrados

Comparación de resultados									
Clase		0			1			modelo	
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
Parámetros por Defecto									
XGBoost	Ninguna	0,92	0,98	0,95	0,50	0,18	0,26	0,90	0,63
	Subsampling en la clase mayoritaria	0,85	0,12	0,21	0,09	0,81	0,16	0,18	0,37
	Oversampling de la clase minoritaria	0,95	0,85	0,90	0,31	0,61	0,41	0,83	0,67
	Combinamos resampling con Smote-Tomek	0,93	0,92	0,93	0,34	0,39	0,36	0,87	0,65
	Ensamble de Modelos con Balanceo	0,97	0,78	0,86	0,28	0,77	0,41	0,78	0,68
Parámetros Calibrados									
XGBoost	Por defecto	0,91	0,98	0,95	0,55	0,18	0,28	0,90	0,64
	Subsampling en la clase mayoritaria	0,83	0,11	0,19	0,09	0,81	0,17	0,18	0,34
	Oversampling de la clase minoritaria	0,96	0,81	0,88	0,31	0,73	0,43	0,80	0,70
	Combinamos resampling con Smote-Tomek	0,93	0,91	0,92	0,34	0,41	0,37	0,86	0,68
	Ensamble de Modelos con Balanceo	0,97	0,76	0,85	0,27	0,80	0,41	0,76	0,69

La calibración ha mejorado el desempeño de los modelos al compararlo con los hiperparámetros por defecto y se evidencia que la estrategia de oversampling de la clase minoritaria y el ensamblaje de Modelos con balanceo son las estrategias que tienen mejor desempeño al comparar los valores de la métrica de recall siendo estos los candidatos para utilizar en la predicción de la variable 'Adición'.

7 Modelo prórrogas

Con este modelo se pretende dar respuesta al Artefacto de predicción de prórrogas (Desarrollar un artefacto que permita predecir el riesgo de ocurrencia de prórrogas en los que se incurría en el contrato, el cual se podría utilizar para tomar decisiones para mejorar la eficiencia en la contratación y generar ahorros.), por lo que se busca, por medio de un modelo de clasificación, predecir si un contrato tendrá prórrogas.

7.1 Procesos de calibración o entrenamiento

Para el procesamiento de los datos, se llevaron a cabo varios procedimientos adicionales además de la selección de las variables más significativas y la limpieza de los datos. En primer lugar, se creó una nueva variable llamada Prórroga, que se define como 1 si la suma de los campos Tiempo Adiciones en Dias y Tiempo Adiciones en Meses en el conjunto de datos SECOP I es diferente de cero, y como 0 en caso contrario. También se eliminaron las filas con observaciones nulas y se dividieron las variables predictoras (X) y la variable objetivo de adición (1 y 0) (y) utilizando un 33% para el conjunto de prueba y el 67% restante para el conjunto de entrenamiento.

Además, se llevó a cabo la conversión de variables a dummies y el proceso de factorización de variables. A partir de ahí, se utilizaron diferentes modelos para correr los modelos, dependiendo de si se utilizó datos factorizados o datos dummies. Para los modelos factorizados se utilizaron Decision Tree Classifier, Random Forest, AdaBoost y Gradient Boosting, mientras que para los modelos Dummies se utilizaron LDA, QDA, Naive Bayes, Regresión Logística, Decision Tree Classifier, Random Forest, AdaBoost y Gradient Boosting. Los modelos fueron evaluados utilizando diferentes métricas de desempeño y se ajustaron sus parámetros para mejorar el desempeño. Se implementaron diferentes estrategias de desbalance y el desempeño del modelo se midió utilizando las diferentes métricas descritas en el capítulo de [Definición de métricas adecuadas de desempeño](#), con énfasis en el Recall.

7.1.1 Selección de variables

A partir de las variables seleccionadas en el numeral de [Preselección de variables explicativas](#), se realizó una verificación inicial de las variables que pueden tener correlación con la variable creada de Prórroga, por lo que se eliminaron las variables de: Prorroga, Tiempo Adiciones en Dias, Tiempo Adiciones en Meses, Adicion, Valor Total de Adiciones y Valor Contrato con Adiciones.

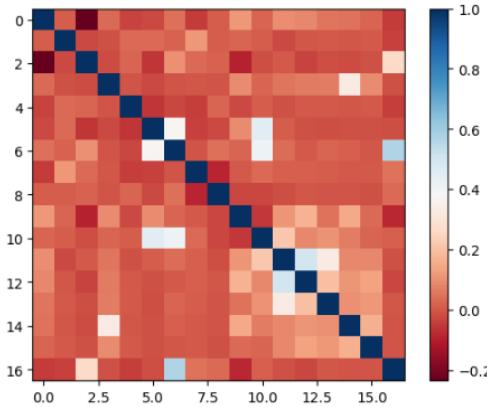
Las variables utilizadas para el modelo son: Orden Entidad, Modalidad de Contratacion, Estado del Proceso, Municipio de Obtencion, Cuantia Proceso, Nombre Grupo, Tipo Identifi del Contratista, Dpto y Muni Contratista, Anno Ini Ejec Contrato, Mes Ini Ejec Contrato, Plazo de Ejec del Contrato, Rango de Ejec del Contrato, Cuantia Contrato, Departamento Entidad.

7.1.2 Pruebas de verificación de los supuestos requeridos

Supuesto para modelos de clasificación lineal

- Análisis de independencia de observaciones:

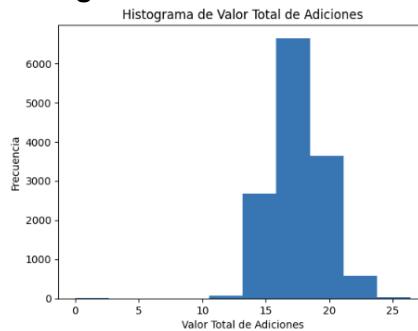
Gráfica: Matriz de correlación de las observaciones



Al observar los valores de la matriz de correlación, se puede notar que la mayoría de estos están cercanos a cero. Este hecho sugiere que no hay una relación lineal fuerte entre las variables en el conjunto de datos. En otras palabras, si el valor de una variable aumenta, no necesariamente implica que el valor de la otra variable aumente o disminuya en la misma proporción. Es posible que las variables sean independientes entre sí. Por lo tanto, la correlación cercana a cero indica que las variables no están relacionadas de manera significativa o que la relación entre ellas no es lineal. Cabe destacar que esta correlación no lineal puede no ser capturada por la matriz de correlación lineal.

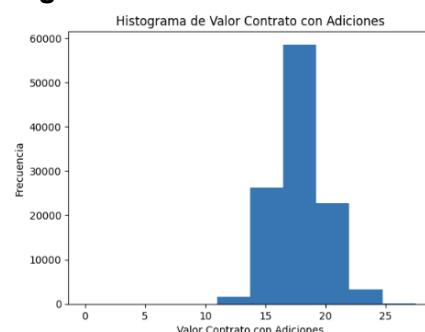
- Supuesto de normalidad de datos numéricos
 - Valor total de adicciones

Gráfica: Histograma del Valor total de adiciones



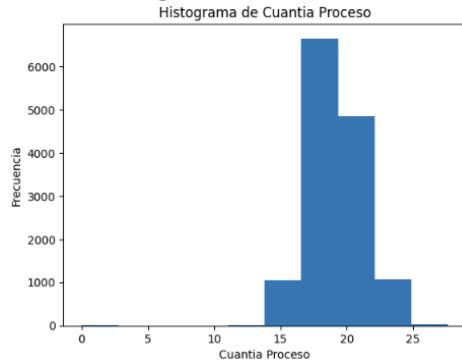
- Valor contrato con adiciones

Gráfica: Histograma del Valor contrato con adiciones



- Cuantia proceso

Gráfica: Histograma de Cuantía proceso



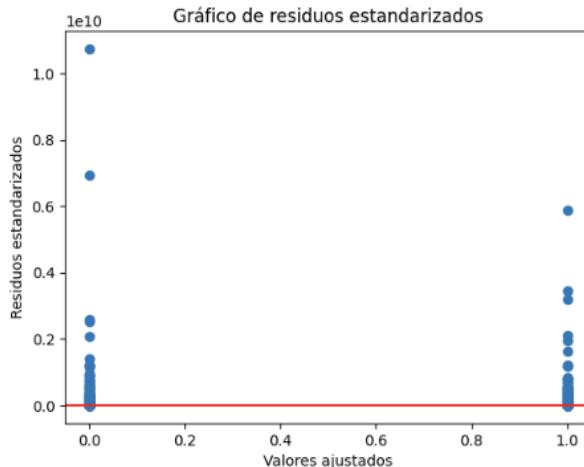
Al observar el comportamiento de los histogramas de las variables numéricas se puede apreciar que tienen una distribución parecida a la normal, por lo que se puede inferir que cumplen con el supuesto de normalidad.

- Linealidad:
 - Rainbow statistic: 1.2964988568819171
 - Rainbow p-value: 4.825511641860705e-57

Se cumple el supuesto. El rainbow statistic es cercano a 1, sugiriendo que se ajusta al supuesto de linealidad y el p-value es menor que 0.05, lo que significa que no se puede rechazar la hipótesis nula

- Homocedasticidad

Gráfica: Residuos estandarizados



Breusch-Pagan test:
 Estadístico de prueba: 24136.58138357534
 Valor p: 0.0
 Estadístico de prueba alternativo: 7259.9985694274665
 Valor p alternativo: 0.0

Hay evidencia estadística para rechazar la hipótesis nula de homocedasticidad, y que la varianza del error no es constante. En otras palabras, existe heterocedasticidad en el modelo de regresión por lo que no se cumple con el supuesto.

- Supuesto de independencia de los errores

Estadística de Durbin-Watson: 1.991797820266946
No hay autocorrelación

El resultado de la estadística sugiere que el supuesto de independencia de errores se cumple, ya que los errores no están correlacionados entre sí y, por lo tanto, son independientes.

Resultados:

Después de realizar las validaciones de los supuestos necesarios para el análisis de datos, se llevaron a cabo diversas técnicas para verificar si se cumplían o no. En concreto, se validó la normalidad mediante la visualización de un gráfico de histograma, la homocedasticidad mediante un gráfico de residuos estandarizados, y la independencia entre las observaciones mediante un gráfico de autocorrelación. Se verificó el tamaño de muestra adecuado y la existencia de clases equilibradas mediante el balanceo de las muestras. Para la linealidad se utilizó la prueba Rainbow statistic, mientras que la ausencia de multicolinealidad se verificó con un gráfico de autocorrelación. Se llevó a cabo un diagnóstico para comprobar la presencia o ausencia de valores atípicos, y se convirtieron las variables categóricas predictoras en 1 y 0 para la variable dependiente binaria. La imputación de datos faltantes se realizó con ceros, mientras que la factorización y estandarización de las entradas numéricas se hizo para mejorar la calidad de los datos.

En la siguiente tabla se muestra una síntesis de los supuestos que se cumplieron y los que no en el análisis de datos:

Gráfica: Validación de supuestos del modelo de prorrrogas

Supuestos	LDA	QDA	Naive Bayes	Regresión Logística	Decision Tree Classifier	Random Forest	AdaBoost	Gradient Boosting	XGBoost	K Vecinos
Normalidad	cumple	cumple	cumple	cumple						
Homocedasticidad	no cumple	no cumple								
Independencia	cumple	cumple		cumple			cumple	cumple	cumple	cumple
Matrices de covarianza iguales	cumple									
Tamaño de muestra adecuado		cumple	cumple	cumple						
Independencia condicional de clase			cumple							
Clases equilibradas			cumple				cumple			
Linealidad				cumple					cumple	
Ausencia de multicolinealidad				cumple	cumple					
Sin valores atípicos				cumple		cumple		cumple	cumple	
Variable dependiente binaria				cumple	cumple					

Supuestos	LDA	QDA	Naive Bayes	Regresión Logística	Decision Tree Classifier	Random Forest	AdaBoost	Gradient Boosting	XGBoost	K Vecinos
Sin datos faltantes						cumple			cumple	
Entradas numéricas								cumple	cumple	
Estandarizados										cumple

Después de realizar las validaciones correspondientes, se lograron verificar la mayoría de los supuestos necesarios para el análisis de datos. En particular, se pudo validar la normalidad, la independencia entre las observaciones, el tamaño de muestra adecuado, las clases equilibradas, la linealidad, la ausencia de multicolinealidad, la ausencia de valores atípicos, la variable dependiente binaria y la ausencia de datos faltantes. Sin embargo, se encontró que no se cumplió el supuesto de homocedasticidad, según los resultados obtenidos del test de Breusch-Pagan, lo cual solo afecta los modelos de LDA y QDA.

7.1.3 Parametrización

La revisión inicial de los modelos se realizó con los parámetros por defecto, aplicando las siguientes técnicas de balanceo según el modelo:

1. Ninguna: no se implementa ninguna técnica de balanceo de clases.
2. Penalización para compensar
3. Subsampling en la clase mayoritaria: se implementa la técnica de submuestreo 'NearMiss' en el conjunto de datos de entrenamiento seleccionando las muestras de la clase mayoritaria más cercanas a las de la clase minoritaria.
4. Oversampling de la clase minoritaria: se implementa la técnica de sobremuestreo 'RandomOverSampler' en el conjunto de datos de entrenamiento generando muestras sintéticas de la clase minoritaria para equilibrar la distribución de clases.
5. Combinamos resampling con Smote-Tomek: se implementa la técnica de sobremuestreo y submuestreo combinados SMOTETomek generando muestras sintéticas de la clase minoritaria y eliminando las muestras de la clase mayoritaria que están cerca de las muestras de la clase minoritaria.
6. Ensamble de Modelos con Balanceo: se implementa el uso de 'BalancedBaggingClassifier()' para abordar el problema de desequilibrio de clases en el conjunto de datos de entrenamiento mediante el muestreo de la clase minoritaria para equilibrar la distribución de clases.

Los resultados se pueden ver en el siguiente numeral de [Análisis de resultados](#).

A partir de estos resultados se identificó el modelo con mejor desempeño general según las métricas planteadas, que en este caso fue el modelo XGBoost.

La calibración del modelo XGBoost es un proceso de ajustar a los hiperparámetros del modelo para mejorar su rendimiento y reducir el sobreajuste. Se decide utilizar una calibración por pasos de los diferentes hiperparámetros hasta obtener la mejor combinación posible de acuerdo a la métrica definida.

Se toma como base el modelo XGBClassifier().

El proceso de calibración se realizó de la siguiente manera:

Paso 1 - Preparación de los Datos

Tomando como base el resultado de la selección de variables se continúa el trabajo de preparación de los datos para aplicarlos al modelo.

Las variables categóricas son convertidas en numéricas utilizando 'OneHotEncoder()'.

Las variables numéricas se estandarizan utilizando 'StandardScaler'.

Se hace una separación de variables predictoras (X) y variable de interés (y) y se divide la muestra en un set de entrenamiento y un set de pruebas test usando la función 'train_test_split'.

Paso 2 - Definición de los parámetros a calibrar

Se definen en un valor inicial los hiperparámetros a calibrar dentro del modelo XGBoost con los valores preestablecidos por defecto por el modelo:

- objective = 'binary:logistic'
- eval_metric = "logloss"
- learning_rate=None
- max_depth = 6
- n_estimators=100
- colsample_bytree = 1
- gamma = 0
- min_child_weight = 1
- reg_alpha=None
- reg_lambda=None

Paso 3 - Calibración Iterativa de los hiperparámetros

La estrategia es implementar una búsqueda en cuadrícula con el método 'GridSearchCV()' para ajustar los hiperparámetros utilizando la métrica de puntuación 'recall' para evaluar el rendimiento del modelo, y una validación cruzada de 5 veces (cv=5) para evitar el sobreajuste.

La búsqueda se desarrolla de la siguiente manera.

1. Selecciona uno o dos hiperparámetros para calibrar los demás permanecen constantes.
2. Prueba valores dentro de un rango predefinido para cada hiperparámetro.
3. Evalúa cada combinación de hiperparámetros contra la métrica seleccionada.
4. Guarda temporalmente la combinación de mejor desempeño.
5. Prueba nuevamente con una grilla más detallada alrededor de los parámetros seleccionados como temporales.
6. Guarda el resultado de esta última prueba como el valor definitivo para los hiperparámetros.

Tabla: Calibración hiperparámetros

Hiperparametro	Rango de prueba inicial	Óptimo Temporal	Rango de prueba detallado	Óptimo final
'max_depth'	range(3,10,2)	max_depth': 5	[max_depth-1, max_depth+1]	max_depth': 5

Hiperparametro	Rango de prueba inicial	Óptimo Temporal	Rango de prueba detallado	Óptimo final
'min_child_weight'	[1,3,5]	'min_child_weight': 3	[min_child_weight-1, min_child_weight-0.5, min_child_weight, min_child_weight+0.5, min_child_weight+1]	'min_child_weight': 2.5
'gamma'	[i/10.0 for i in range(1,10,2)]	'gamma': 0.7	['gamma'-1, 'gamma'-0.5, 'gamma', 'gamma'+0.5, 'gamma'+1]	'gamma': 0.7
'subsample'	'subsample': [i/10.0 for i in range(6,11)]	'subsample': 0.9	[i/100.0 for i in range(int((subsample-0.1)*100.0), min(int((subsample+0.1)*100),105) , 5)]	'subsample': 0.9
'colsample_bytree':	'colsample_bytree': [i/10.0 for i in range(6,11)]	'colsample_bytree': 0.6	[i/100.0 for i in range(int((colsample_bytree-0.1)*100.0), min(int((subsample+0.1)*100),105), 5)]	'colsample_bytree': 0.75
'reg_alpha':	'reg_alpha': [1e-5, 1e-2, 0.1, 1, 100]	'reg_alpha': 0.1	[reg_alpha*0.2, reg_alpha*0.5, reg_alpha*2, reg_alpha*5]	'reg_alpha': 0.01
'reg_lambda'	'reg_lambda': [1e-5, 1e-2, 0.1, 1, 100]	'reg_lambda': 0.01	[reg_lambda*0.2, reg_lambda*0.5, reg_lambda*2, reg_lambda*5]	'reg_lambda': 0.05
'learning_rate':	'learning_rate': [0.0001, 0.001, 0.01, 0.1, 0.2, 0.3]	'learning_rate': 0.3	[learning_rate*0.2, learning_rate*0.5, learning_rate, learning_rate*2, learning_rate*5]	'learning_rate': 0.3
'n_estimators':	'n_estimators': [50, 100, 200]	'n_estimators': 100	[(n_estimators*3/4), n_estimators, n_estimators*5/4]	'n_estimators': 100

Paso 4 - Aplicación de estrategia de balanceo

Se aplican las 5 técnicas de balanceo de clases que permite XGBoost para probar mejoras en los resultados del modelo y los resultados comparativos se pueden ver en el siguiente numeral de [Análisis de resultados](#).

7.2 Análisis de resultados

Para analizar los resultados, es importante tener en cuenta que se realizaron dos transformaciones de las variables durante el procesamiento de los datos: la conversión de las variables en dummies y la transformación por factorización. Estas transformaciones dieron lugar a diferentes resultados en los modelos entrenados.

Factorize:

A continuación se presentan los resultados de los modelos implementados con el uso de diferentes estrategias de calibración de datos.

Tabla: Resultados de las métricas para parámetros por defecto

Modelo	Estrategia	0			1			accuracy	prom
		precision	recall	f1	precision	recall	f1		
Decision Tree Classifier	Ninguna	0,93	0,92	0,92	0,35	0,38	0,36	0,86	0,64
	Penalización para compensar	0,92	0,93	0,92	0,35	0,34	0,35	0,87	0,64
	Subsampling en la clase mayoritaria	0,80	0,11	0,19	0,09	0,77	0,17	0,18	0,36
	Oversampling de la clase minoritaria	0,92	0,93	0,93	0,36	0,33	0,35	0,87	0,64
	Combinamos resampling con Smote-Tomek	0,94	0,87	0,90	0,32	0,51	0,39	0,83	0,66
	Ensamble de Modelos con Balanceo	0,96	0,84	0,90	0,34	0,68	0,45	0,83	0,70
Random Forest	Ninguna	0,92	0,98	0,95	0,63	0,27	0,38	0,91	0,69
	Penalización para compensar	0,91	0,99	0,95	0,63	0,22	0,33	0,90	0,67
	Subsampling en la clase mayoritaria	0,91	0,27	0,42	0,11	0,77	0,19	0,32	0,45
	Oversampling de la clase minoritaria	0,93	0,97	0,95	0,56	0,34	0,43	0,90	0,70
	Combinamos resampling con Smote-Tomek	0,94	0,91	0,93	0,41	0,53	0,46	0,87	0,70
	Ensamble de Modelos con Balanceo	0,97	0,79	0,87	0,31	0,81	0,45	0,79	0,70
AdaBoost	Ninguna	0,91	0,98	0,95	0,58	0,21	0,31	0,90	0,66
	Subsampling en la clase mayoritaria	0,90	0,28	0,43	0,11	0,73	0,19	0,33	0,44
	Oversampling de la clase minoritaria	0,97	0,75	0,85	0,28	0,82	0,41	0,76	0,68
	Combinamos resampling con Smote-Tomek	0,95	0,83	0,89	0,30	0,64	0,41	0,81	0,67
	Ensamble de Modelos con Balanceo	0,97	0,75	0,84	0,28	0,82	0,41	0,75	0,68
Gradient Boosting	Ninguna	0,92	0,98	0,95	0,63	0,24	0,34	0,90	0,68
	Subsampling en la clase mayoritaria	0,80	0,12	0,21	0,09	0,75	0,16	0,19	0,36
	Oversampling de la clase minoritaria	0,97	0,76	0,85	0,29	0,82	0,43	0,77	0,69
	Combinamos resampling con Smote-Tomek	0,95	0,85	0,90	0,33	0,63	0,43	0,83	0,68
	Ensamble de Modelos con Balanceo	0,97	0,76	0,85	0,29	0,83	0,43	0,76	0,69
XGBoost	Ninguna	0,92	0,97	0,95	0,57	0,30	0,39	0,90	0,68
	Subsampling en la clase mayoritaria	0,82	0,12	0,20	0,10	0,79	0,17	0,19	0,37
	Oversampling de la clase minoritaria	0,96	0,87	0,91	0,38	0,68	0,48	0,85	0,71
	Combinamos resampling con Smote-Tomek	0,84	0,92	0,93	0,41	0,50	0,45	0,87	0,68
	Ensamble de Modelos con Balanceo	0,97	0,80	0,88	0,32	0,82	0,46	0,80	0,71

Se utilizaron diferentes modelos para decidir cuál de ellos ofrecía una mejor predicción de los datos. En el análisis, se emplearon cuatro modelos de clasificación: Decision Tree Classifier, Random Forest, AdaBoost y Gradient Boosting. Se evaluaron seis estrategias para balancear los datos: ninguna, penalización para compensar, subsampling en la clase mayoritaria, oversampling de la clase minoritaria, combinación de resampling con Smote-Tomek y ensamble de modelos con balanceo.

Para el modelo Decision Tree Classifier, las estrategias de penalización para compensar y la combinación de resampling con Smote-Tomek mostraron un mejor desempeño en términos de

precisión, recall, f1 y exactitud, mientras que el subsampling en la clase mayoritaria tuvo el peor rendimiento.

Para el modelo Random Forest, las estrategias de ninguna y oversampling de la clase minoritaria tuvieron un mejor rendimiento en términos de precisión, recall, f1 y exactitud, mientras que el subsampling en la clase mayoritaria tuvo el peor rendimiento.

Para el modelo AdaBoost, la estrategia de oversampling de la clase minoritaria tuvo el mejor rendimiento en términos de precisión, recall, f1 y exactitud, mientras que la estrategia de ninguna tuvo el peor desempeño.

Para el modelo Gradient Boosting, la estrategia de oversampling de la clase minoritaria tuvo el mejor rendimiento en términos de precisión, recall, f1 y exactitud, mientras que el subsampling en la clase mayoritaria tuvo el peor rendimiento.

Notamos que la estrategia de oversampling de la clase minoritaria es efectiva para mejorar el desempeño de los modelos en términos de precisión, recall, f1 y exactitud, mientras que el subsampling en la clase mayoritaria suele tener el peor rendimiento. Además, las estrategias de penalización para compensar y la combinación de resampling con Smote-Tomek también mostraron un buen rendimiento en algunos modelos.

Tras haber realizado un análisis exhaustivo del rendimiento de distintos modelos aplicando diversas estrategias, se concluye que los resultados más destacados se obtienen mediante el uso de XGBoost en combinación con la técnica de oversampling aplicada a la clase minoritaria. Ya que con este es el que se obtiene un mejor balance entre los resultados del recall, el f1 score y la predicción del modelo.

Dummies:

Los resultados obtenidos a través de estas variables mostraron un rendimiento inferior al de la factorización, por lo tanto, se presentarán en los anexos (ver [Análisis de resultados](#)) para centrarnos en los resultados de la factorización.

El paso final, fue implementar una búsqueda en cuadrícula con el método 'GridSearchCV()' para ajustar los hiperparámetros utilizando la métrica de puntuación 'recall' para evaluar el rendimiento del modelo, y una validación cruzada de 5 veces (cv=5) para evitar el sobreajuste.

Este proceso se aplica para las 5 estrategias de tratamiento de muestra desbalanceadas y se compara su desempeño:

Tabla: Comparativos de resultados por defecto vs calibrados

Comparación de resultados									
Clase		0			1			modelo	
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
Parámetros por Defecto									
XGBoost	Ninguna	0,92	0,97	0,95	0,57	0,30	0,39	0,90	0,68
	Subsampling en la clase mayoritaria	0,82	0,12	0,20	0,10	0,79	0,17	0,19	0,37
	Oversampling de la clase minoritaria	0,96	0,87	0,91	0,38	0,68	0,48	0,85	0,71
	Combinamos resampling con Smote-Tomek	0,84	0,92	0,93	0,41	0,50	0,45	0,87	0,68
	Ensamble de Modelos con Balanceo	0,97	0,80	0,88	0,32	0,82	0,46	0,80	0,71

Comparación de resultados									
Clase		0			1			modelo	
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
Calibrado									
XGBoost	Por defecto	0,92	0,98	0,95	0,62	0,28	0,39	0,91	0,72
	Subsampling en la clase mayoritaria	0,86	0,14	0,24	0,10	0,80	0,18	0,21	0,36
	Oversampling de la clase minoritaria	0,96	0,83	0,89	0,34	0,72	0,46	0,82	0,72
	Combinamos resampling con Smote-Tomek	0,94	0,91	0,93	0,41	0,51	0,45	0,87	0,72
	Ensamble de Modelos con Balanceo	0,97	0,80	0,88	0,32	0,79	0,46	0,80	0,72
Calibrado									
XGBoost	Por defecto	0,92	0,91	0,92	0,35	0,37	0,36	0,86	0,67
	Subsampling en la clase mayoritaria	0,83	0,14	0,24	0,10	0,76	0,17	0,21	0,35
	Oversampling de la clase minoritaria	0,95	0,86	0,90	0,34	0,60	0,43	0,83	0,70
	Combinamos resampling con Smote-Tomek	0,94	0,86	0,90	0,31	0,51	0,39	0,83	0,68
	Ensamble de Modelos con Balanceo	0,96	0,80	0,87	0,30	0,71	0,42	0,79	0,69

La calibración ha mejorado el desempeño de los modelos al compararlo con los hiperparámetros por defecto y se evidencia que la estrategia de oversampling de la clase minoritaria y el ensamble de Modelos con balanceo son las estrategias que tienen mejor desempeño al comparar los valores de la métrica de recall siendo estos los candidatos para utilizar en la predicción de la variable Prórroga.

8 Implementación de ajustes

8.1 Modelos

En el modelo descriptivo es necesario realizar el ajuste en los nombres de los clusters encontrados, ya que es importante tener un buen nombre para los clusters porque ayuda a comunicar de manera clara y efectiva los resultados del análisis de clustering. Los nombres adecuados y descriptivos permiten que los resultados sean fácilmente interpretados por las personas que no están familiarizadas con la técnica de clustering, lo que a su vez puede ayudar a tomar decisiones informadas basadas en los resultados del análisis. Además, un nombre apropiado también puede facilitar la comunicación y colaboración entre diferentes equipos y áreas dentro de una organización que puedan estar trabajando con los mismos conjuntos de datos.

Tomando en cuenta esto, en primera instancia se sugieren los siguientes nombres para los clusters calculados:

- Contratos de alta complejidad: Agrupará contratos que presentan características complejas en términos de tecnología, financiación, tiempo de ejecución, entre otros.
- Contratos de baja complejidad: Agrupará contratos que son sencillos en cuanto a sus requerimientos técnicos, de tiempo y de financiación, podrías nombrarlo así.
- Contratos en regiones remotas: Contratos cuyas obras se ejecutan en zonas remotas o de difícil acceso, este nombre podría ser adecuado.

- Contratos con retrasos: Contratos que han presentado retrasos en su ejecución.

En cuanto al modelo de estados finales es necesario ajustarlo y complementarlo con la información que obtendrá el usuario al momento de consultar la predicción sobre el estado final del proceso para un contrato dado se decide integrar al reporte un informe sobre sanciones que tenga el contratista.

Esto se realiza mediante la creación de un robot que ingresará a la página web de Portal Anticorrupción de Colombia - PACO introduce todos los nit y cédulas del representante legal del contratista y descarga el reporte de sanciones: disciplinarias, contractuales y fiscales. Este reporte es guardado en una csv que será consultado por el aplicativo cada vez que se introduzca un nit o cédula en la predicción.

Es fundamental la creación de robot que extraiga la información ya que el portal PACO no dispone de una base de datos descargable y disponible para consulta pública. La forma de consultar la información en el portal limita los reportes a consultas individuales. Esta forma de presentar la información para consultas públicas es muy común a través de las diferentes entidades gubernamentales que podrían tener información relevante para estos análisis.

8.2 Esquema general de solución

La solución planteada inicialmente tiene un esquema que contempla un tablero de control compuesto de tres páginas con información de: Estadísticas descriptivas, Alertas de Contratación y el aplicativo de la predicción.

Tabla: Esquema general de la solución inicial

Tablero de control

Estadísticas descriptivas

Este tablero está diseñado para introducir al usuario a la problemática de la contratación, darle contexto sobre el tipo y el contenido de los datos, presentar un resumen de los datos más utilizados en los análisis, mostrar patrones o tendencias más relevantes halladas y en general dar al usuario la información suficiente y de forma autocontenido al usuario para que pueda hacer uso de la herramienta.

Alertas en contratación

El tablero mostrará en formato de alertas los contratos que podrían terminar en estados considerados problemáticos.

Aplicativo de Predicción

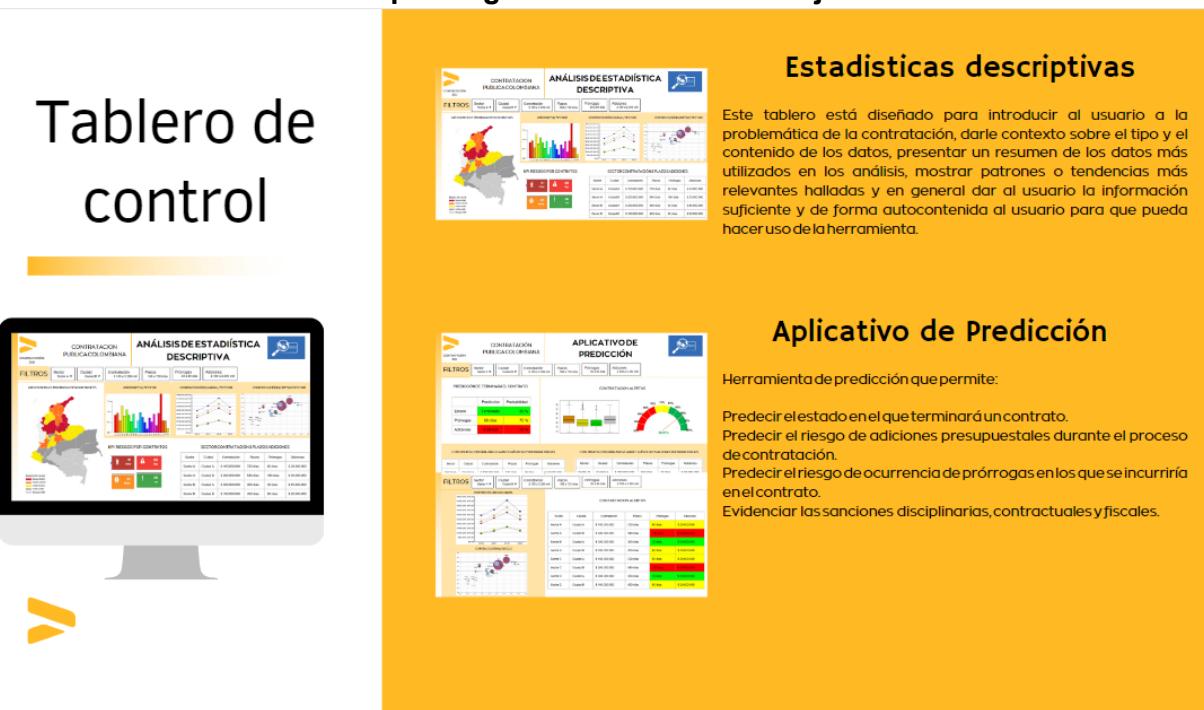
Este aplicativo busca que un usuario pueda consultar el estado final en el que terminará un contrato a partir de ciertos atributos conocidos.

Durante el desarrollo de la herramienta y los modelos a implementar se evidencia la necesidad de ajustar el esquema de la solución inicialmente propuesto con el fin de simplificar algunas

características para facilitar la consulta de los usuarios y complementar el aplicativo de predicción con alertas de sanciones.

El esquema ajustado quedaría de la siguiente forma:

Tabla: Esquema general de la solución ajustada



Las estadísticas descriptivas se complementan con los resultados de los métodos de clustering en dónde se pueden identificar grupos de contratos con características muy similares entre sí lo que genera un insumo adicional para el análisis al incorporar un componente de comportamiento de grupos a las estadísticas descriptivas generales.

El aplicativo de predicción se unifica en un sólo página para simplificar el uso del artefacto para el usuario. Se incorporan todos los resultados de la predicción y se unen las alertas por sanciones en un sólo dashboard permitirá las siguientes funciones:

- Predecir el estado en el que terminará un contrato.
- Predecir el riesgo de adiciones presupuestales durante el proceso de contratación.
- Predecir el riesgo de ocurrencia de prórrogas en los que se incurría en el contrato.
- Evidenciar las sanciones disciplinarias, contractuales y fiscales.

8.3 Problema de negocio

Se considera que los ajustes mencionados son suficientes para la fase uno del prototipo, donde se busca desarrollar una herramienta que identifique alertas y predicciones en los contratos en etapas tempranas, para prevenir problemas y sobrecostos en su cierre.

Para lograr esto, se utilizarán datos históricos para analizar la eficiencia en la contratación, comparando lo presupuestado con lo ejecutado e identificando características de riesgo como adiciones presupuestales y prórrogas de tiempo. Se incorpora la predicción sobre el estado en

el que terminará el contrato y se complementa con alertas sobre sanciones disciplinarias, contractuales y fiscales.

Estas alertas pueden ser utilizadas por los encargados de la ejecución y la auditoría, así como por los entes de control para enfocar sus esfuerzos y recursos limitados en los contratos que podrían presentar problemas haciendo una gestión más eficiente y metódica del seguimiento contractual.

9 Plan de implementación del prototipo

En función de las características o requerimientos pendientes por implementar en las etapa de prototipo y la etapa de implementación se propone el siguiente plan:

Tabla: Plan de implementación

TAREA	PROGRESO	INICIO	FIN	Semanas de desarrollo del proyecto									
				S1	S2	S3	S4	S5	S6	S7	S8	S9	S10 +
Fases de la Etapa 1 del proyecto - Prototipo													
Fase 1: Análisis de requisitos y diseño	100%												
Revisión y validación de los requerimientos propuestos en la tabla de requerimientos. Identificar los requerimientos detallados que el artefacto y sus componentes deben satisfacer.	100%	S1	S2										
Diseño del prototipo de la herramienta. Mediante un diagrama esquemático que relacione sus requerimientos, forma de uso, modelos de análisis, y procesos de adquisición y tratamiento de datos.	100%	S1	S2										
Selección de métricas y pruebas de evaluación. Validar que el prototipo y sus componentes satisfacen los requerimientos establecidos.	100%	S1	S2										
Selección de las fuentes de datos necesarias para el entrenamiento de los modelos.	100%	S1	S2										
Selección de las técnicas de aprendizaje supervisado a utilizar.	100%	S1	S2										
Diseño de los modelos de predicción.	100%	S1	S2										
Fase 2: Modelos y parametrización	82%												
Preparación de los datos base para ejecutar los modelos.	100%	S3	S5										
Implementación de los modelos de predicción. Con datos de entrenamiento y test.	100%	S3	S3										
Ejecución exhaustiva de experimentos para todos los modelos y datos de interés. Ajuste de los parámetros del modelo.	100%	S4	S4										
Evaluación de completitud de los modelos.	100%	S5	S5										
Definir las estadísticas descriptivas relevantes.	100%	S3	S5										
Hallar y recopilar nuevos datos complementarios. Principalmente portal.paco.gov.co donde se descargaran los contratos que hayan tenido alguna sanción en el transcurso de su desarrollo.	80%	S3	S6										
Fase 3: Creación de la herramienta	0%												
Conexión de los datos con la herramienta.	0%	S6	S6										
Desarrollar el tablero de control hoja de gráficas descriptivas y alertas.	0%	S6	S6										
Desarrollar el tablero de control Configurar predicciones.	0%	S7	S7										
Validar prototipo. Pruebas para garantizar que el tablero sea fácil de usar y que las visualizaciones sean claras y fáciles de entender.	0%	S7	S7										
Ensamblar la solución, adicionar elementos para la interfaz interactiva, automatizaciones en adquisición de datos, u otros módulos de procesamiento, además del refinamiento, ajuste y escalamiento de los modelos analíticos.	0%	S7	S7										
Fase 4: Entrega	0%												
Entrega del artefacto al cliente.	0%	S8	S8										
Capacitación del cliente en el uso del artefacto.	0%	S8	S8										
Soporte y mantenimiento del artefacto para corregir errores y mejorar su funcionalidad.	0%	S8	S8										

TAREA	PROGRESO	INICIO	FIN	Semanas de desarrollo del proyecto									
				S1	S2	S3	S4	S5	S6	S7	S8	S9	S10 +
Fase 5: Etapa 2 proyecto - implementación y seguimiento	0%												
Implementar una métrica o validar con datos nuevos para verificar la generación de ahorros en la contratación a través de la predicción del riesgo de gastos adicionales y la activación de un control preventivo. De esta manera, se podrá evaluar la efectividad de esta estrategia en la gestión de contrataciones y tomar medidas adicionales en caso de ser necesario.	0%	S9	S10 +										
Identificar la cantidad de contratos que se encuentran en estado de adición para validar la mejora en la eficiencia del control de gastos a través de la predicción de estados con riesgo de sobrecostos en los contratos. De esta manera, se podrá medir el impacto y la efectividad de la predicción en la reducción de costos adicionales en los contratos en estado de adición.	0%	S9	S10 +										
Identificar la cantidad de contratos en estado de prórroga para validar la mejora en el desempeño de la contratación a través de la predicción de estados con riesgo de ocurrencia de prórrogas en los contratos. De esta manera, se podrá evaluar el impacto y la eficacia de la predicción en la reducción de las prórrogas en los contratos y en la mejora de la planificación de la contratación.	0%	S9	S10 +										
Realizar una encuesta de satisfacción a los usuarios para evaluar su comprensión del comportamiento de la contratación. De esta manera, se podrá obtener una retroalimentación valiosa y mejorar la comprensión y el uso de los datos en la planificación y la toma de decisiones de la contratación.	0%	S9	S10 +										
Realizar una encuesta de satisfacción a los usuarios es esencial para evaluar la pertinencia del artefacto desarrollado con el objetivo de mejorar el entendimiento de los datos de la contratación. De esta manera, se podrá obtener retroalimentación valiosa de los usuarios y mejorar el diseño del artefacto, su funcionalidad y su facilidad de uso para asegurar una mayor eficacia en la comprensión y la toma de decisiones basadas en los datos de la contratación	0%	S9	S10 +										
Mantenimiento de modelo predicción del estado en el que terminará el contrato (Terminado, Suspenido)	0%	S9	S10 +										
Mantenimiento de modelo de predicción de estados con riesgo de sobrecostos (Prórrogas, Adiciones, Cedido) en los que incurría en el contrato.	0%	S9	S10 +										
Mantenimiento de modelo de predicción días de prórroga en los que incurría en el contrato.	0%	S9	S10 +										
Mantenimiento de modelo de predicción de valores de adición en los que incurría en el contrato.	0%	S9	S10 +										
Mantener o mejorar el tiempo óptimo en el procesamiento y generación de resultados.	0%	S9	S10 +										
Actualizar la documentación de ETL, Modelos y Tablero de Control	0%	S9	S10 +										

10 Referencias

Alfaro Rojas, D., Leguizamo Almanza, K. A., Londoño Galvis, D. A., & Maturana Córdoba, M. A. (2022). *Análisis de datos históricos para la eficiencia en la contratación Pública - Anteproyecto*. Bogotá, Colombia.

ALGORITMO DE SELECCIÓN Y VALIDACIÓN DEL MÉTODO DE CLUSTERIZACIÓN ÓPTIMO PARA DATOS NO SUPERVISADOS Universidad Tecnológica. (n.d.). Universidad Tecnológica de Pereira. Retrieved May 6, 2023, from <https://repositorio.utp.edu.co/server/api/core/bitstreams/9c69e97a-7742-4553-8c4f-df3609245065/content>

Algoritmo Meta-Heurístico para Clustering Particional de Datos basado en Global-Best Harmony Search, K-means y Restricted Growth Strings. (n.d.). Repositorio. Retrieved May 6, 2023, from <http://repositorio.unicauca.edu.co:8080/xmlui/bitstream/handle/123456789/1733/ALGORITMO%20META-HEUR%C3%88DSTICO%20PARA%20CLUSTERING%20PARTICIONAL%20DE%20DATOS%20BASADO%20EN%20GLOBAL-BEST.pdf?sequence=1>

Aprende Machine Learning. (2019, May 16). *Clasificación con datos desbalanceados*. Aprende Machine Learning. Retrieved April 27, 2023, from <https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/>

Colombia Compra Eficiente. (2020, enero). *Capacitación Secop*. Colombia Compra Eficiente. Retrieved April 21, 2023, from <https://colombiacompra.gov.co/secop/que-es-el-secop-i/capacitacion-secop>

Elizabeth León Guzmán, E. (n.d.). *Métricas para la validación de Clustering*. Ingeniería. Retrieved May 6, 2023, from https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

Suryakanthi, T. (2020, 01). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2), 612-619. <https://www.researchgate.net/journal/International-Journal-of-Advanced-Computer-Science-and-Applications-2156-5570>

Trefethen, L. N., & Bau, D. (2009, May). *Principal Component Analysis*. Duke People. Retrieved May 6, 2023, from <https://people.duke.edu/~hpgavin/SystemID/References/Richardson-PCA-2009.pdf>

11 Anexos

11.1 Tratamiento previo de los datos

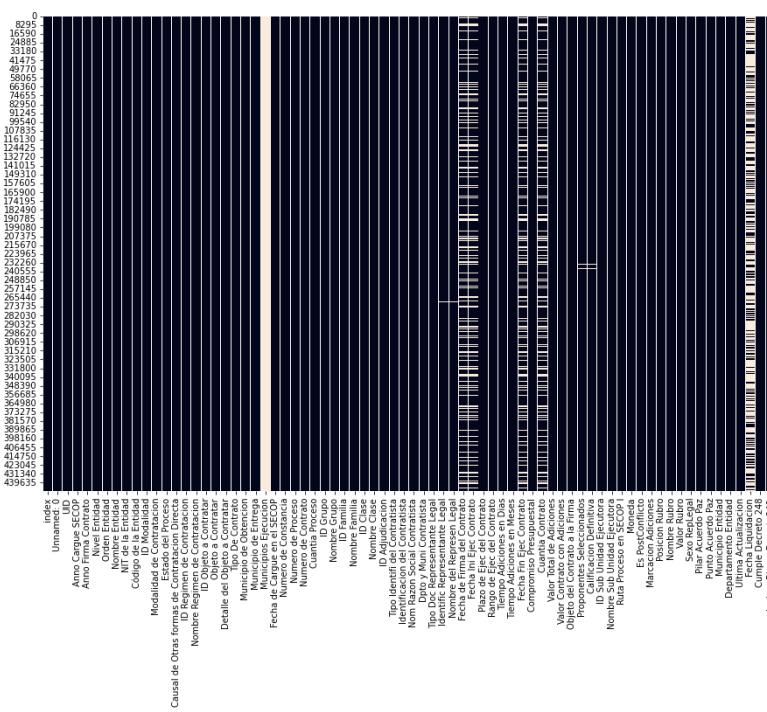
11.1.1 Valores nulos

Para identificar los valores nulos se utilizaron 2 técnicas.

Contar la cantidad de registros con información de cada variable, adicionalmente un conteo total de filas de la base de datos, determinando que si la cantidad de registros de la variable era mejor al número de filas de la base de datos esa variable contaba con valores nulos o faltantes.

Una gráfica, donde visualmente se logra identificar (espacios en blanco) las variables que cuentan con valores nulos.

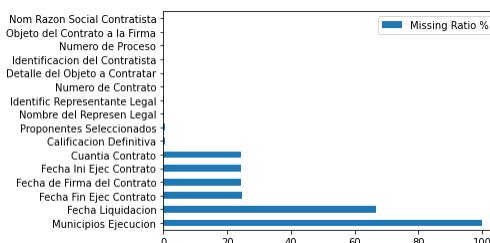
Gráfica: Valores nulos



11.1.1.1 Valores nulos SECOP I

Se revisan valores nulos por columna en las siguientes gráficas y tablas.

Gráfica: Porcentaje de datos nulos por columna en SECOP I



Se identifica que la columna Municipios Ejecución presenta un 100% de datos nulos, seguida

de Fecha de liquidación con un 65% de datos nulos.

Tabla: Datos no nulos SECOP I

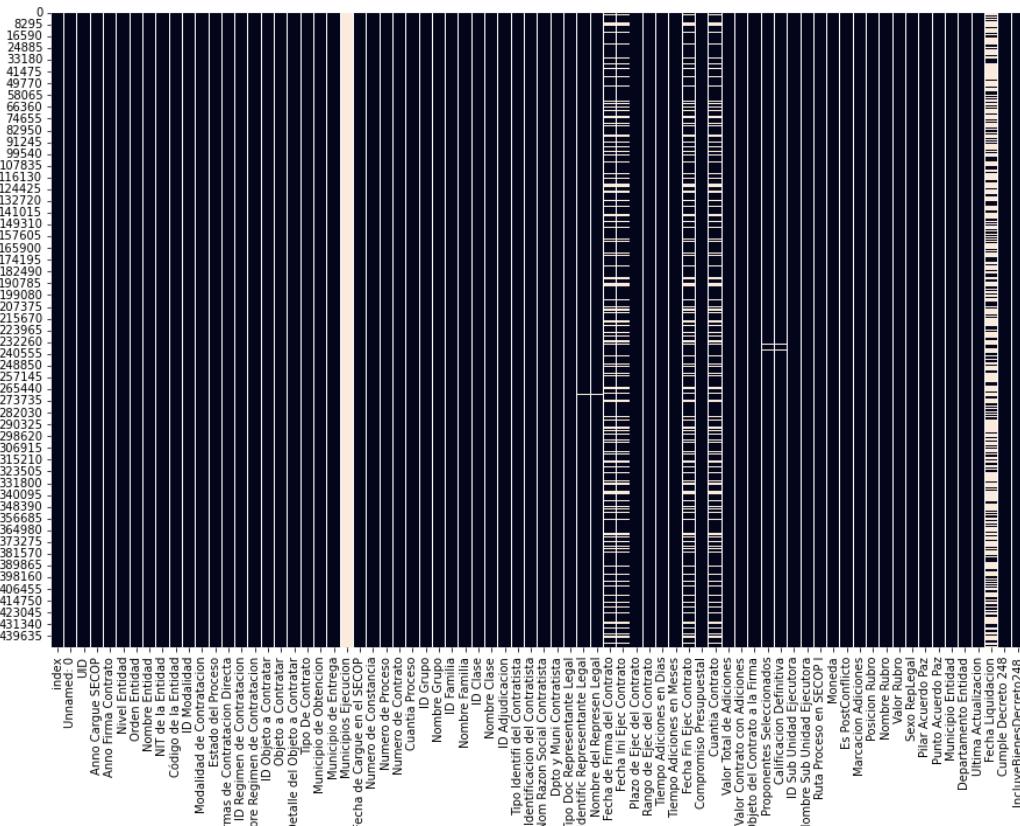
#	Columna	Non-Null	#	Columna	Non-Null
1	UID	447876	37	Dpto y Muni Contratista	447876
2	Anno Cargue SECOP	447876	38	Tipo Doc Representante Legal	447876
3	Anno Firma Contrato	447876	39	Identific Representante Legal	447385
4	Nivel Entidad	447876	40	Nombre del Represen Legal	447239
5	Orden Entidad	447876	41	Fecha de Firma del Contrato	338272
6	Nombre Entidad	447876	42	Fecha Ini Ejec Contrato	338272
7	NIT de la Entidad	447876	43	Plazo de Ejec del Contrato	447876
8	Código de la Entidad	447876	44	Rango de Ejec del Contrato	447876
9	ID Modalidad	447876	45	Tiempo Adiciones en Días	447876
10	Modalidad de Contratación	447876	46	Tiempo Adiciones en Meses	447876
11	Estado del Proceso	447876	47	Fecha Fin Ejec Contrato	337498
12	Causal de Otras formas de Contratación Directa	447876	48	Compromiso Presupuestal	447876
13	ID Regimen de Contratacion	447876	49	Cuantía Contrato	338272
14	Nombre Regimen de Contratacion	447876	50	Valor Total de Adiciones	447876
15	ID Objeto a Contratar	447876	51	Valor Contrato con Adiciones	447876
16	Objeto a Contratar	447876	52	Objeto del Contrato a la Firma	447865
17	Detalle del Objeto a Contratar	447851	53	Proponentes Seleccionados	446187
18	Tipo De Contrato	447876	54	Calificación Definitiva	445606
19	Municipio de Obtención	447876	55	ID Sub Unidad Ejecutora	447876
20	Municipio de Entrega	447876	56	Nombre Sub Unidad Ejecutora	447876
21	Municipios Ejecucion	0	57	Ruta Proceso en SECOP I	447876
22	Fecha de Cargue en el SECOP	447876	58	Moneda	447876
23	Número de Constancia	447876	59	Es PostConflict	447876
24	Número de Proceso	447863	60	Marcación Adiciones	447876
25	Número de Contrato	447734	61	Posición Rubro	447876
26	Cuantia Proceso	447876	62	Nombre Rubro	447876
27	ID Grupo	447876	63	Valor Rubro	447876
28	Nombre Grupo	447876	64	Sexo RepLegal	447876
29	ID Familia	447876	65	Pilar Acuerdo Paz	447876
30	Nombre Familia	447876	66	Punto Acuerdo Paz	447876
31	ID Clase	447876	67	Municipio Entidad	447876
32	Nombre Clase	447876	68	Departamento Entidad	447876

#	Columna	Non-Null	#	Columna	Non-Null
33	ID Adjudicación	447876	69	Última Actualización	447876
34	Tipo Identifi del Contratista	447876	70	Fecha Liquidación	149370
35	Identificación del Contratista	447852	71	Cumple Decreto 248	447876
36	Nom Razon Social Contratista	447868	72	IncluyeBienesDecreto248	447876

En la anterior tabla se resaltan en rojo las columnas con más del 20% de valores nulos y con amarillo las que tienen menos del 20%, lo cual también se puede apreciar en la siguiente gráfica, y se observa que 56 de 72 variables cuentan con el total de datos, es decir el 77,7%.

Los datos restantes cuentan con algún porcentaje de datos faltantes, siendo Fecha Liquidación el que cuenta con menos registros ya que su completitud es del 33,33% y Municipios Ejecución que no cuenta con ningún dato.

Gráfica: Datos nulos SECOP I

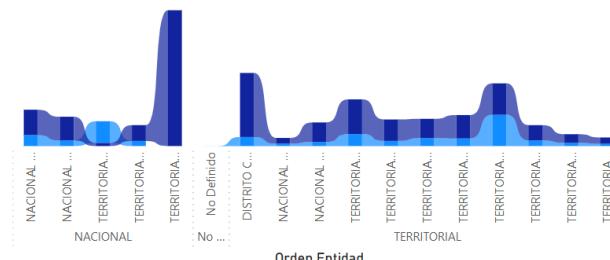


11.1.2 Datos para predicción de adiciones y prórrogas

Gráfica: SECOP I - Prórroga - Cuantía

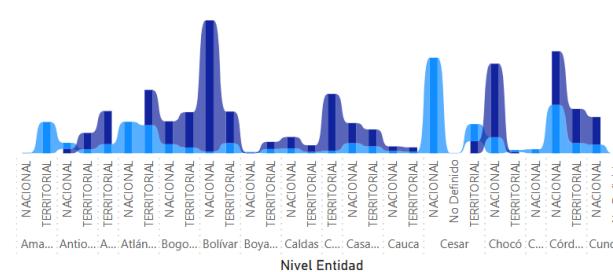
Promedio de Cuantia Proceso por Nivel Entidad, Orden Entidad y Prorroga

Prorroga ● No Prorrogado ● Prorrogado



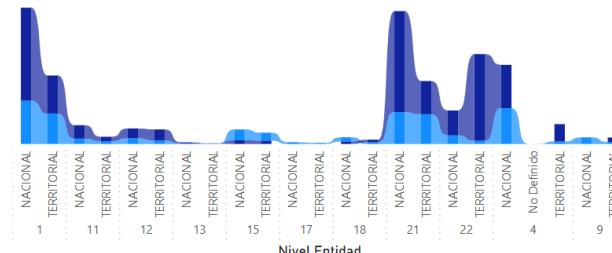
Promedio de Cuantia Proceso por Departamento Entidad, Nivel Entidad y Prorroga

Prorroga • No Prorrogado • Prorrogado



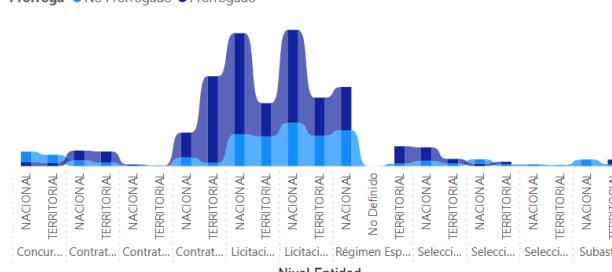
Promedio de Cuantia Proceso por ID Regimen de Contratacion, Nivel Entidad y Prorroga

Prorroga ● No Prorrogado ● Prorrogado



Promedio de Cuantia Proceso por Modalidad de Contratacion, Nivel Entidad y

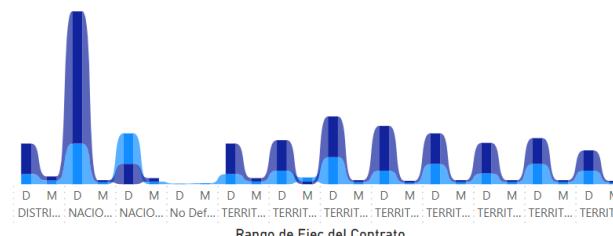
Prorroga



Gráfica: SECOP I - Prórroga - Plazo

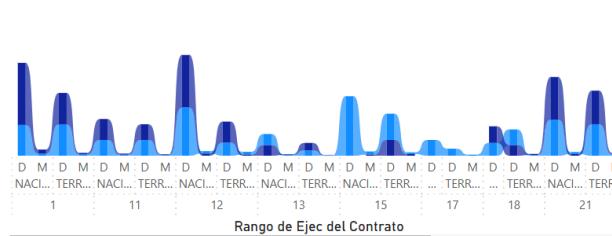
Mediana de Plazo de Ejec del Contrato por Orden Entidad, Rango de Ejec del Contrato y Prorroga

Prorroga ● No Prorrogado ● Prorrogado



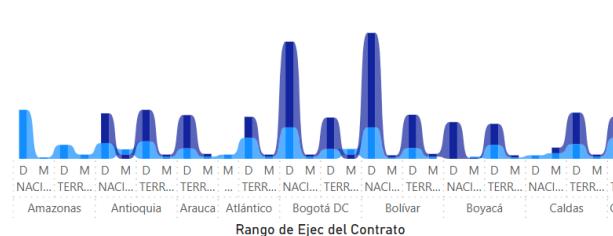
Mediana de Plazo de Ejec del Contrato por ID Regimen de Contratacion, Nivel Entidad, Rango de Ejec del Contrato y Prorroga

Prorroga ● No Prorrogado ● Prorrogado



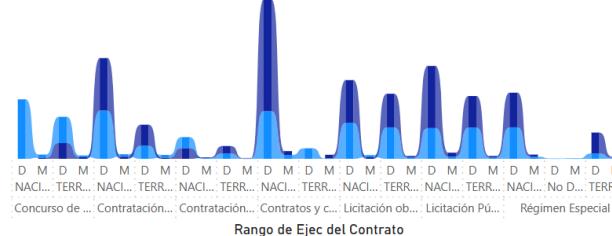
Mediana de Plazo de Ejec del Contrato por Departamento Entidad, Nivel Entidad, Rango de Ejec del Contrato y Prorroga

Prorroga ● No Prorrogado ● Prorrogado

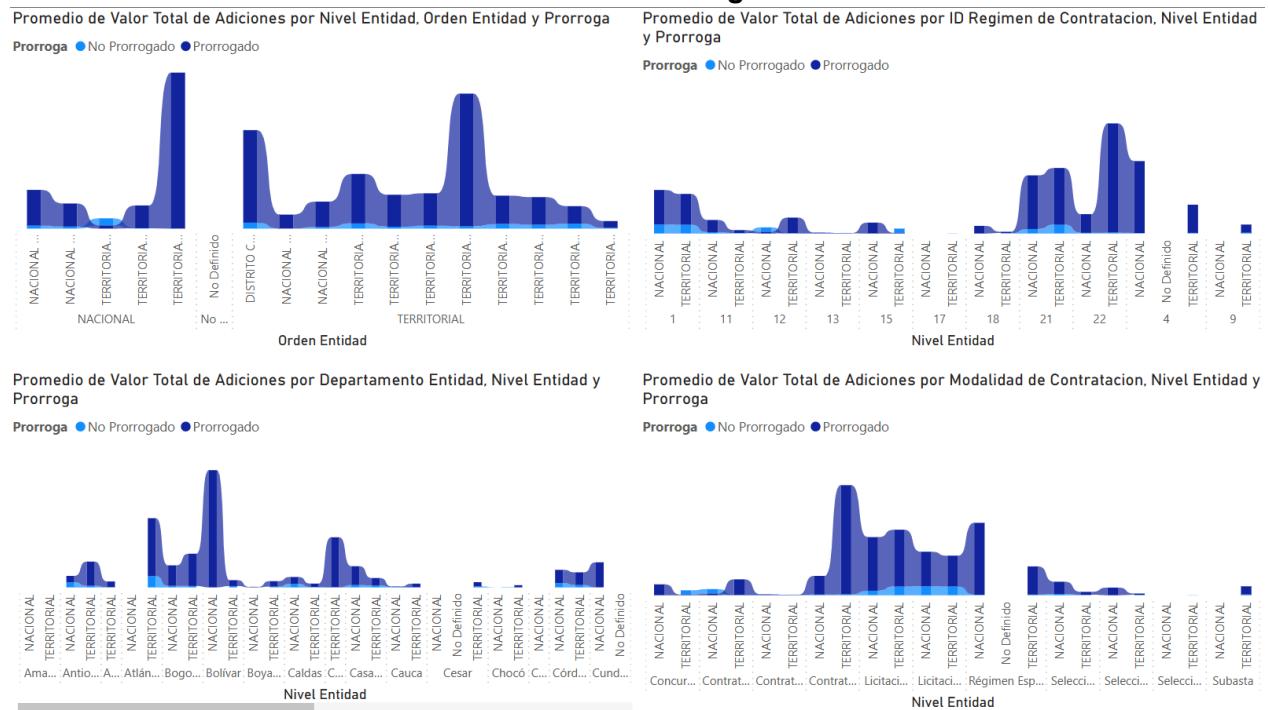


Mediana de Plazo de Ejec del Contrato por Modalidad de Contratacion, Nivel Entidad, Rango de Ejec del Contrato y Prorroga

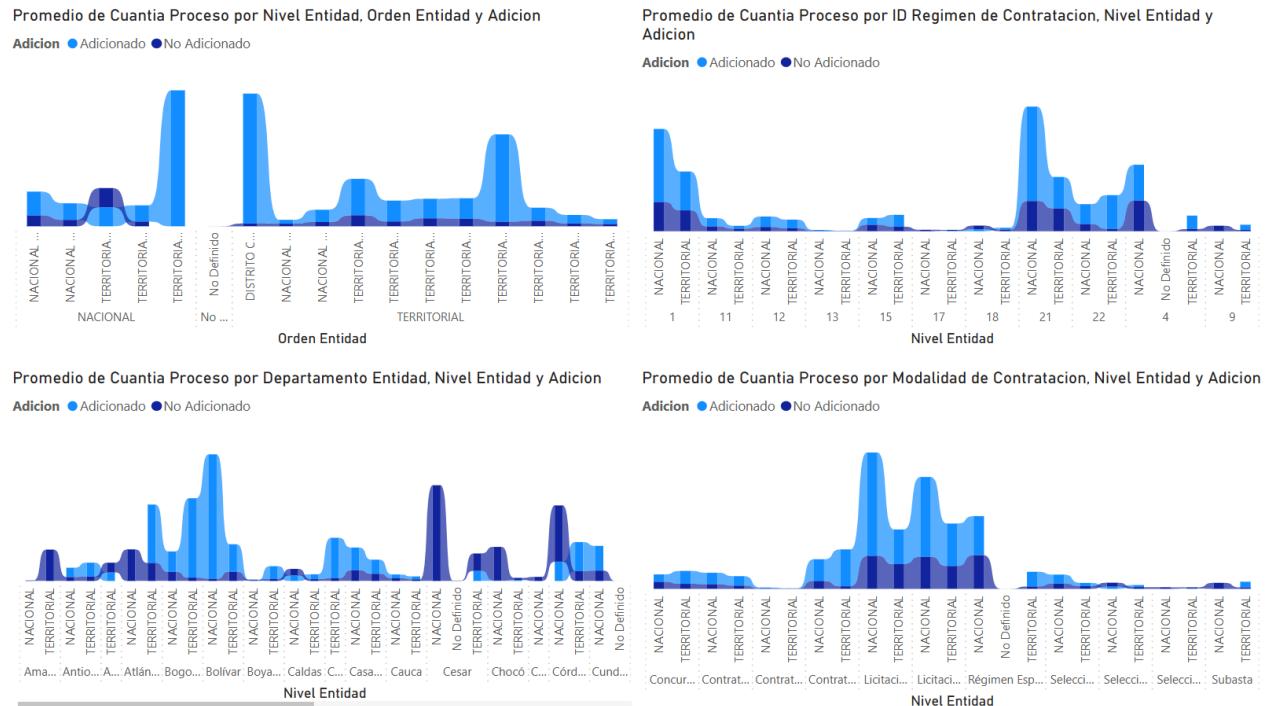
Prorroga • No Prorrogado • Prorrogado



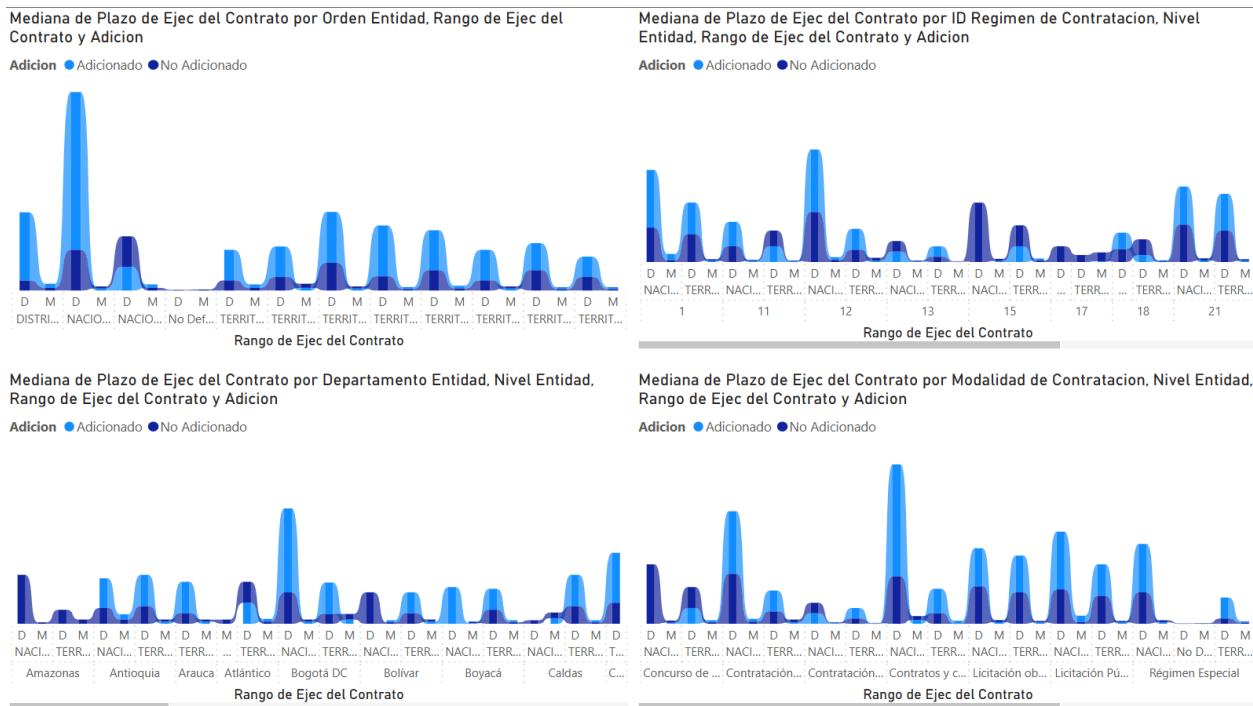
Gráfica: SECOP I - Prórroga - Adición



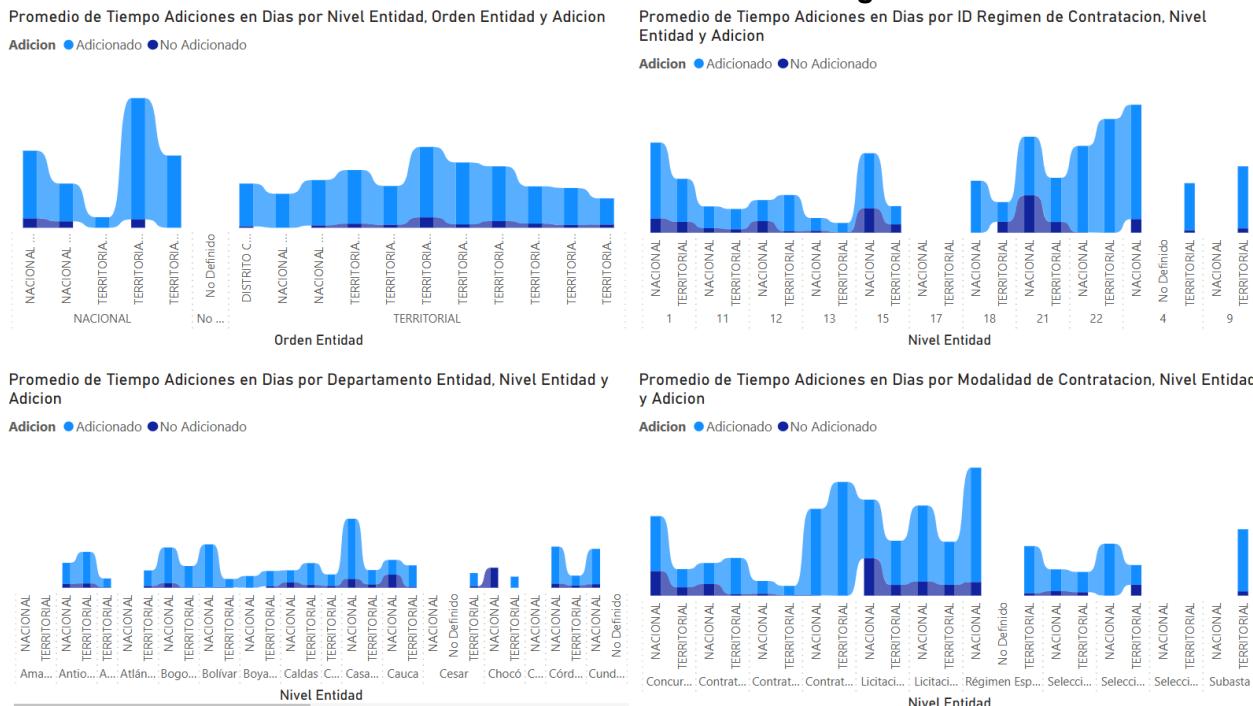
Gráfica: SECOP I - Adición - Cantidad



Gráfica: SECOP I - Adición - Plazo



Gráfica: SECOP I - Adición - Prórroga



11.1.3 Datos para predicción de estados finales

Imagen: Estados de finalización de los contratos SECOP I - Terminado Anormalmente después de Convocado

Fuente: (Colombia Compra Eficiente, 2020)

• Estados de los procesos de contratación



Terminado Anormalmente después de convocado

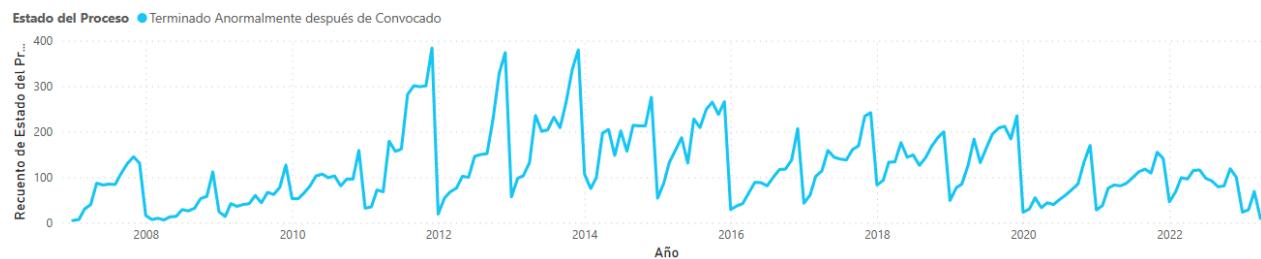


Es un estado de proceso que se utiliza para aquellos casos en que una vez abierta la convocatoria la entidad requiere terminar el proceso anormalmente, por ejemplo cuando debe revocar el acto administrativo que ordena la apertura o declararlo desierto.

Cuando un proceso es terminado anormalmente la entidad debe especificar la fecha y el motivo de terminación y anexar el acto administrativo de declaratoria de desierta o de terminación del proceso.

Gráfica: Estados de finalización de los contratos SECOP I - Mensual - (Terminado Anormalmente después de Convocado)

Recuento de Estado del Proceso por Año, Mes y Estado del Proceso

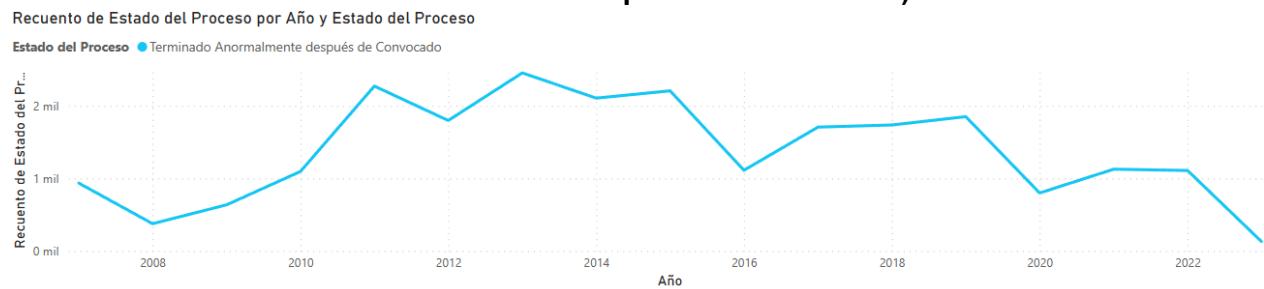


Recuento de Tiempo Adiciones en Días por Año, Mes y Estado del Proceso

Recuento de Tiempo Adi...
Recuento de Valor Total ...

Recuento de Valor Total de Adiciones por Año, Mes y Estado del Proceso

Gráfica: Estados de finalización de los contratos SECOP I - Anual - (Terminado Anormalmente después de Convocado)



Recuento de Tiempo Adiciones en Días por Año y Estado del Proceso

Recuento de Tiempo Adi...
Recuento de Valor Total ...

Recuento de Valor Total de Adiciones por Año y Estado del Proceso

Imágen: Estados de finalización de los contratos SECOP I - Liquidado
Fuente: (Colombia Compra Eficiente, 2020)

• Estados de los procesos de contratación



Liquidado

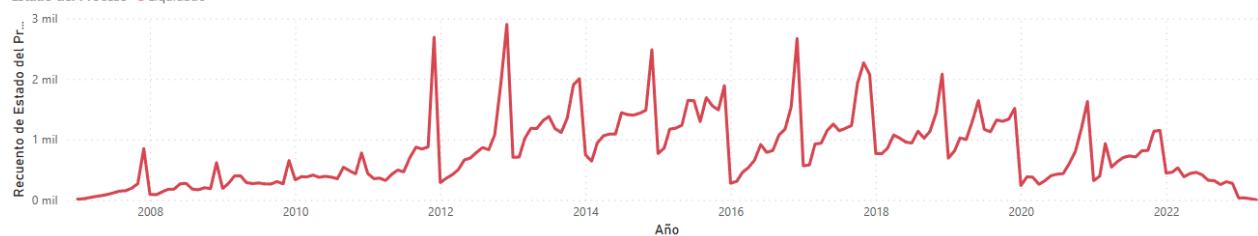


En el estado liquidado se deben reflejar los procesos de contratación que de acuerdo con la normatividad vigente requieran liquidación y hayan sido liquidados. En este estado se requiere la publicación del acta de liquidación o del acto administrativo de liquidación unilateral si es del caso.

Gráfica: Estados de finalización de los contratos SECOP I - Mensual - (Liquidado)

Recuento de Estado del Proceso por Año, Mes y Estado del Proceso

Estado del Proceso ● Liquidado



Recuento de Tiempo Adiciones en Días por Año, Mes y Estado del Proceso

Estado del Proceso ● Liquidado



Recuento de Valor Total de Adiciones por Año, Mes y Estado del Proceso

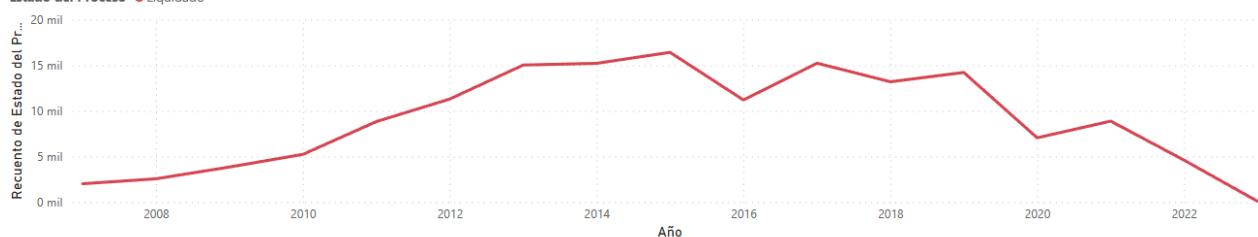
Estado del Proceso ● Liquidado



Gráfica: Estados de finalización de los contratos SECOP I - Anual - (Liquidado)

Recuento de Estado del Proceso por Año y Estado del Proceso

Estado del Proceso ● Liquidado



Recuento de Tiempo Adiciones en Días por Año y Estado del Proceso

Estado del Proceso ● Liquidado



Recuento de Valor Total de Adiciones por Año y Estado del Proceso

Estado del Proceso ● Liquidado



Imagen: Estados de finalización de los contratos SECOP I - Terminado sin Liquidar

Fuente: (Colombia Compra Eficiente, 2020)

• **Estados de los procesos de contratación**



Terminado sin liquidar



En este estado deberán registrarse los procesos de contratación que conforme con la naturaleza del contrato se hayan ejecutado y no requieran liquidación. Así también deberán registrarse los contratos que se hayan ejecutado y vencido el plazo legal para la liquidación los mismos no hayan sido liquidados, y aquellos procesos cuyos contratos hayan sido terminados anormalmente sin que hayan sido liquidados.

Gráfica: Estados de finalización de los contratos SECOP I - Mensual - (Terminado sin Liquidar)

Recuento de Estado del Proceso por Año, Mes y Estado del Proceso

Estado del Proceso ● Terminado sin Liquidar



Recuento de Tiempo Adiciones en Días por Año, Mes y Estado del Proceso

Estado del Proceso ● Terminado sin Liquidar



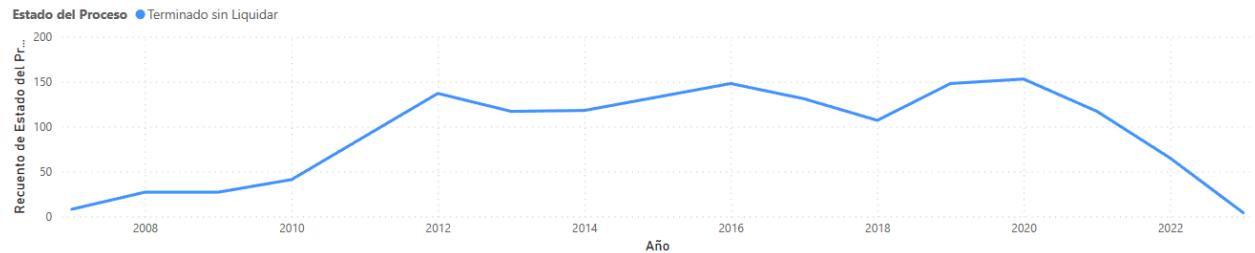
Recuento de Valor Total de Adiciones por Año, Mes y Estado del Proceso

Estado del Proceso ● Terminado sin Liquidar

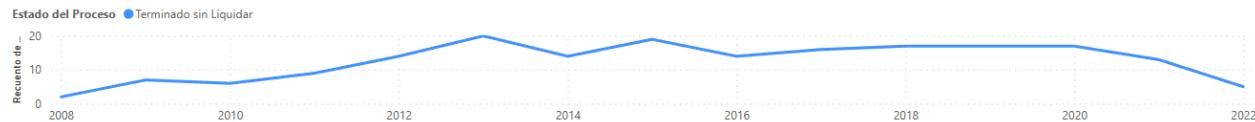


Gráfica: Estados de finalización de los contratos SECOP I - Anual - (Terminado sin Liquidar)

Recuento de Estado del Proceso por Año y Estado del Proceso



Recuento de Tiempo Adiciones en Dias por Año y Estado del Proceso

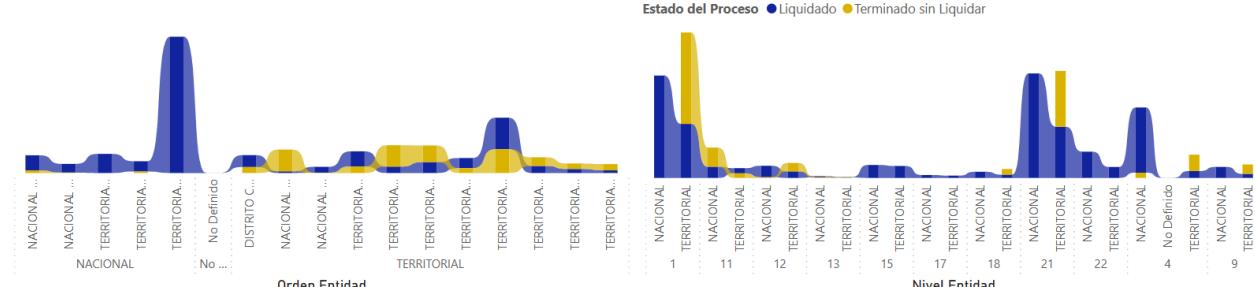


Recuento de Valor Total de Adiciones por Año y Estado del Proceso

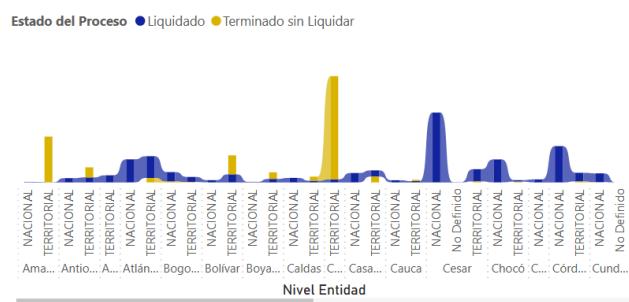


Gráfica: SECOP I - Estado finalizado - Cuantía

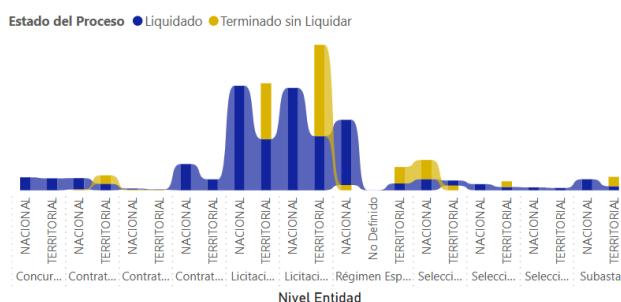
Promedio de Cuantia Proceso por Nivel Entidad, Orden Entidad y Estado del Proceso Promedio de Cuantia Proceso por ID Regimen de Contratacion, Nivel Entidad y Estado del Proceso



Promedio de Cuantia Proceso por Departamento Entidad, Nivel Entidad y Estado del Proceso



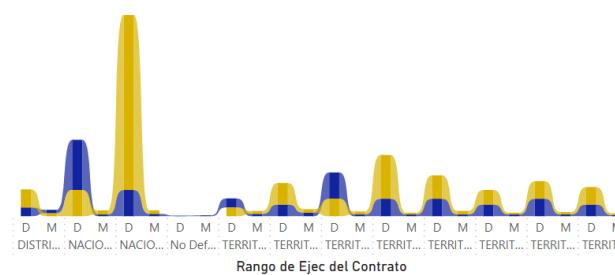
Promedio de Cuantia Proceso por Modalidad de Contratacion, Nivel Entidad y Estado del Proceso



Gráfica: SECOP I - Estado finalizado - Plazo

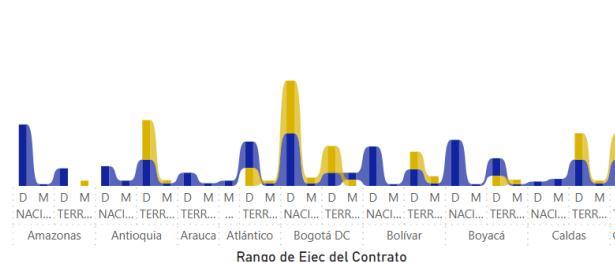
Mediana de Plazo de Ejec del Contrato por Orden Entidad, Rango de Ejec del Contrato y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



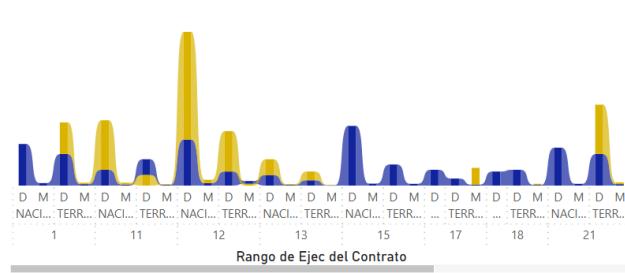
Mediana de Plazo de Ejec del Contrato por Departamento Entidad, Nivel Entidad, Rango de Ejec del Contrato y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



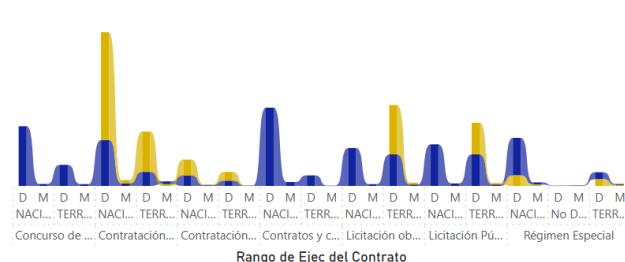
Mediana de Plazo de Ejec del Contrato por ID Regimen de Contratacion, Nivel Entidad, Rango de Ejec del Contrato y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



Mediana de Plazo de Ejec del Contrato por Modalidad de Contratacion, Nivel Entidad, Rango de Ejec del Contrato y Estado del Proceso

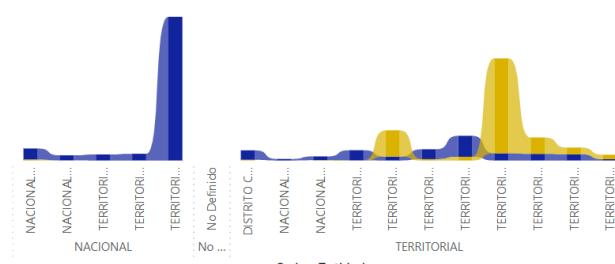
Estado del Proceso ● Liquidado ● Terminado sin Liquidar



Gráfica: SECOP I - Estado finalizado - Adición

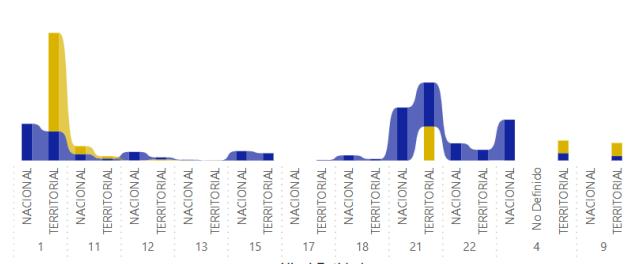
Promedio de Valor Total de Adiciones por Nivel Entidad, Orden Entidad y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



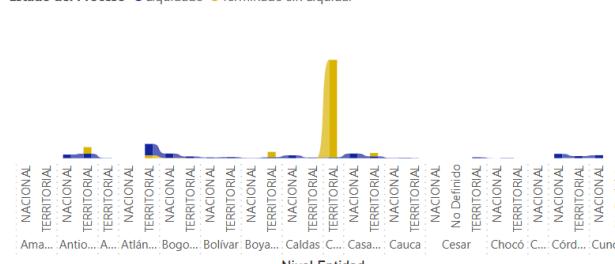
Promedio de Valor Total de Adiciones por ID Regimen de Contratacion, Nivel Entidad y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



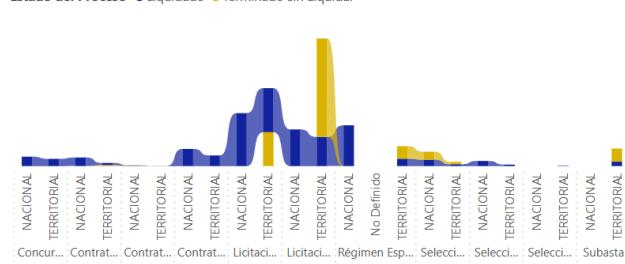
Promedio de Valor Total de Adiciones por Departamento Entidad, Nivel Entidad y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar

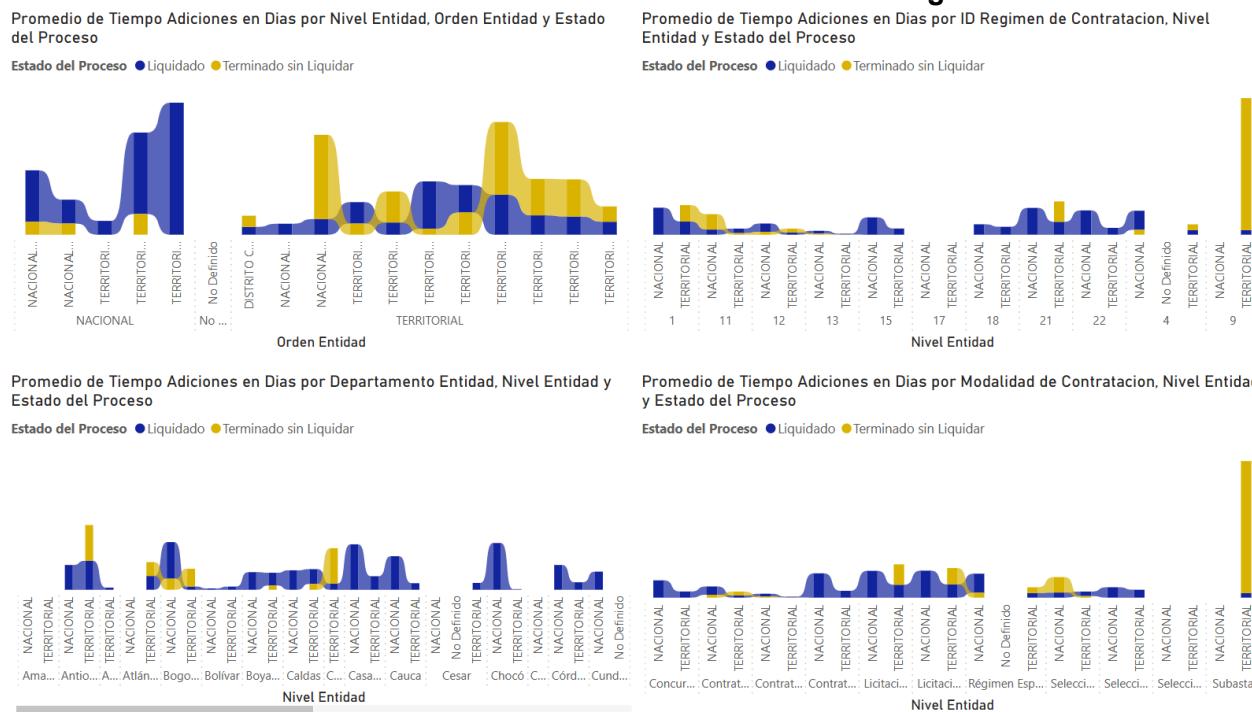


Promedio de Valor Total de Adiciones por Modalidad de Contratacion, Nivel Entidad y Estado del Proceso

Estado del Proceso ● Liquidado ● Terminado sin Liquidar



Gráfica: SECOP I - Estado finalizado - Prórroga



11.2 Introducción a los modelos

11.2.1 Modelos de clasificación lineal

LDA - Análisis discriminante lineal

El análisis discriminante lineal (LDA) es un método estadístico utilizado para tareas de clasificación. LDA hace suposiciones sobre la distribución de los datos para crear límites lineales entre las clases.

Los siguientes son los principales supuestos del modelo LDA cuando se utiliza para la clasificación:

- Suposición de normalidad: LDA asume que las variables predictoras se distribuyen normalmente dentro de cada clase. Esto significa que los datos deben seguir una curva en forma de campana, lo cual es importante para calcular la probabilidad de los datos dada la clase.
- Suposición de homocedasticidad: LDA también asume que la varianza de las variables predictoras es la misma para todas las clases. Esto significa que la distribución de los datos es la misma para todas las clases.
- Suposición de independencia: LDA asume que las variables predictoras son independientes entre sí dentro de cada clase. Esto significa que no debería haber correlación entre las variables predictoras dentro de cada clase.
- Suposición de matrices de covarianza iguales: LDA asume que la matriz de covarianza de las variables predictoras es la misma para todas las clases. Esto significa que la forma de los datos es la misma para todas las clases.

LDA puede proporcionar buenos resultados de clasificación incluso cuando los supuestos no se cumplen por completo, siempre que los datos no estén demasiado lejos de la normalidad o la homocedasticidad.

QDA - Análisis discriminante cuadrático

El análisis cuadrático discriminante (QDA) es un algoritmo de clasificación que se utiliza para clasificar las observaciones en una de varias clases predefinidas. Es similar al Análisis Discriminante Lineal (LDA), pero en QDA, la suposición de matriz de covarianza igual no se hace para todas las clases. Más bien, cada clase tiene su propia matriz de covarianza, lo que convierte a QDA en un modelo más flexible que LDA.

Cuando se usa QDA para la clasificación, hay algunas suposiciones que se deben cumplir:

- Suposición de normalidad: QDA asume que las variables independientes siguen una distribución normal multivariante dentro de cada clase.
- Suposición de homocedasticidad: se supone que la varianza de cada variable independiente es la misma en todas las clases.
- Suposición de independencia: se supone que cada observación es independiente de todas las demás observaciones.
- Tamaño de muestra adecuado: QDA asume que el número de observaciones es mayor que el número de variables independientes.

Si no se cumplen estos supuestos, el rendimiento del modelo QDA puede verse afectado. Por ejemplo, violar el supuesto de normalidad puede llevar a estimaciones sesgadas de los parámetros, mientras que violar el supuesto de homocedasticidad puede resultar en una estimación ineficiente del modelo.

Naive Bayes gaussiano

El algoritmo Gaussian Naive Bayes (GNB) es un modelo de clasificación que se basa en el teorema de Bayes. El modelo asume que los predictores o características son independientes entre sí, dadas las etiquetas de clase. En otras palabras, el modelo asume que cada característica contribuye por igual e independientemente a la probabilidad de una clase en particular.

Algunas de las suposiciones clave del modelo Gaussian Naive Bayes incluyen:

- Las características siguen una distribución normal (gaussiana): GNB asume que las características continuas en los datos siguen una distribución gaussiana. Si los datos no se distribuyen normalmente, el rendimiento del modelo puede ser subóptimo.
- Independencia condicional de clase: GNB asume que las características son condicionalmente independientes dadas las etiquetas de clase. Esto significa que el valor de una característica no influye en el valor de otra característica, dada la etiqueta de clase. Esta suposición no siempre es cierta en los datos del mundo real, pero el modelo sigue funcionando bien en muchos casos.
- Tamaño de muestra lo suficientemente grande: GNB funciona mejor con una gran cantidad de muestras de entrenamiento. La precisión del modelo puede ser pobre cuando el tamaño de la muestra es pequeño o cuando las clases están muy desequilibradas.
- Clases equilibradas: GNB funciona mejor cuando las clases están aproximadamente equilibradas. Si una clase tiene muchos más ejemplos que la otra, el modelo puede estar sesgado hacia la clase mayoritaria.

En general, el algoritmo Gaussian Naive Bayes es un modelo simple y eficiente que puede funcionar bien en muchos escenarios prácticos, especialmente cuando las características son independientes y siguen una distribución gaussiana.

11.2.2 Modelos lineales generalizados

Regresión Logística

La regresión logística es un algoritmo de aprendizaje automático ampliamente utilizado para problemas de clasificación binaria. Es un tipo de análisis de regresión donde la variable de respuesta es binaria (0 o 1) y el objetivo es predecir la probabilidad de la clase positiva dado un conjunto de características de entrada. El modelo de regresión logística tiene los siguientes supuestos:

- Linealidad: La relación entre las variables independientes y las probabilidades logarítmicas de la variable dependiente es lineal.
- Independencia: Las observaciones son independientes entre sí. No debe haber correlación entre los residuos.
- Ausencia de multicolinealidad: Las variables independientes no deben estar altamente correlacionadas entre sí.
- Normalidad: el término de error debe tener una distribución normal.
- Sin valores atípicos: no debe haber valores atípicos ni observaciones influyentes en los datos.
- Datos suficientes: debe haber suficientes datos para estimar los parámetros del modelo de manera confiable.
- Variable dependiente binaria: La variable dependiente debe ser binaria (0 o 1).

Las violaciones de estos supuestos pueden generar estimaciones sesgadas de los parámetros del modelo y predicciones inexactas. Por ejemplo, las violaciones de la suposición de linealidad pueden generar relaciones no lineales entre las variables independientes y la variable dependiente, lo que puede llevar a predicciones por encima o por debajo de la probabilidad de la clase positiva. De manera similar, las violaciones de la suposición de independencia pueden dar como resultado estimaciones sesgadas de los parámetros del modelo, lo que puede conducir a predicciones inexactas.

11.2.3 Modelos basados en particiones

Clasificador de árboles de decisión

Los árboles de decisión son un modelo popular para tareas de clasificación debido a su capacidad para manejar límites de decisión complejos e interacciones entre entidades. Estos son los supuestos clave de los modelos de árboles de decisión:

- Divisiones binarias: los árboles de decisión se basan en divisiones binarias en cada nodo, lo que significa que cada característica está por encima o por debajo de un cierto umbral. Esta suposición puede limitar la flexibilidad del modelo, ya que es posible que no pueda capturar relaciones más complejas entre características.
- Sin interacciones de características: los árboles de decisión asumen que las características interactúan de forma independiente para determinar la variable de destino. Sin embargo, en problemas del mundo real, las características pueden

interactuar de formas más complejas, lo que puede generar errores en las predicciones del modelo.

- Búsqueda codiciosa: los árboles de decisión utilizan un algoritmo de búsqueda codiciosa para determinar la mejor división en cada nodo, lo que significa que elige la solución óptima localmente en cada paso sin considerar el óptimo global. Esto puede conducir a divisiones subóptimas y a un árbol que no es tan efectivo como podría ser.
- Sobreajuste: los árboles de decisión son propensos al sobreajuste, especialmente cuando el árbol es profundo y complejo. Esto significa que el modelo puede ajustarse muy bien a los datos de entrenamiento, pero no generalizar a datos nuevos e invisibles.

Clasificador de Bagging

El clasificador de embolsado es un método de ensamble que combina varios clasificadores de base y se puede utilizar tanto para tareas de clasificación como de regresión. Asume que cada clasificador base tiene el mismo peso y se entrena de forma independiente utilizando un subconjunto de los datos originales. Estos son los supuestos del clasificador de embolsado:

- Clasificadores de base independientes: Se supone que los clasificadores de base utilizados en el embolsado son independientes entre sí. Esto significa que los errores cometidos por un clasificador no deberían influir en los errores cometidos por otros clasificadores.
- Diversidad de clasificadores básicos: para obtener el máximo beneficio del embolsado, los clasificadores básicos deben ser diversos en cuanto a las características que utilizan y los límites de decisión que aprenden. Esto se puede lograr mediante el uso de diferentes algoritmos o diferentes subconjuntos de datos para entrenar a cada clasificador.
- Clasificadores base imparciales: los clasificadores base deben ser imparciales, lo que significa que no deben sobreestimar o subestimar sistemáticamente las verdaderas probabilidades de clase. Los clasificadores sesgados pueden afectar negativamente el rendimiento del embolsado.
- Gran cantidad de clasificadores base: el rendimiento del ensacado generalmente mejora con la cantidad de clasificadores base utilizados. Sin embargo, hay un punto de rendimientos decrecientes en el que agregar más clasificadores no mejora significativamente el rendimiento pero aumenta el costo computacional.

En general, el clasificador de embolsado supone que la combinación de varios clasificadores básicos independientes, diversos, imparciales y de buen rendimiento puede mejorar la precisión y la solidez del modelo de clasificación final en comparación con el resultado de un clasificador de árboles de decisión.

Clasificador de Random Forest

Random Forest Classifier es un método de aprendizaje conjunto para tareas de clasificación, donde se construyen y agregan múltiples árboles de decisión para mejorar el rendimiento del modelo. Los siguientes son los supuestos del Random Forest Classifier:

- Independencia de los árboles: los árboles en el bosque deben ser independientes entre sí, lo que significa que cada árbol debe construirse sobre una muestra diferente y la selección de características también debe ser diferente para cada árbol.
- Los datos de entrenamiento deben ser representativos de la población: Los datos utilizados para entrenar el modelo deben ser representativos de la población en la que se pretende utilizar.

- Sin datos faltantes: Random Forests no puede manejar datos faltantes en las características.
- Sin valores atípicos: los bosques aleatorios son sensibles a los valores atípicos y pueden darles un peso indebido, lo que puede resultar en un ajuste excesivo.
- Las variables deben tener cierto poder predictivo: los bosques aleatorios que se entrena con variables predictoras que no son informativas o son redundantes pueden afectar negativamente el rendimiento del modelo.

Ventajas:

- Los bosques aleatorios tienen menos probabilidades de sobreajustarse que un modelo de árbol de decisión único.
- Pueden manejar una gran cantidad de características de entrada y no requieren escalado de características.
- Random Forests puede manejar valores perdidos y valores atípicos.

Desventajas:

- Pueden ser computacionalmente costosos y lentos de entrenar especialmente en conjuntos de datos grandes .
- No son fáciles de interpretar en comparación con otros modelos más de clasificación más simples..
- La salida del modelo puede estar sesgada hacia la clase mayoritaria en conjuntos de datos desbalanceados.

Clasificador de Gradient Boosting

Gradient Boosting Classifier es un popular algoritmo de aprendizaje automático utilizado para tareas de clasificación. Es un método de aprendizaje conjunto que crea una secuencia de árboles de decisión y luego los combina para producir un modelo más potente y preciso.

Los siguientes son los supuestos del clasificador de Gradeint Boosting:

- Variables independientes: el clasificador Gradeint Boosting asume que las variables predictoras son independientes entre sí. Si existe un alto grado de correlación entre las variables independientes, es posible que el modelo no funcione bien.
- Entradas numéricas: el clasificador de aumento de gradiente asume que las características de entrada son numéricas y tienen una relación lineal significativa con la variable de salida. Es posible que no funcione bien con características categóricas, y la codificación de características categóricas puede conducir a un espacio de características de alta dimensión.
- Valores atípicos: el clasificador de aumento de gradiente es sensible a los valores atípicos. Es importante identificar y manejar los valores atípicos en los datos de entrada para evitar el sobreajuste.
- Datos equilibrados: el clasificador de aumento de gradiente asume que los datos de entrada están equilibrados, es decir, el número de observaciones en cada clase es aproximadamente igual. Si los datos están muy desequilibrados, es posible que el modelo no pueda predecir la clase minoritaria con precisión.
- Hiperparámetros apropiados: Gradient Boosting Classifier tiene varios hiperparámetros que deben ajustarse adecuadamente para obtener un buen rendimiento. Estos hiperparámetros incluyen la tasa de aprendizaje, la cantidad de árboles, la profundidad máxima de los árboles y la cantidad mínima de muestras requeridas para dividir un nodo.

En resumen, Gradient Boosting Classifier asume entradas numéricas independientes con una relación lineal con la variable de salida, datos equilibrados y un ajuste de hiperparámetro adecuado. Es sensible a los valores atípicos y es posible que no funcione bien con características categóricas.

Clasificador de Adaboosting

El clasificador Adaboost (Adaptive Boosting) es un algoritmo de aprendizaje automático que se utiliza en tareas de clasificación. Es una técnica de aprendizaje en conjunto que combina múltiples clasificadores débiles para formar un clasificador fuerte. El algoritmo Adaboost tiene las siguientes suposiciones:

- Los datos utilizados en el conjunto de entrenamiento deben ser diversos: Adaboost funciona mejor cuando el conjunto de entrenamiento contiene una amplia gama de diferentes tipos de datos. El algoritmo crea múltiples clasificadores débiles que se entranen en diferentes subconjuntos de datos. Si los datos son demasiado similares, puede ser difícil para el algoritmo crear diversos clasificadores débiles.
- Los clasificadores débiles deben ser precisos: Adaboost está diseñado para funcionar con clasificadores débiles que funcionan mejor que al azar. Si los clasificadores débiles no son precisos, es posible que el algoritmo no pueda crear un clasificador fuerte.
- Los clasificadores débiles deben ser independientes: Adaboost asume que los clasificadores débiles son independientes entre sí. Esto significa que el error de un clasificador no debería afectar el error de otro clasificador.
- Las funciones utilizadas en los clasificadores débiles deben ser relevantes: Adaboost funciona mejor cuando las funciones utilizadas en los clasificadores débiles son relevantes para la tarea de clasificación. Las características irrelevantes pueden hacer que el algoritmo cree clasificadores débiles que no son útiles.

Los datos utilizados en el conjunto de entrenamiento deben estar balanceados: Adaboost asume que el conjunto de entrenamiento está balanceado, con el mismo número de muestras en cada clase. Si el conjunto de entrenamiento está desequilibrado, el algoritmo puede producir un clasificador que esté sesgado hacia la clase mayoritaria.

Clasificador XG Boosting

XGBoost (Extreme Gradient Boosting) es una extensión del aumento de gradiente que se ha vuelto cada vez más popular en los últimos años debido a su alto poder predictivo y eficiencia computacional. Al igual que el aumento de gradiente, XGBoost es un algoritmo de aprendizaje supervisado para clasificación y regresión que crea un conjunto de árboles de decisión de forma iterativa.

Hay varias suposiciones y requisitos a tener en cuenta al usar XGBoost:

- La variable dependiente debe ser categórica o numérica.
- Las variables independientes pueden ser numéricas o categóricas, pero las variables categóricas deben codificarse como valores numéricos.
- Los datos de entrada deben tener la forma de una matriz, donde cada fila corresponde a una observación y cada columna corresponde a una característica.
- XGBoost asume que no faltan valores en el conjunto de datos. Por lo tanto, es necesario preprocessar los datos e imputar los valores faltantes antes de ajustar el modelo.
- XGBoost asume que el conjunto de datos está libre de valores atípicos. Los valores atípicos pueden tener un impacto significativo en el rendimiento del modelo y se recomienda detectar y manejar los valores atípicos antes de entrenar el modelo.

- XGBoost funciona mejor cuando las funciones tienen una relación lineal o monótona con la variable dependiente. Por lo tanto, se recomienda preprocesar los datos y transformar las características si tienen una relación no lineal con la variable dependiente.
- XGBoost asume que los datos son independientes y están distribuidos de manera idéntica (i.i.d.) y que las observaciones no dependen del tiempo. Si los datos violan estas suposiciones, puede dar lugar a resultados sesgados o sobreajustados.

11.2.4 Modelos no paramétricos

K-Vecinos más cercanos

K-Nearest Neighbors (KNN) es un algoritmo no paramétrico y basado en instancias, lo que significa que no hace suposiciones sobre la distribución de datos subyacente. Por lo tanto, no asume ningún modelo específico o distribución de los datos.

Consideraciones que deben tenerse en cuenta al utilizar el clasificador KNN:

- KNN asume que los datos están normalizados o estandarizados. Por lo tanto, se recomienda normalizar los datos antes de aplicar KNN, para que todas las funciones contribuyan por igual al cálculo de la distancia.
- KNN asume que todas las características son igualmente importantes. Por lo tanto, es importante eliminar las funciones irrelevantes o realizar una selección de funciones antes de aplicar KNN.
- KNN asume que se debe elegir el valor óptimo de k para evitar el sobreajuste o el ajuste insuficiente. Por lo tanto, se recomienda utilizar técnicas de validación cruzada para encontrar el valor óptimo de k.
- KNN asume que la métrica de distancia utilizada para calcular la distancia entre dos puntos de datos es adecuada para los datos. Por lo tanto, es importante elegir una métrica de distancia adecuada según la naturaleza de los datos.

En resumen, KNN es un clasificador simple y versátil que no hace suposiciones sobre la distribución de datos subyacente, pero es importante normalizar los datos, eliminar características irrelevantes, elegir el valor óptimo de k y elegir una métrica de distancia apropiada para obtener mejores resultados.

11.2.5 Métricas de desempeño

- **RMSE:** RMSE mide la cantidad de error entre los valores predichos y los valores reales. Si los valores predichos son similares a los valores reales, entonces el valor de RMSE será bajo, lo que indica que el modelo de regresión es preciso en la predicción. Por otro lado, si los valores predichos son muy diferentes de los valores reales, entonces el valor de RMSE será alto, lo que indica que el modelo de regresión no es preciso en la predicción.

La fórmula que se utiliza para calcular el RMSE es la siguiente:

$$\text{RMSE} = \sqrt{(\sum(y_{\text{pred}} - y_{\text{real}})^2 / n)}$$

Donde:

- y_{pred} es el valor predicho por el modelo
- y_{real} es el valor real observado en los datos

- Σ es el símbolo de sumatoria
- n es el número de observaciones o datos

El RMSE se calcula tomando la diferencia entre los valores predichos y los valores reales al cuadrado, sumando estos valores y dividiendo entre el número de observaciones o datos. La raíz cuadrada de este valor se toma para obtener el valor final del RMSE, que se expresa en las mismas unidades que los valores reales. Una vez que se ha calculado el RMSE, se puede utilizar para evaluar la precisión de un modelo de regresión.

- **Curva ROC:** La curva ROC es una herramienta comúnmente usada para evaluar cuán bien un modelo de clasificación puede distinguir entre dos clases. La curva representa la tasa de verdaderos positivos (TPR) en el eje y y la tasa de falsos positivos (FPR) en el eje x para diferentes umbrales de clasificación. La curva ROC se genera al ajustar el umbral de clasificación y luego calcular el TPR y FPR correspondiente para cada umbral. Cuanto más cerca esté la curva ROC de la esquina superior izquierda, mejor será el modelo para distinguir entre las dos clases. Una curva ROC perfecta tendría un área bajo la curva (AUC) de 1, lo que indica que el modelo de clasificación es capaz de distinguir perfectamente entre las dos clases.

La fórmula para realizar el cálculo de la curva ROC:

$$TPR = TP / (TP + FN)$$

$$FPR = FP / (FP + TN)$$

Donde:

- TP (True Positive) es el número de casos en los que el modelo clasificó correctamente una instancia positiva.
- FN (False Negative) es el número de casos en los que el modelo clasificó incorrectamente una instancia positiva.
- FP (False Positive) es el número de casos en los que el modelo clasificó incorrectamente una instancia negativa.
- TN (True Negative) es el número de casos en los que el modelo clasificó correctamente una instancia negativa.

La curva ROC se construye variando el umbral de clasificación para cada observación en los datos y calculando TPR y FPR para cada umbral. Luego, se traza un gráfico de TPR vs. FPR para cada umbral. El área bajo la curva (AUC) de la curva ROC se puede calcular integrando la curva. Un AUC de 1 indica un modelo de clasificación perfecto, mientras que un AUC de 0.5 indica un modelo que es igualmente bueno en la clasificación que una elección aleatoria.

- **Matriz de confusión:** es una tabla que se usa para evaluar la precisión de un modelo de clasificación en una clasificación binaria. La tabla muestra la cantidad de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) que un modelo predijo. Los verdaderos positivos son los casos en los que el modelo predijo correctamente una clase positiva, los falsos positivos son los casos en los que el modelo predijo incorrectamente una clase positiva, los verdaderos negativos son los casos en los que el modelo predijo correctamente una clase negativa y los falsos negativos son los casos en los que el modelo predijo incorrectamente una clase

negativa. La precisión del modelo se puede calcular a partir de los valores de la matriz de confusión, lo que permite evaluar la calidad del modelo en la clasificación de datos.

La matriz de confusión se construye utilizando la clasificación del modelo en comparación con los valores reales. La matriz tiene una estructura de tabla 2x2, y sus elementos son los siguientes:

Tabla: Matriz de confusión

	Valores reales positivos	Valores reales negativos
Predicciones positivas	Verdaderos positivos (TP)	Falsos positivos (FP)
Predicciones negativas	Falsos negativos (FN)	Verdaderos negativos (TN)

La fórmula para calcular la precisión de un modelo a partir de una matriz de confusión es la siguiente:

$$\text{Precisión} = (TP + TN) / (TP + FP + TN + FN)$$

Donde:

- TP (True Positive) es el número de casos en los que el modelo clasificó correctamente una instancia positiva.
- FN (False Negative) es el número de casos en los que el modelo clasificó incorrectamente una instancia positiva.
- FP (False Positive) es el número de casos en los que el modelo clasificó incorrectamente una instancia negativa.
- TN (True Negative) es el número de casos en los que el modelo clasificó correctamente una instancia negativa.

La precisión se define como la proporción de casos clasificados correctamente por el modelo en comparación con el total de casos. Se puede calcular a partir de los valores de la matriz de confusión y se utiliza para evaluar la calidad de un modelo en la clasificación de datos. Además de la precisión, otras medidas comunes que se pueden calcular a partir de la matriz de confusión son la sensibilidad (TPR) y la especificidad (TNR). (James et al., 2013, #)

11.3 Modelo estados finales

11.3.1 Análisis de resultados

Tabla: Resultados de las métricas para parámetros por defecto (dummie)

Modelo	Estrategia	0			1			accuracy	prom
		precision	recall	f1	precision	recall	f1		
LDA	Ninguna	0,99	0,99	0,99	0,20	0,24	0,22	0,98	0,61
	Subsampling en la clase mayoritaria	1,00	0,13	0,23	0,01	0,95	0,02	0,14	0,39
	Oversampling de la clase minoritaria	1,00	0,85	0,92	0,04	0,65	0,07	0,85	0,59
	Combinamos resampling con Smote-Tomek	0,99	0,98	0,99	0,14	0,26	0,18	0,98	0,59

		0			1				
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
QDA	Ninguna	1,00	0,34	0,51	0,01	0,84	0,02	0,35	0,45
	Subsampling en la clase mayoritaria	0,99	0,31	0,47	0,01	0,66	0,02	0,31	0,41
	Oversampling de la clase minoritaria	1,00	0,75	0,86	0,03	0,70	0,05	0,75	0,57
	Combinamos resampling con Smote-Tomek	0,99	1,00	0,99	0,13	0,05	0,08	0,99	0,54
	Ensamble de Modelos con Balanceo	0,99	0,13	0,23	0,01	0,91	0,02	0,14	0,38
Naive Bayes	Ninguna	0,99	1,00	1,00	0,00	0,00	0,00	0,99	0,50
	Subsampling en la clase mayoritaria	0,99	0,02	0,04	0,01	0,98	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	0,03	0,06	0,01	0,98	0,02	0,04	0,35
	Combinamos resampling con Smote-Tomek	0,99	0,03	0,06	0,01	0,97	0,02	0,04	0,35
	Ensamble de Modelos con Balanceo	0,99	0,90	0,94	0,01	0,10	0,02	0,89	0,49
Regresión Logistica	Ninguna				0,00	0,00	0,00	0,99	0,50
Decision Tree Classifier	Ninguna	0,99	0,99	0,99	0,18	0,21	0,20	0,98	0,59
	Penalización para compensar	0,99	0,99	0,99	0,11	0,18	0,13	0,98	0,57
	Subsampling en la clase mayoritaria	0,99	0,01	0,02	0,01	0,99	0,02	0,02	0,34
	Oversampling de la clase minoritaria	0,99	0,99	0,99	0,12	0,18	0,14	0,98	0,57
	Combinamos resampling con Smote-Tomek	0,99	0,99	0,99	0,16	0,19	0,17	0,98	0,58
	Ensamble de Modelos con Balanceo	1,00	0,86	0,92	0,04	0,58	0,07	0,86	0,58
Random Forest	Ninguna	0,99	1,00	1,00	0,78	0,08	0,14	0,99	0,67
	Penalización para compensar	0,99	1,00	1,00	0,88	0,08	0,14	0,99	0,68
	Subsampling en la clase mayoritaria	0,99	0,02	0,04	0,01	0,98	0,02	0,03	0,34
	Oversampling de la clase minoritaria	0,99	1,00	1,00	0,75	0,10	0,17	0,99	0,67
	Combinamos resampling con Smote-Tomek	0,99	1,00	1,00	0,82	0,10	0,18	0,99	0,68
	Ensamble de Modelos con Balanceo	1,00	0,86	0,92	0,04	0,58	0,07	0,85	0,58
AdaBoost	Ninguna	0,99	1,00	1,00	0,50	0,02	0,04	0,99	0,59
	Subsampling en la clase mayoritaria	0,99	0,01	0,03	0,01	0,98	0,02	0,02	0,34
	Oversampling de la clase minoritaria	1,00	0,81	0,89	0,03	0,64	0,06	0,81	0,57
	Combinamos resampling con Smote-Tomek	0,99	0,97	0,98	0,05	0,19	0,08	0,96	0,54
	Ensamble de Modelos con Balanceo	1,00	0,78	0,87	0,03	0,64	0,05	0,78	0,56
Gradient Boosting	Ninguna	0,99	1,00	0,99	0,24	0,13	0,17	0,99	0,59
	Subsampling en la clase mayoritaria	0,99	0,01	0,03	0,01	0,99	0,02	0,02	0,34
	Oversampling de la clase minoritaria	1,00	0,87	0,93	0,04	0,58	0,07	0,86	0,58
	Combinamos resampling con Smote-Tomek	0,99	0,99	0,99	0,12	0,15	0,13	0,98	0,56
	Ensamble de Modelos con Balanceo	1,00	0,80	0,89	0,03	0,63	0,05	0,80	0,57
K Vecinos	Ninguna	0,99	0,99	0,99	0,02	0,02	0,02	0,98	0,51

11.4 Modelo adiciones

11.4.1 Análisis de resultados

Tabla: Resultados de las métricas para parámetros por defecto (dummie)

Modelo	Estrategia	0			1			accuracy	prom
		precision	recall	f1	precision	recall	f1		
LDA	Ninguna	0,92	0,96	0,94	0,39	0,24	0,30	0,89	0,63
	Subsampling en la clase mayoritaria	0,94	0,43	0,59	0,13	0,77	0,22	0,46	0,51
	Oversampling de la clase minoritaria	0,97	0,75	0,84	0,25	0,76	0,37	0,75	0,66
	Combinamos resampling con Smote-Tomek	0,91	0,99	0,95	0,47	0,09	0,15	0,90	0,59
QDA	Ninguna	0,92	0,80	0,86	0,16	0,35	0,22	0,76	0,55
	Subsampling en la clase mayoritaria	0,96	0,40	0,56	0,13	0,84	0,23	0,44	0,52
	Oversampling de la clase minoritaria	0,90	0,99	0,95	0,20	0,02	0,03	0,90	0,52
	Combinamos resampling con Smote-Tomek	0,98	0,31	0,47	0,13	0,93	0,23	0,37	0,51
	Ensamble de Modelos con Balanceo	0,93	0,75	0,83	0,18	0,51	0,27	0,72	0,58
Naive Bayes	Ninguna	0,90	1,00	0,95	0,33	0,02	0,04	0,90	0,54
	Subsampling en la clase mayoritaria	0,90	1,00	0,95	0,00	0,00	0,00	0,90	0,48
	Oversampling de la clase minoritaria	0,90	0,99	0,95	0,35	0,04	0,06	0,90	0,55
	Combinamos resampling con Smote-Tomek	0,90	0,99	0,95	0,34	0,04	0,07	0,90	0,55
	Ensamble de Modelos con Balanceo	0,90	0,99	0,94	0,30	0,06	0,09	0,89	0,55
Regresión Logística	Ninguna	0,90	1,00	0,95	0,06	0,00	0,00	0,90	0,49
Decision Tree Classifier	Ninguna	0,92	0,94	0,93	0,35	0,29	0,32	0,88	0,63
	Penalización para compensar	0,93	0,90	0,91	0,29	0,36	0,32	0,85	0,62
	Subsampling en la clase mayoritaria	0,85	0,10	0,18	0,09	0,84	0,17	0,18	0,37
	Oversampling de la clase minoritaria	0,93	0,91	0,92	0,30	0,36	0,33	0,85	0,63
	Combinamos resampling con Smote-Tomek	0,92	0,91	0,92	0,27	0,29	0,28	0,85	0,60
	Ensamble de Modelos con Balanceo	0,96	0,80	0,87	0,28	0,70	0,40	0,79	0,67
Random Forest	Ninguna	0,91	0,99	0,95	0,57	0,12	0,20	0,90	0,62
	Penalización para compensar	0,91	0,99	0,95	0,61	0,14	0,23	0,91	0,64
	Subsampling en la clase mayoritaria	0,93	0,27	0,42	0,11	0,82	0,19	0,33	0,46
	Oversampling de la clase minoritaria	0,92	0,97	0,95	0,48	0,26	0,34	0,90	0,65
	Combinamos resampling con Smote-Tomek	0,92	0,97	0,94	0,46	0,26	0,33	0,90	0,65
	Ensamble de Modelos con Balanceo	0,97	0,76	0,86	0,27	0,80	0,40	0,77	0,68
AdaBoost	Ninguna	0,91	0,99	0,95	0,44	0,07	0,11	0,90	0,58
	Subsampling en la clase mayoritaria	0,93	0,28	0,42	0,11	0,80	0,19	0,33	0,46
	Oversampling de la clase minoritaria	0,97	0,73	0,83	0,24	0,80	0,37	0,73	0,66
	Combinamos resampling con Smote-Tomek	0,93	0,90	0,91	0,29	0,38	0,33	0,85	0,62
	Ensamble de Modelos con Balanceo	0,97	0,73	0,84	0,25	0,81	0,38	0,74	0,66

		0			1				
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
Gradient Boosting	Ninguna	0,91	1,00	0,95	0,66	0,06	0,11	0,90	0,62
	Subsampling en la clase mayoritaria	0,83	0,12	0,21	0,09	0,78	0,16	0,18	0,37
	Oversampling de la clase minoritaria	0,97	0,71	0,82	0,24	0,82	0,37	0,72	0,66
	Combinamos resampling con Smote-Tomek	0,93	0,92	0,93	0,34	0,38	0,36	0,87	0,64
	Ensamble de Modelos con Balanceo	0,97	0,70	0,82	0,24	0,83	0,37	0,72	0,66
K Vecinos	Ninguna	0,91	0,92	0,91	0,22	0,22	0,22	0,84	0,61

11.5 Modelo prorrrogas

11.5.1 Análisis de resultados

Tabla: Resultados de las métricas para parámetros por defecto (dummie)

		0			1				
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
LDA	Ninguna	0,92	0,95	0,94	0,47	0,34	0,40	0,89	0,67
	Subsampling en la clase mayoritaria	0,94	0,47	0,63	0,15	0,76	0,25	0,50	0,53
	Oversampling de la clase minoritaria	0,97	0,77	0,86	0,29	0,80	0,43	0,77	0,69
	Combinamos resampling con Smote-Tomek	0,91	0,98	0,94	0,48	0,17	0,25	0,89	0,62
	Ensamble de Modelos con Balanceo	0,94	0,80	0,86	0,25	0,57	0,35	0,78	0,63
QDA	Ninguna	0,89	1,00	0,94	0,33	0,00	0,00	0,89	0,53
	Subsampling en la clase mayoritaria	0,99	0,29	0,44	0,14	0,96	0,24	0,36	0,51
	Oversampling de la clase minoritaria	0,89	1,00	0,94	0,33	0,00	0,00	0,89	0,53
	Combinamos resampling con Smote-Tomek	0,95	0,42	0,58	0,14	0,80	0,24	0,46	0,52
	Ensamble de Modelos con Balanceo	0,94	0,80	0,86	0,25	0,57	0,35	0,78	0,63
Naive Bayes	Ninguna	0,90	1,00	0,94	0,59	0,05	0,10	0,89	0,60
	Subsampling en la clase mayoritaria	0,89	1,00	0,94	0,53	0,02	0,04	0,89	0,57
	Oversampling de la clase minoritaria	0,90	0,99	0,94	0,58	0,08	0,13	0,89	0,60
	Combinamos resampling con Smote-Tomek	0,90	0,99	0,94	0,54	0,10	0,16	0,89	0,61
	Ensamble de Modelos con Balanceo	0,90	0,99	0,94	0,52	0,09	0,16	0,89	0,60
Regresión Logistica	Ninguna	0,89	1,00	0,94	0,00	0,00	0,00	0,89	0,47
Decision Tree Classifier	Ninguna	0,92	0,93	0,93	0,38	0,37	0,38	0,87	0,65
	Penalización para compensar	0,93	0,91	0,92	0,36	0,44	0,40	0,86	0,66
	Subsampling en la clase mayoritaria	0,77	0,10	0,18	0,09	0,75	0,16	0,17	0,34
	Oversampling de la clase minoritaria	0,93	0,91	0,92	0,36	0,45	0,40	0,86	0,66
	Combinamos resampling con Smote-Tomek	0,93	0,91	0,92	0,37	0,42	0,39	0,86	0,66
	Ensamble de Modelos con Balanceo	0,96	0,82	0,89	0,33	0,74	0,46	0,81	0,70
Random Forest	Ninguna	0,92	0,98	0,95	0,65	0,26	0,38	0,91	0,69
	Penalización para compensar	0,92	0,98	0,95	0,64	0,27	0,38	0,91	0,69

		0			1				
Modelo	Estrategia	precision	recall	f1	precision	recall	f1	accuracy	prom
	Subsampling en la clase mayoritaria	0,90	0,26	0,41	0,11	0,76	0,19	0,32	0,44
	Oversampling de la clase minoritaria	0,93	0,95	0,94	0,53	0,42	0,47	0,90	0,71
	Combinamos resampling con Smote-Tomek	0,93	0,96	0,94	0,54	0,44	0,48	0,90	0,72
	Ensamble de Modelos con Balanceo	0,97	0,79	0,88	0,33	0,83	0,47	0,80	0,71
AdaBoost	Ninguna	0,91	0,98	0,95	0,63	0,23	0,34	0,90	0,67
	Subsampling en la clase mayoritaria	0,89	0,27	0,42	0,11	0,72	0,19	0,32	0,43
	Oversampling de la clase minoritaria	0,98	0,75	0,85	0,29	0,85	0,43	0,76	0,69
	Combinamos resampling con Smote-Tomek	0,94	0,90	0,92	0,37	0,48	0,42	0,85	0,67
	Ensamble de Modelos con Balanceo	0,97	0,75	0,85	0,29	0,84	0,43	0,76	0,69
Gradient Boosting	Ninguna	0,91	0,99	0,95	0,69	0,19	0,30	0,90	0,67
	Subsampling en la clase mayoritaria	0,76	0,11	0,20	0,09	0,70	0,15	0,18	0,34
	Oversampling de la clase minoritaria	0,98	0,75	0,85	0,29	0,85	0,43	0,76	0,69
	Combinamos resampling con Smote-Tomek	0,94	0,90	0,92	0,40	0,54	0,46	0,86	0,69
	Ensamble de Modelos con Balanceo	0,98	0,74	0,84	0,29	0,86	0,43	0,75	0,69
K Vecinos	Ninguna	0,91	0,91	0,91	0,26	0,26	0,26	0,84	0,62