Resumen ejecutivo del *Modelo Probabilístico del Éxito de las Demandas en Contra del Estado* de QUANTIL.

Este estudio presenta dos modelos probabilísticos del éxito de la Nación en procesos judiciales en su contra. El objetivo de los modelos es brindar a la Agencia Nacional de Defensa Jurídica del Estado y a las entidades públicas del orden nacional una herramienta que, a través del cálculo de una probabilidad confiable de éxito, les permita priorizar los procesos en los que desean intervenir y definir estrategias de defensa y actuación procesal que redunden en mejores resultados para el Estado.

Los resultados fueron obtenidos con dos metodologías que buscan encontrar patrones en grandes volúmenes de datos. Las dos metodologías fueron aplicadas a dos bases de datos, una con todas las variables del proceso que serían conocidas al finalizar el mismo y otra que incluye solo las variables que se conocen a su inicio. La primera metodología aplicada es la de regresión logística, cuyos resultados se pueden encontrar en el archivo Excel que acompaña este estudio. Los resultados de la segunda metodología, boosting de árboles, se encuentran en interfaz del programa estadístico de libre acceso "R" cuyo manejo se explica en la página 45 del documento. Para realizar estos análisis se utilizó la base de datos de un censo de casos activos en la jurisdicción contencioso administrativa en 2012 a nivel nacional realizado por el Banco Mundial. Este censo incluye información de más de 150.000 casos activos, de estos, cerca de 16.000 tenían información para el fallo en primera instancia.

A su vez, se encontró que algunas características de los procesos favorecen el éxito procesal de la Nación. Entre más se prolonga un proceso, es más probable que el Estado lo gane; presentar apelación al fallo de primera instancia, aumenta la probabilidad de éxito para la Nación; las demandas interpuestas en Pereira, Villavicencio o Tunja tienen con mayor frecuencia un fallo favorable al Estado; algunos jueces y magistrados tienden a fallar a favor del Estado; las demandas de reparación directa por falla en la prestación del servicio público, las de carácter salarial, retiro del servicio y contra sanciones de órganos de inspección y vigilancia son falladas a favor del Estado con mayor probabilidad; el cambio de abogado de la entidad demandada tiene un impacto a favor del éxito del Estado, haciendo que sus probabilidades de ganar aumenten; en la medida en que el monto pretendido por el demandante sea mayor, se incrementa la probabilidad de éxito de la Nación, aunque el impacto no es significativo en la probabilidad de éxito. Por último, la presentación de un mayor número de pruebas se relaciona con una menor probabilidad de que se conceda la pretensión al demandante.

Por otro lado, se encontraron variables que se relacionan con un mayor riesgo de obtener fallos desfavorables para el Estado. La presentación de los alegatos de conclusión por cualquiera de las partes, aumenta la probabilidad de éxito del demandante; el aumento en el número de demandantes, cuando estos son personas naturales, o si es una persona jurídica tiene un impacto negativo en el éxito de la Nación;

las demandas interpuestas en Bucaramanga, Ibagué o Santa Marta presentan más fallos desfavorables al Estado; cuando el fallo de primera instancia se produce en un juzgado disminuye la probabilidad de éxito de la Nación; los fallos de ciertos jueces y magistrados generan con mayor frecuencia condenas contra el Estado. Finalmente, las demandas de reparación directa por daños generados por la fuerza pública, o sufridos por conscriptos, las acciones de nulidad y restablecimiento de tipo tributario y las laborales, sobre todo las de tipo pensional y prestacional, tienen mayor probabilidad de que el Estado pierda.

Además de las variables procesales anteriormente presentadas, la entidad que es demandada juega un papel considerable en la probabilidad de éxito que tenga la Nación. La base de datos analizada incluye demandas en contra de sesenta entidades, Al analizar el efecto que la entidad tiene sobre el éxito del proceso, se encontró que ocho entidades afectan en algún sentido la probabilidad de éxito de la Nación. Las entidades que tienen una mayor probabilidad de éxito son la Superintendencia de Servicios Públicos Domiciliarios, el Ministerio de la Protección Social y la Caja de Sueldos de Retiro de la Policía Nacional, mientras que las entidades con una menor probabilidad de éxito son CAJANAL – en liquidación, la Fiscalía, CREMIL, el Ejército Nacional y la Fiduprevisora S.A. Esta última es la que tiene la menor tasa de éxito. Junto a esto se encontró que el 40% de los procesos del INPEC y el 38% de los procesos del Ministerio de Transporte fueron fallados en contra de las entidades y que para los Ministerios de Educación y Agricultura el número de procesos es muy bajo. Por esta razón no fue posible extraer información concluyente sobre estas dos entidades. Por último, se encontró que existen quince abogados que tienen efecto en la probabilidad de éxito del Estado al fungir como apoderados de los demandantes, siete de estos abogados afectan la probabilidad en favor del Estado mientras ocho se relacionan con una mayor probabilidad de que la Nación pierda el proceso.

En el modelo de éxito en que se incluyen sólo las variables conocidas al comienzo del proceso, los resultados son similares, aquellos diferentes al modelo con todas las variables se presentan a continuación. Con respecto a las entidades que resultan relevantes, aparecen las mismas que en el modelo anterior excepto por el Consejo Superior de la Judicatura, que pierde las demandas por encima del promedio; las ciudades Popayán y Santa Rosa de Viterbo inciden perjudicialmente en el éxito del Estado. Entre las razones para demandar, los daños causados por la fuerza pública y por y a conscriptos dejan de ser características determinantes del éxito, mientras que las demandas de tipo laboral continúan teniendo una mayor probabilidad de pérdida para el Estado. Por último, también se presentan cambios en los jueces o Magistrados que tienden a condenar y exonerar al Estado de forma sistemática.

quantil

Matemáticas Aplicadas

INFORME FINAL

Modelo Probabilístico del Éxito de las Demandas en Contra del Estado

PRESENTADO A:

Agencia Nacional de Defensa Jurídica del Estado

Diciembre 2013

Los datos e información usados por Quantil en el desarrollo y presentación de este reporte han sido obtenidos o derivados de fuentes que Quantil considera confiables, pero Quantil no se responsabiliza por su completitud y precisión.

1. Introducción

El pago anual por sentencias en contra del Estado subió de forma consistente de 400 mil millones en 2008 a 800 mil millones en 2011. Con la expectativa de que este crecimiento seguirá esta tendencia, la proyección de los pagos esperados en años futuros reviste una alta importancia desde un punto de vista fiscal, convirtiéndose en un rubro relevante para la elaboración del Presupuesto General de la Nación. Más aún, una valoración de este pasivo contingente es un cálculo valioso para cualquier análisis financiero del costo que para el Estado tienen las demandas en su contra.

Por otro lado, considerando que el número agregado de procesos abiertos en contra del Estado supera los doscientos mil, se vuelve imperativo para la Agencia Nacional de Defensa Jurídica del Estado (en adelante, ANDJE) contar con los datos y herramientas necesarios para poder otorgar una probabilidad confiable del éxito de la defensa de distintos procesos, de forma que se apoye el análisis de priorización de la implementación de estas defensas.

Este documento presenta el informe final del análisis de características de variables en procesos en contra del Estado y del desarrollo de un modelo probabilístico del éxito de la defensa del Estado. El objetivo es identificar las demandas en contra del Estado que tienen una alta probabilidad de éxito (para el Estado), utilizando la información histórica de los casos archivados. Partiendo de la premisa técnica que dificulta juzgar un modelo como bueno o malo, y guía el enfoque de evaluación hacia la comparación con otros modelos, a lo largo de este trabajo se han desarrollado varios modelos que cuantifican la probabilidad de éxito buscada: regresión logística, redes neuronales, *boosting* de árboles, y varias familias que usan técnicas de análisis semi-supervisado. Para cada paradigma se identifican las variables relevantes, y se implementan soluciones de reducción de dimensionalidad consistente con estos conjuntos. Como resultado final, para cada caso se establece una función que toma las características de una demanda en contra del Estado (en efecto, el valor de todas las variables descriptivas), y calcula la probabilidad de éxito.

A lo largo del trabajo se distinguen dos tipos de análisis: el primero toma las características conocidas en el momento en que arranca un proceso, y el segundo incluye las variables que se van conociendo en la medida en que avanza el proceso (por ejemplo, la duración del proceso, o la presentación de alegatos). Aunque los datos históricos incorporan toda la información, un usuario de la ANDJE en ocasiones solo contará con la información parcial que se conoce al principio, razón por la cual se analizan las dos situaciones.

En los dos casos, el modelo que mejor permite ajustar los datos observados es el de *boosting* de árboles. Para el caso en que se usan todas las variables, tomando 70 % de los datos para calibrar el modelo, y evaluándolo sobre el 30 % de datos fuera de muestra, el área bajo la curva ROC es de 76 %, comparado con 74 % para redes neuronales y 74 % para regresión logística, que son los que le siguen en bondad. En el segundo caso, en que solo se toman los datos conocidos al principio de los

casos, los números de comparación son 75 % para *boosting* de árboles, 74 % para redes neuronales y 72 % para regresión logística. En el capítulo cinco se elabora más profundamente sobre las definiciones e interpretaciones de estos números.

Como resultado adicional, este trabajo exhibe las variables estadísticamente relevantes para el éxito procesal de la Nación a partir del uso de las herramientas desarrolladas, con lo cual se permite segmentar el universo de casos según las características que sean predictivas del éxito. Particularmente, algunas variables se encuentran frecuentemente con alta relevancia de predicción, independientemente del modelo usado. Entre ellas se pueden resaltar entidades demandadas, participación en el proceso de determinados abogados del demandante o magistrados, motivos de la pretensión y ubicación geográfica de ésta. En los capítulos cuatro y cinco se profundiza en esta lista y en la descripción de su incidencia en la probabilidad de éxito.

Naturalmente, existen varias formas de comparar la bondad de los modelos implementados; en el presente trabajo se ha optado por usar el área bajo la curva ROC, que es la gráfica que compara la tasa de verdaderos positivos contra la de falsos positivos para todo el continuo de posibilidades, variando el nivel de confianza tolerado. Intuitivamente, la curva ROC exhibe, para un modelo dado, el número falsos positivos que deben tolerarse para obtener un número determinado de verdaderos positivos, donde los falsos positivos son aquellos procesos que el modelo marca como pertenecientes a una clase de manera errónea, mientras los verdaderos positivos son aquéllos que en efecto pertenecen a la clase asignada por el modelo. Teniendo en cuenta que el método de clasificación asigna un valor entre 0 y 1 a cada proceso, el área bajo esta curva suele interpretarse como "la probabilidad de asignar una alta probabilidad a los casos pertenecientes a la clase representada por 1", que en este caso corresponden a aquéllos en los que se otorga la pretensión.

En adición, en este trabajo se realiza un análisis de reglas de asociación para identificar relaciones entre variables que pueden escapar revisiones preliminares de los datos. Este análisis tiende a ser útil en varios sentidos: permite priorizar situaciones que pueden ameritar investigación, resalta valores particulares de ciertas variables que tienden a repetirse más de lo normal en ciertas situaciones específicas, y sugiere la definición de nuevas variables explicativas para posterior análisis y seguimiento.

La base de datos usada fue creada por el Banco Mundial para un estudio realizado en 2012, a manera de censo de la jurisdicción contencioso-administrativa; es importante notar, sin embargo, que es incompleta considerando que solo agrupó casos en una Jurisdicción. Sin embargo, el volumen de registros (26.000 marcados y otros 130.000 no marcados)[1] es suficiente para poder proceder con el análisis descrito. De especial interés para la ANDJE se encuentran aquéllos casos relacionados con instituciones del orden nacional. El número de casos marcados cae a 15.000,

[1]Los procesos *marcados* hacen referencia a aquellos de los cuales se conoce el fallo. Para los *no marcados* no se conoce si la pretensión se otorgó o no.

manteniendo así una riqueza de datos adecuada para los análisis propuestos.

El documento se compone de nueve secciones, comenzando por esta introducción. En la segunda sección se describe la base de datos usada en el análisis. En la tercera sección se describe la metodología de reglas de asociación, y se exhiben los resultados encontrados en los datos. En la cuarta sección se exhibe la metodología de selección de variables que permite reducir la dimensionalidad del problema. La quinta sección resume la técnica de regresión logística, describiendo la preselección de variables y los resultados obtenidos. En la sexta sección se desarrollan las metodologías de *boosting* de árboles y de redes neuronales, describiendo las técnicas usadas, y comparando los resultados encontrados. La séptima sección extiende el universo de modelos usados al marco de análisis semi-supervisado, comparando resultados de distintos modelos con los encontrados anteriormente. La octava sección describe la implementación de dos metodologías (regresión logística y *boosting* de árboles, que resulta ser la de mejor precisión) para uso de funcionarios de la ANDJE en el cálculo de probabilidades de éxito de nuevos procesos. Finalmente, la novena sección concluye.

En conjunto con este informe, forma parte integral de la entrega (i) un archivo de EXCEL donde se exhiben los resultados de la regresión logística, (ii) un archivo de EXCEL que resume las reglas de asociación halladas, y (iii) una interfaz de usuario donde se implementa un motor para calcular la probabilidad de éxito según el modelo de *boosting* de árboles. En ambos casos la interfaz es útil para un usuario que desea calcular la probabilidad de éxito de un proceso dadas las características descriptivas del mismo. La funcionalidad incluye la posibilidad de seleccionar el caso en estudio: un caso en sus comienzos, o en etapas avanzadas.

2. Descripción de los Datos

La base de datos utilizada corresponde a un censo realizado por el Banco Mundial, que incluye todos los casos abiertos en el 2012 con la información relativa al proceso en primera instancia. El total de demandas incluidas es de 153.268, de los cuales 26.637 incluyen la información sobre si se otorgó o no la pretensión. Adicionalmente, dado que la ANDJE tiene un interés particular en aquellos casos en los que el demandado es una entidad de orden nacional, se trabaja también con una porción de la base en la que se cumple esta regla. En esta base de procesos de orden nacional se incluyen 88.825 casos, de los cuales 15.945 incluyen información sobre si se otorgó o no la pretensión.

En los datos del Banco Mundial existen 348 variables para describir cada uno de los procesos judiciales, sin embargo, no todos los campos son relevantes para la construcción del modelo. Por esta razón el primer paso en la limpieza de los datos es seleccionar las variables descriptivas que serán tenidas en cuenta. Este proceso se divide en dos partes: primero, se realiza una preselección de variables tomando en cuenta la riqueza de información de cada una (es decir, se cuantifica el número de procesos para los que existe la información de cada variable); segundo, se incluyen variables adicionales (o se crean nuevas a partir de existentes, buscando un uso más eficiente de la información disponible) a partir de la relevancia percibida por el grupo de trabajo de la ANDJE, según la dinámica de los procesos. A continuación se explican en detalle estas dos partes.

2.1. Preselección según la frecuencia de ocurrencia en los procesos históricos

Para construir un modelo basado en aprendizaje de máquinas[2], es fundamental contar con información lo más completa posible. Ahora, en la base del Banco Mundial los procesos no tienen toda la información, y se reconoce que la inclusión de atributos para los cuales no se tiene información en un gran número de casos puede ser perjudicial. De las 348 variables que describen cada uno de los procesos, hay muchas que solo están disponibles para un número reducido de cas os; incluirlas puede sacrificar la precisión del modelo. Por esto, se ignoran las variables que se desconocen para muchos de los casos.

Es importante recordar que hay dos bases: una de alrededor de 27.000 procesos sobre los cuales se conoce el fallo, utilizada para realizar aprendizaje supervisado [3],

[2]Una definición para *aprendizaje de máquinas* se puede encontrar en el anexo I, junto con otras definiciones.

[3]Una definición para *aprendizaje supervisado* se puede encontrar en el anexo I.

y otra de alrededor de 120.000 procesos no marcados, que se pueden utilizar para realizar aprendizaje no supervisado.

A partir de estas dos bases se crean dos listas de atributos. La primera incluye 29 atributos que se conocen para más de 25.000 procesos de la primera base *y* más de 120.000 de la segunda. Inicialmente, todos estos atributos son candidatos para ser incluidos en el análisis. La segunda lista contiene 16 sobre los cuales se conoce la información para más de 20.000 de la base de aprendizaje supervisado *o* más de 90.000 de la de aprendizaje no supervisado. Estos atributos serán incluidos solo en la medida en que el equipo de trabajo de la ANDJE los consideren relevantes para el análisis. Estas listas se pueden ver en el anexo II.

2.2. Preselección según relevancia subjetiva

A partir de discusiones centradas alrededor de la potencial relevancia de distintas variables en la significancia de la regresión logística, se eliminan variables que podrían contar con alta riqueza de información, y se incluyen otras que cuentan con información parcial; asimismo, se determinan variables que deben ser creadas a partir de las existentes, con el objeto de resumir la información que puede tener un efecto notable en el análisis. La lista final de atributos usados es la siguiente (descripciones en itálicas corresponden a modificaciones de variables existentes o a creación de nuevas):

1. **departamento**: Departamento donde se consultó el proceso.
2. **ciudad**: Municipio donde se consultó el proceso.
3. **c_municipio_pres**: Lugar donde se presentó la demanda.
4. **corp_pres**: Nombre de la corporación de inicio: 1) Juzgados; 2) Juzgados Descongestión; 3) Tribunales; 4) Tribunales Descongestión; 5) Consejo de Estado.
5. **Magistrado**: Nombre del juez o magistrado ponente, como aparece en los documentos (nombres, apellido paterno y apellido materno). SOLAMENTE se llena para tribunal (regular o de descongestión) o Consejo de Estado.
6. **c_dmte**: Tipo de demandante: 1) Persona natural; 3) Varias naturales; 4) Persona jurídica; 5) Varias personas jurídicas; 6) Alguna combinacion de una o varias personas juridicas y naturales.
7. **c_gen_dmte**: Sexo del demandante si es persona natural: 1) Hombre; 2) Mujer.

8. **abogado_tarjeta1**: Tarjeta profesional más antigua que aparece en el escrito de los representantes del demandante, si hay al menos tres abogados. Si hay menos de tres entre todos los demandantes, se registran en las casillas "abogado_tarjeta".

9. **c_demdo_1**: Entidad estatal demandada. *Se han eliminado los casos en los que el demandado es una persona natural sin tratarse de un caso de acción de repetición.*

10. **n_demdos**: Número de demandados.

11. **n_cambiorep_dmte**: Número de veces que el demandante cambia de abogado durante el proceso.

12. **n_cambiorep_dmdo**: Número de veces que el demandado cambia de abogado durante el proceso.

13. **accion**: Acción por la que demandan.

14. **subtipo1**: Primer subtipo de la demanda. *Cuando no existe un subtipo, esta casilla se llena con la acción.*

15. **subtipo2**: Segundo subtipo de la demanda. *Si no hay un segundo subtipo, se llena con el primer subtipo.*

16. **c_monto_salario_pret**: Monto de la pretensión.

17. **c_litisconsorcio**: Si el juez ordena vincular a otra parte o si lo solicitan las partes y el juez lo concede. 1) Si; 0) No.

18. **c_suspension**: Si hubo o no suspensión del proceso: 1) Sí hubo; 0) No hubo.

19. **c_aclaracion**: Si hay o no adición o corrección de la demanda. 1) Si; 0) No. (En caso que haya adición o corrección se toma la información de ésta en lo concerniente a la admisión y notificación de la admisión).

20. **Contestacion**: *Indica si hubo o no contestación de la demanda. Se afirma que hubo contestación cuando alguna de las siguientes casillas está llena: f_rad_cont_dmdo y f_info_sec_cont_dmdo.*

21. **c_llamamiento**: Si alguno de los demandados llama en garantía a otra parte. 1) Si; 0) No.

22. **c_excepciones**: Si alguna de las partes presenta excepciones. 1) Sí presenta; 0) No presenta.

23. **N_PRUEBAS_TOTAL**: *Variable creada sumando 5 tipos de pruebas: documentales, testimoniales, interrogatorios, oficios y periciales.*

$$N\_PRUEBAS\_TOTAL = n\_documentales + n\_testimoniales + n\_interrogat\_parte + n\_oficios + n\_periciales$$

24. **c_alegatos_dmte**: Alegatos de conclusión del demandante. 1) Sí presentó; 0) No presentó.

25. **c_alegatos_dmdo**: Alegatos de conclusión del demandado. 1) Sí presentó; 0) No presentó.

26. **APELACION**: *Variable creada a partir de tres atributos de la base: c_ap_1_tipo, c_ap_2_tipo, c_ap_3_tipo. La nueva variable puede tener valor* $0, 1, 2$. *Si no hubo apelación, le corresponde el valor 0; si apeló algo diferente a la sentencia se asigna un 1, y si apeló la sentencia tiene un 2.*

27. **Duracion**: *Define una medida de la longitud del proceso:*

$$Duracion = f\_final - f\_inicio,$$

donde:

- f_inicio: *Fecha que indica el inicio del proceso. Se toma elab_admision_mand (fecha en que se elabora el auto de admisión por parte del despacho). Para los casos en los que está vacía, se considera f_not_admision_mand (fecha en que se notifica el auto de admisión o mandato de pago por parte del despacho), y si ésta tampoco está disponible, se incluye la fecha correspondiente a f_rad_esc_dem (fecha en que se radica el escrito de demanda).*

- f_final: *Como fecha final se toma f_elab_sentencia (fecha de elaboración de la sentencia); en los casos en los que ésta no está disponible, se utiliza f_ult_act (fecha de la última actuación).*

28. **nivel_entidad**: Nivel de la entidad gubernamental demandada. 1) BOGOTA - DISTRITAL; 2) Departamental; 3) Municipal; 4) NACIONAL.

Una vez se ha realizado este proceso, se puede comenzar a extraer información de los datos. En el siguiente capítulo se presentan los resultados de la aplicación de reglas de asociación, que usan esta base de datos refinada. Posteriormente, para crear modelos de clasificación se deberá volver a reducir el número de variables por medio de algoritmos de ganancia de información, los cuales determinan las variables que inciden en la probabilidad de éxito de una pretensión.

3. Reglas de Asociación

Una gran parte de los aportes de la aplicación de métodos de minería de datos a la base del Banco Mundial radican en los modelos de asignación de probabilidad que se puedan construir. Sin embargo, la minería de datos puede ser entendida como la extracción de información oculta en grandes volúmenes de datos, y por lo tanto parte de su valor está en la exposición de relaciones y tendencias que no resultan evidentes a simple vista. Las reglas de asociación se enfocan en esto. Si bien no crean un modelo de probabilidad, sí revelan relaciones entre diferentes características de una demanda.

El ejemplo clásico de reglas de asociación hace referencia a la forma como éstas se utilizan en mercadeo para determinar cómo organizar los elementos de un supermercado. Tomando las listas de compra, se encuentran los artículos que con frecuencia la gente compra juntos; por ejemplo, cuando la gente compra leche tiende a comprar también cereal. Con base en esta información, se organizan las repisas del supermercado. Así, las reglas de asociación, como su nombre lo indica, asocian diferentes elementos que con frecuencia coocurren. En este caso se busca determinar los atributos de los procesos en contra del Estado que tienden a estar vinculados.

3.1. Métricas para clasificación de reglas de asociación

En contraste con los métodos de clasificación, que buscan encontrar un pequeño conjunto de reglas que conforman un clasificador preciso, las reglas de asociación tienen como objetivo hallar todas las reglas que cumplen con un umbral de confianza y de soporte. Estas reglas son de la forma

$$A \Longrightarrow B.$$

Inicialmente se define el **soporte** como el número de casos en los que se apoya la regla; es decir, cuenta el subconjunto de casos en los que ocurre el antecedente y la conclusión. Su significancia radica en aceptar que es diferente tener una regla que en los datos de prueba ocurre para 200 casos, que una que ocurre para 20. Seguidamente, se define la **confianza**, que hace referencia a la solidez de la regla, tomando en cuenta el porcentaje de los casos en los que ocurre la consecuencia, limítandose al universo de casos en los que se tiene la premisa. Es la probabilidad de que ocurra la consecuencia, condicionando a la premisa. Estas son las dos nociones básicas para analizar reglas de asociación. De estas se derivan otras medidas que se presentan a continuación (la primera exhibe matemáticamente la definición de la confianza).

- **Confianza:** Proporción de casos cubiertos por la premisa y aquellos cubiertos por la premisa y por la conclusión.

$$conf = \frac{P(A \bigcap B)}{P(A)}$$

- **Lift:** Índice de confianza dividido por el número de casos cubiertos por la consecuencia.

$$lift = \frac{P(A \bigcap B)}{P(A)P(B)}$$

- **Conviction:** Probabilidad de que ocurra la premisa y que no ocurra la consecuencia, dividido por la probabilidad de que ocurra la premisa considerando únicamente los casos en los que no ocurre la consecuencia. Esta es una medida de independencia en las variables, comparando el comportamiento real con el comportamiento esperado si fueran independientes. Matemáticamente, es el inverso del lift de la regla $A \Longrightarrow \neg B$.

$$conv = \frac{P(A)P(\neg B)}{P(A, \neg B)}$$

3.2. Datos

Las reglas de asociación se hallaron para los procesos de orden nacional en los que se conoce la decisión judicial. Sin embargo, no se pueden incluir todas las variables seleccionadas en el capítulo dos, pues la inclusión de variables continuas en las reglas de asociación es todavía un tema de investigación, por lo que las siguientes variables no se consideran:

- Monto pretendido por el demandante.
- Duración del proceso.
- Número de demandados.
- Número de veces que cambia de representante el demandante.
- Número de pruebas presentadas por el demandante.

Por otro lado, ciertas variables, como el género y el tipo de demandante, se prestan para relaciones obvias, que dificultan el hallazgo de vínculos verdaderamente interesantes, por lo que también se decidió excluirlas. Una vez se ha hecho esto, quedan las siguientes variables (entre paréntesis se indica el nombre de la variable en la base modificada):

- Entidad demandada (*c_demdo_otras_5*).
- Ciudad (*ciudad*).
- Corporación que preside (*corp_pres*).
- Magistrado o juez que lleva el caso (*Magistrado_final_10*).
- Abogado del demandante (*abogado_final*).
- Motivo de la demanda, indicado por el subtipo 2. En caso de no haber subtipo 2, aparece el subtipo 1, y en caso de no existir subtipo 1, se indica la acción (*subtipo2*).
- Si el juez ordena vincular a otra parte o si lo solicitan las partes y el juez lo concede (*c_litisconsorcio*).
- Si hubo o no suspensión del proceso (*c_suspension*).
- Si hay o no adición o corrección de la demanda (*c_aclaracion*).
- Si hay o no contestación de la demanda (*Contestacion*).
- Si alguno de los demandados llama en garantía a otra parte (*c_llamamiento*).
- Si alguna de las partes presenta excepciones (*c_excepciones*).
- Si hay alegatos de conclusión del demandado (*c_alegatos_dmdo*).
- Si hay alegatos de conclusión del demandante (*c_alegatos_dmte*).
- Si el demandante apeló la sentencia, apeló algo más (*otro*), o no apeló (*APELACION*).
- Si la pretensión se otorga o no al demandante (*c_otorg_pret*).

Dado que las reglas de asociación no son un método de aprendizaje supervisado, la variable de otorgamiento de la pretensión se utiliza como cualquier otra variable, y no como un *target* para clasificación.

3.3. Resultados

A continuación se describe la forma como se contruyó cada sistema, y se describen los resultados encontrados. Sin embargo, por claridad de exhibición no se incluyen las tablas de resultados, las cuales se presentan detalladamente en el documento de Excel *Reglas de asociación*, que forma parte integral de esta entrega final.

El criterio principal para calificar la importancia de las reglas es el *lift*. Esta medida resulta útil pues tiene en cuenta la proporción entre el nivel de confianza de la regla y el número total de procesos para los que se tiene la consecuencia. Es decir, una regla puede tener un alto nivel de confianza, pero su importancia debería ser reducida si la conclusión se presenta también frecuentemente entre el universo total de casos; el *lift* permite interpretar esta frecuencia en exceso de forma directa.

Sin embargo, a la hora de leer los resultados, se debe tener en cuenta todas las métricas. Particularmente, *conviction* permite establecer una comparación con el comportamiento que existiría si las variables fueran independientes; a medida que este valor aumenta, el vínculo entre éstas se hace mayor.

Tabla 1.1

Al aplicar reglas de asociación a todas las variables antes mencionadas (estableciendo un soporte mínimo de 0.05, es decir, que los casos que soporten la regla representen al menos un 5 % del total de los casos), se encuentran las reglas de la hoja 1.1 en el documento de Excel.

Sin embargo, esta primera serie de reglas no aportan mucha información, pues en todas aparece la variable en la que el magistrado es 0,0, lo que quiere decir que la información del magistrado o juez que llevó el caso no estaba disponible. Excluir todos los casos en los que no se reporta quién es el magistrado podría resultar en la pérdida de mucha información relevante, por lo que se plantea un enfoque diferente: se construye un sistema de reglas excluyendo todos los casos en los que el magistrado o el abogado no se conocen o están marcados como 'otro' por ser comunes a menos de 10 casos, y un segundo sistema excluyendo la variable de abogado y magistrado, para ver el comportamiento de las demás variables. Al hacer esto, en el sistema de reglas del primer modelo no aparece ningún abogado o magistrado, por lo que se decide omitir este modelo y considerar únicamente el presentado en la tabla 1.2. Esta decisión se toma pues el hecho de que en el primer caso no aparezca ningún abogado o magistrado indica que no se está dejando de lado ninguna regla relevante al excluir estas variables del modelo; sin embargo, sí se puede perder información si se limita la base a aquellos casos en los que tanto abogado como magistrado están reportados.

Tabla 1.2

Esta tabla muestra los resultados cuando se excluyen las variables de abogado y magistrado, tal como se acaba de explicar. Lo que se muestra son unos vínculos entre la suspensión del proceso, la presentación de alegatos por parte del demandante, la apelación, el llamamiento en garantía por parte de alguno de los involucrados y la vinculación de otra parte por decisión del juez o por pedido de alguna de las partes. De todas formas, el valor del *lift* en estos casos, si bien es mayor a 1 indicando que el comportamiento no es normal, no es tan alto, lo que nos indica que estas reglas no se desvían mucho del comportamiento esperado. En la primera, vemos que en los casos en los que el demandante presenta alegatos y no hay una suspensión del proceso, el 83 % de estos tienen una apelación a la sentencia (de alguna de las partes), sin vincular a otra parte ni llamarla en garantía. En la segunda regla vemos la implicación opuesta, la cual tiene una confianza de 62 %. Las otras tres reglas muestran diferentes formas en las que se relacionan estas cinco variables.

Tabla 1.3

Este sistema de reglas excluye todas las variables binarias (las que resultaban protagónicas en el modelo anterior, excepto por apelaciones, que no es binaria), para poder capturar la forma como las demás se relacionan. La única variable binaria que se incluye es si se otorga o no la pretensión.

La primera regla indica que el 28 % por ciento de los procesos ante juzgados son por nulidad y restablecimiento del derecho, de tipo laboral y más especificamente con subtipo 2 prestacional. En la segunda regla se observa que el 94 % de las demandas por esta causa se llevan en un juzgado.

En esta tabla también se encuentra información relativa a las demandas de este mismo grupo pero que tienen como subtipo 2 pensión. De los 4.922 procesos por este motivo, el 68 % se ha presentado en un juzgado y ha tenido apelación. A su vez, el 38 % de los procesos en juzgado que tienen apelación son por pensión. Las siguientes reglas muestran relaciones similares entre este tipo de demanda, el hecho de que la corporación que preside sea un juzgado y la posible apelación de la decisión judicial. Las dos últimas reglas dan información sobre el comportamiento de los procesos en Bogotá. De acuerdo con la última regla, el 90 % de los casos presentados en esta ciudad en los que no se otorga la pretensión al demandante tienen una apelación.

Tabla 1.4

Incluye las mismas variables de la tabla anterior, excepto la corporación que preside el caso y si se presenta o no apelación. Es decir, esta serie de reglas muestra

relaciones entre la entidad demandada, la ciudad, el motivo de la demanda y si se otorga o no la pretensión. En esta, se hacen evidentes los motivos principales por las que ciertas entidades tienden a ser demandadas, y en algunos casos se puede ver si estas pretensiones tienden a otorgarse o no.

El 76 % de las demandas en contra de la Fiscalía son por privación injusta de la libertad, y en el sentido opuesto se encuentra que el 81 % de las demandas por este motivo son en contra de la Fiscalía.

Por otra parte, el 47 % de las demandas cuyo motivo es prestacional, en una acción por nulidad y restablecimiento del derecho, que sí son otorgadas, son en contra de la Caja de Retiro de las Fuerzas Militares. Estas demandas constituyen el 42 % de los procesos en contra de esta entidad. En este caso, el *lift* de 3.87 indica que esta entidad presenta una anomalía que dista bastante del comportamiento normal de las demandas de tipo prestacional que son otorgadas. En la Caja de Sueldos de Retiro de la Policía Nacional sucede lo contrario: el 72 % de las demandas de este tipo que no se otorgan son en contra de esta entidad. Así, las reglas muestran que si bien ambas cajas tienen una gran cantidad de demandas por este motivo, la Caja de Retiro de las Fuerzas Militares tiene una tendencia a perder mayor a la del promedio, mientras la Caja de Sueldos de Retiro de la Policía Nacional abarca un gran porcentaje de las demandas por este motivo en las que el Estado ha ganado.

Por último, de las demandas en contra de la Caja Nacional de Previsión Social e.i.c.e. - en liquidación, el 93 % son por pensión. Además, la regla en la dirección opuesta indica que el 22 % de los procesos por este motivo son en contra de esta entidad. Sin embargo, las reglas no muestran ninguna tendencia que relacione esta entidad con que la pretensión sea o no otorgada.

Tabla 1.5

La tabla 1.5 establece los abogados y magistrados que coinciden en los procesos de una forma que se aleja del comportamiento normal. Para encontrar estas reglas, se utilizaron 1.321 casos en los que se conoce tanto el abogado como el magistrado. Las primeras líneas nos muestran una relación con valores de *lift* muy elevados; es decir, distan mucho del comportamiento normal. De los 57 casos que ha llevado el abogado de tarjeta profesional 31614, en 15 ha coincidido con la juez Amparo Oviedo Pinto y no le ha sido otorgada la pretensión. Adicionalmente, el 58 % de las pretensiones que esta juez no ha otorgado han tenido como representante del demandante al abogado con esta tarjeta profesional.

Por su parte, el abogado de tarjeta profesional 45113 coincide con los jueces Luz Elena Sierra Valencia, Álvaro Montenegro Calvachy y Melva Giraldo Londoño.

Por ejemplo, de los 31 casos de Luz Elena Sierra Valencia, este abogado ha estado presente en la mitad. Como se puede ver en la tabla, estas coincidencias tienden a ir acompañadas de que un no otorgamiento de la pretensión.

Aquí, el papel del analista es fundamental, pues los factores geográficos, por ejemplo, no se están teniendo en cuenta. Si estos procesos corresponden a un municipio pequeño, esta correlación puede ser normal, mientras en Bogotá posiblemente sea un comportamiento atípico que deba ser investigado; es importante el insumo del analista para llegar a estas conclusiones.

Tabla 1.6

Las tablas 1.6 y 1.7 corresponden a análisis de relaciones adicionales solicitadas por la ANDJE. La primera busca establecer si hay abogados especializados en demandar a una entidad. Si bien en un primer momento se habían incluido también los motivos de la demanda, las reglas resultaban más obvias y correspondían más a un comportamiento normal, por lo que se decidió buscar relaciones entre abogados y entidades estatales. En los resultados se puede observar que hay algunos abogados que demandan a una única entidad, así como casos en los que un alto porcentaje de las demandas contra la entidad son representadas por un mismo abogado del demandante. Las diez reglas presentadas son todas las que tienen un nivel de soporte mayor al 3 %.

Entre la Caja de Retiro de las Fuerzas Militares y el abogado de tarjeta profesional 170560 hay una relación que no corresponde al promedio. El 84 % de los casos de este abogado son en contra de esta entidad, pero llama aun más la atención que el 25 % de los procesos en contra de esta entidad tienen como representante del demandante a este abogado.

Una relación similar existe entre la Fiduciaria La Previsora S.A. y el abogado de tarjeta profesional 97002; el 81 % de los casos de este abogado son contra esta entidad, y el 20 % de las demandas contra la fiduciaria son representados por este abogado. Lo mismo ocurre para la Caja de Sueldos de Retiro de la Policía Nacional y el abogado de tarjeta profesional 31614, el cual representa el 16 % de los procesos contra la entidad, los cuales corresponden al 93 % de los procesos que este abogado tiene contra la Nación. En una situación similar, el 26 % de de los casos contra la Caja Nacional de Previsión Social e.i.c.e.- en liquidación, son representados por el abogado de tarjeta profesional 31571. El mismo valor de 26 % aparece cuando se ven los casos de la Caja de Sueldos de Retiro de la Policía Nacional representados por el abogado de tarjeta profesional 45113, el cual concentra el 92 % de las demandas que tiene contra el Estado en esta entidad.

Todas estas reglas van acompañadas de índices estadisticos que corresponden a anomalías que distan mucho de lo que sería un comportamiento normal. Sin embargo, es importante enfatizar la necesidad de análisis e investigación por parte de analistas, pues si bien estos comportamientos son irregulares desde una perspectiva estadística, puede tener explicaciones lógicas como, por ejemplo, que las demandas contra estas entidades requieren de altos niveles de experticia.

Tabla 1.7

La tabla 1.7 establece la relación entre la contestación, la presentación de alegatos por parte del demandado y las entidades estatales. La primera parte de la tabla presenta reglas de asociación entre contestación y presentación de alegatos. La primera regla indica que en el 95 % de los casos en los que la entidad demandada presenta alegatos también ha presentado contestación, mientras la segunda indica que el 56 % de los casos en los que hay contestación también hay alegatos. Ambos tienen un *lift* muy cercano a 1, lo que indica que la regla tiene una atipicidad baja; sin embargo, puede ser de interés para la ANDJE. La segunda regla describe la relación contraria: el 57 % de los casos en los que hay contestación también hay alegatos del demandado.

La Caja de Sueldos de Retiro de la Policía Nacional tiene una tendencia por encima del promedio a no presentar alegatos, tanto en los casos en los que contesta como en los que no lo hace. El 39 % de los casos que no son constestados ni tienen alegatos del demandado corresponden a esta entidad; a su vez, de los casos contestados pero sin alegatos el 36 % corresponden a esta caja de retiro. En general, el 36 % de los procesos sin alegatos del demandante son en contra de esta entidad estatal. Las seis reglas de la tabla que involucran a esta entidad muestran estas relaciones más en detalle.

También llama la atención la alta tendencia de la Fiduciaria la Previsora S.A. a no contestar las demandas. El 19 % de sus casos no reciben contestación, lo que equivale al 23 % de los casos que no son contestados. El 82 % de los casos en su contra son contestados e incluyen alegatos, y de los casos que contestan, presentan alegatos en el 86 % de estos, los cuales representan el 11 % del total de casos en los que el demandado presenta alegatos.

La Superintendencia de Servicios Públicos Domiciliarios tiene un comportamiento similar, pues presenta alegatos en el 80 % de los casos en los que es demandada.

Las anteriores son las reglas que muestran relaciones entre entidades demandadas, presentación de alegatos y contestación, con un *lift* mayor a 1.5, lo que indica que su comportamiento varía en relación al promedio. Estos resultados, si bien son bastante dicientes sobre el comportamiento de cada entidad, deben ser entendidos

como indicios para los analistas e investigadores de la ANDJE, quienes deben determinar a qué se debe cada comportamiento. Por ejemplo, la falta de presentación de alegatos por parte de una entidad puede deberse a que ésta no invierte los recursos suficientes en su defensa, pero también puede ser síntoma de prácticas irregulares dentro de una entidad, o podría ser explicada por el motivo por el que la demandan y el contexto legal no se prestan para que presente alegatos. Estas son solo algunas de las posibilidades que el analista debe estudiar, de modo que pueda obtener conclusiones de la investigación utilizando las reglas de asociación como brújula.

4. Selección de Variables para Modelos de Clasificación

Según lo obtenido en el capítulo 2, se están considerando menos de 30 variables para describir los procesos. Sin embargo, esto no es computacionalmente exacto; de las 26 variables elegidas en este momento, muchas son categóricas, con amplios rangos de valores. Por ejemplo, los campos *Abogado*, *Magistrado* y *Demandado* pueden tomar cientos de valores diferentes cada uno.

En el proceso de construcción de modelos de clasificación es necesario convertir estas variables en dummies. Es decir, para cada posible valor en una categoría se crea una nueva variable, a la cual se le asigna un valor de 0 o 1, según el valor asumido en cada proceso dado. Así, la base pasa a tener más de 9.000 variables. Esto presenta un gran reto computacional, pues con la nueva forma de representar cada proceso, se tiene una matriz de $27{,}000 \times 9{,}000$ para todos los procesos marcados y de $15{,}000 \times 9{,}000$ para aquellos que además son de orden nacional. La memoria de los programas utilizados para realizar el aprendizaje de máquinas no soporta tantos datos, por lo que se hace una reducción del número de variables. Este proceso se divide en dos etapas: la primera anula aquellos valores que sólo están presentes para un número muy reducido de procesos y la segunda se basa en algoritmos de ganancia de información e influencia relativa que arrojan un número reducido de variables a tener en cuenta utilizando como criterio la influencia que tienen en que se otorgue o no la pretensión. A continuación, se presentan ambos pasos en detalle.

4.1. Disminución del número de variables

Con la nueva forma de representar cada proceso, se tiene una matriz de $27{,}000 \times 9{,}000$. Para facilitar el manejo computacional de los datos, se disminuye el número de variables tomando en cuenta los siguientes criterios:

- Todos los abogados con menos de 20 casos se anulan, y el número de su tarjeta es reemplazado por la palabra “otro”.
- Las entidades demandadas en menos de 20 casos se agrupan bajo el título “otra”.
- Los magistrados con menos de 20 casos se reemplazan por la palabra “otro”.

Con esto, el número de variables se reduce a menos de 1.000.

4.2. Selección con algoritmos de ganancia de información e influencia relativa

El paso siguiente es determinar las variables que inciden en la probabilidad de éxito. En el caso de todos los procesos marcados se utilizan dos algoritmos en el proceso; esto dado que el segundo algoritmo provee información más precisa, pero puede tener limitaciones de memoria, lo cual obliga a realizar la selección en dos pasos. Una vez se reduce la base a aquéllos de orden nacional, no es necesario utilizar ambos, por lo que se aplica directamente el segundo. Esto no perjudica el proceso de ninguna forma, pues una variable que hubiera sido excluida por el primer modelo también será rechazada por el segundo.

Utilizando el algoritmo *InfoGainAttributeEval* de WEKA[4] se genera un ranking de los atributos más relevantes para determinar la probabilidad de éxito de una demanda contra el Estado. La fórmula aplicada por este paquete para seleccionar variables es la siguiente:

$$InfoGain(Clase, Atributo) = H(Clase) - H(Clase|Atributo),$$

donde H es la entropía y la clase hace referencia al éxito/fracaso de una demanda[5]. En el anexo III se puede ver el ranking obtenido. A partir de esto, se decide excluir dos variables, que no aportan información al análisis de acuerdo con el algoritmo: *c_alegatos_dmte* y *n_cambiorep_dmdo*.

Para la primera, 15.991 de 25.981 procesos presentan alegatos, mientras 4.603 no presentan y los demás no contienen esta información. La presencia de alegatos, sin embargo, no parece impactar el éxito del proceso; alrededor del 45,35 % de todos los casos incluidos en la base de datos de Banco Mundial, en los que el demandado es una entidad de orden nacional, los gana el demandante. Si se consideran únicamente los casos en los que hay alegatos por parte del demandante, este porcentaje es del 44,74 %. La figura 1 muestra la distribución de los casos ganados y perdidos.

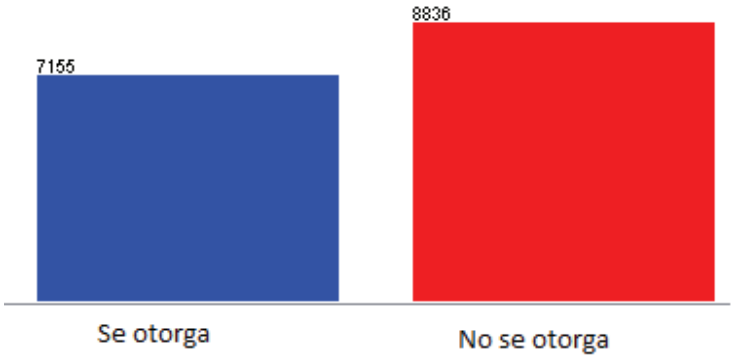
Para entender la variable que recoge el número de veces que cambió de representante el demandado, primero se elimina un proceso que reporta 105 cambios, pues al analizar los otros elementos del proceso, todo indica que esto corresponde a un error.[6] La primera observación del comportamiento de la variable es que los demandados tienden a cambiar de abogado más veces que los demandantes. El valor más alto registrado para el número de cambios realizado por el demandante es 9,

[4]WEKA es un software que reúne una colección de algoritmos de aprendizaje de máquinas para resolver problemas de minería de datos.

[5]Para una mejor descripción del atributo, ver anexo I.

[6]La duración del proceso fue de 211 días, y el monto demandado es de 4.620.959 pesos, lo cual apunta a un error de registro.

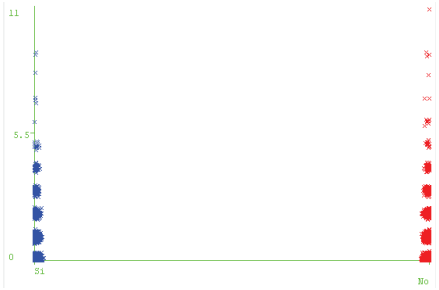
Figura 1: Casos en los que el demandante presenta alegatos



y para el demandado es 11. Esta diferencia no es muy grande; sin embargo, mientras el promedio de cambios para los demandantes es de 0.096, con una desviación estándar de 0.379, para los demandados el promedio es de 0.247 y la desviación estándar es de 0.671. Ahora, el número de cambios por parte del demandado no parece afectar la probabilidad de éxito. De los 3.789 casos en los que no hay cambio de representante, el porcentaje de pretensiones otorgadas es de 44,1 %. En 565 procesos hubo un cambio de representante por parte del demandante; de estos, el 45,84 % los ganó el demandante. Otros 137 casos reportan dos cambios de abogado por parte de la entidad demandada, de los cuales el 50 % los perdió el Estado. Si bien hasta este punto se puede ver un ligero incremento en la probabilidad de éxito para el demandante, al ver los 75 procesos en los que hubo tres cambios de abogado del demandante, esta tendencia se revierte: solo 37,84 % de estos los gana el demandante. Para los 110 casos en los que hay cuatro cambios de representante, el porcentaje de pretensiones otorgadas vuelve a su promedio general, con 44,55 %. De ahí en adelante, hay pocos casos; por ejemplo, solo 37 procesos tienen 5 cambios, y cada vez el número disminuye más. Así, el modelo no considerará relevante esta variable.

Para dar una mejor noción del comportamiento de este atributo, la figura 2 muestra a la izquierda -en azul- los procesos que el Estado ha perdido y a la derecha -en rojo- los que ha ganado. El eje y corresponde al número de cambios de representante que hace el demandado.

Figura 2: Distribución de casos ganados y perdidos de acuerdo con el número de cambios de representante del demandado.



4.3. Selección final con *Generalized Boosted Models*

Utilizando el paquete *GBM* de R Project, se hace la selección final de variables. Este algoritmo utiliza *boosting* de árboles, combinando clasificadores débiles para construir un modelo fuerte, que finalmente entrega la lista de variables con su influencia relativa en el modelo. Esta influencia relativa hace referencia a la importancia de cada una de las variables, de modo que la suma de todas las influencias relativas es 100.[7]

A partir del ranking que realiza el algoritmo, se escogen las variables más importantes, utilizando como criterio que la influencia relativa de estas sumen más de 99. Al hacer esto, se obtiene una lista final de 50 variables, presentada en el anexo V.

Este proceso se repite para los procesos en los que la entidad demandada es de orden nacional. En este caso, la influencia está más dispersa entre las variables, por lo que la lista final la conforman 64 variables que juntas alcanzan más de 97 de la influencia relativa total. A partir de esta lista se construyen todos los métodos de clasificación que se presentan a continuación.

[7]Para una explicación en mayor detalle de esta metodología, ver [1].

5. Regresión Logística

Un modelo logístico se puede escribir como:

$$P(y_i = 1) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}},$$

donde $y_i = 1$ indica que el proceso i-ésimo fue fallado a favor del demandante, x_i es el vector que contiene la información relativa al proceso explicada por las variables seleccionadas, y β es el vector que contiene los coeficientes. Este vector se calcula utilizando *iteratively reweighted least squares (IRLS)*, un algoritmo que optimiza mediante mínimos cuadrados ponderados de forma iterativa. A continuación se presentan dos modelos: el primero incluye todas las variables seleccionadas; para el segundo se hace una nueva selección utilizando *Generalized Boosted Models* tomando en cuenta únicamente las variables que se conocen desde el comienzo del proceso.

5.1. Regresión con todas las variables seleccionadas

A la lista final de 50 variables se le aplica una regresión logística, creando un modelo que permite predecir la probabilidad de éxito de una demanda contra el Estado a partir de estas variables. Los coeficientes de la regresión se pueden ver en el anexo V. Es importante tener en cuenta que la variable que se intenta predecir es el otorgamiento de la pretensión; un valor de 1 indica que la pretensión se otorga al demandante, mientras un valor de 0 significa que el Estado gana. Por lo tanto, una mayor probabilidad de éxito perjudica al Estado.

Primero se observa el intercepto, cuyo valor es 0,052055, lo que implica que antes de añadir cualquier variable, el modelo asigna una probabilidad de éxito al demandante de:

$$\frac{e^{0,052055}}{1+e^{0,052055}} = 51{,}301\,\%.$$

Para el análisis de las variables, coeficientes negativos acompañan variables que se relacionan con una mayor probabilidad de éxito para el Estado, y variables con coeficientes positivos señalan hacia un aumento en la probabilidad de que la pretensión se otorgue al demandante. La duración, por ejemplo, tiene un coeficiente de $-0{,}00021226$. Esto indica que entre más tarda un proceso, más probable es que el Estado lo gane.

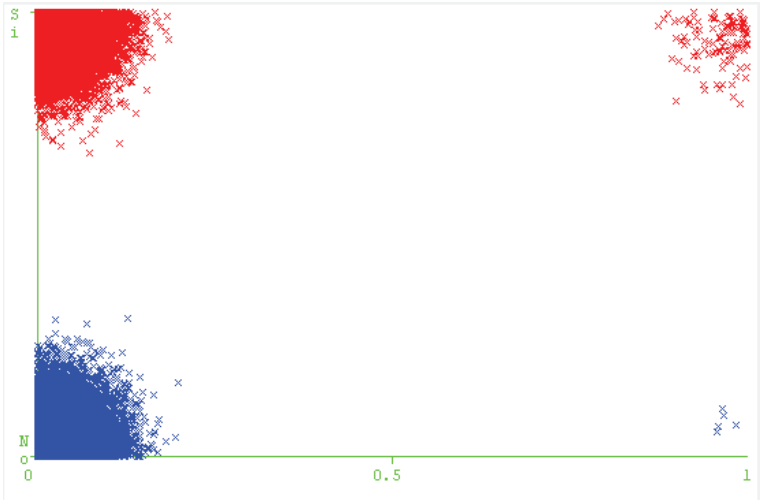
Para cuantificar la sensibilidad de la probabilidad de éxito a cambios unitarios en las variables se define la razón *log-odds* como sigue, donde c es el coeficiente de la variable y x es el intercepto:

$$\frac{e^{x+c}}{1+e^{x+c}}.$$

En el caso de la duración, la medida está en dias. Así, un día adicional en la duración cambia la probabilidad de que el Estado pierda a 51,296 %, es decir, cada 100 días que pasan, la probabilidad de que la pretensión se le otorgue al demandante disminuye 0,5 %.

Otras variables tienen un coeficiente positivo, lo que indica que aumentan la probabilidad de que el demandante gane. Cuando la entidad demandada es el departamento del Atlántico, por ejemplo, el coeficiente es de 3,1144. Si no hay otras variables presentes, la probabilidad de que el Estado pierda en este caso aumenta radicalmente, pasando a ser 95,95 %. Este comportamiento se evidencia en la figura 3, la cual compara el comportamiento de las demandas a otras entidades con las demandas a este departamento.

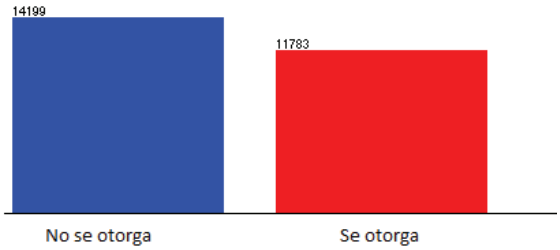
Figura 3: A la izquierda la distribución total de casos ganados y perdidos, a la derecha la distribución de casos ganados y perdidos en Atlántico. Los puntos en la parte superior corresponden a pretensiones otorgadas al demandante y aquellos en la parte inferior de la gráfica son aquellos en los que el Estado gana la demanda.



Los puntos rojos corresponden a aquellos procesos en los que el Estado pierde, y los azules son los que gana. A la derecha se encuentran los casos en los que la entidad demandada es el departamento del Atlántico, donde prácticamente todos los casos son fallados a favor del demandante, y los de la izquierda son todos los demás, donde el comportamiento promedio está dividido de forma casi equitativa, como se muestra en el histograma de la figura 4.

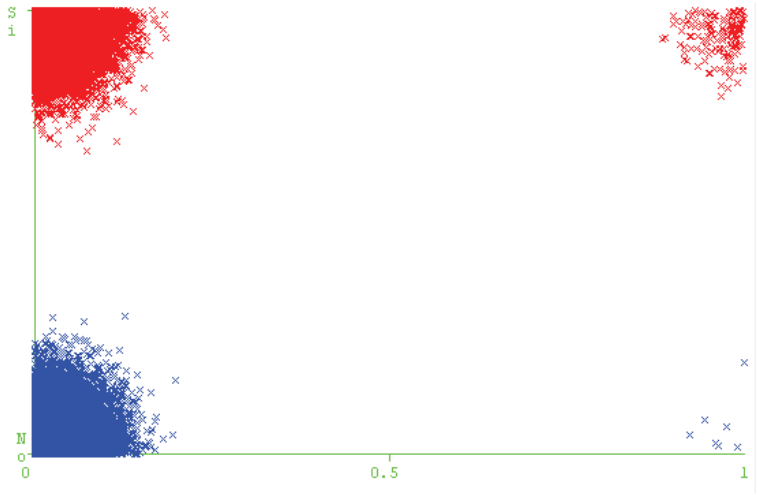
Un comportamiento similar se observa con el departamento de Antioquia, para el que la probabilidad de que el demandante gane, sin tener en cuenta ninguna otra variable, aumenta a 95,112 %. El comportamiento anómalo de los fallos en este

Figura 4: Fallos de todos los procesos incluidos en el modelo.



departamento se observa en la figura 5.

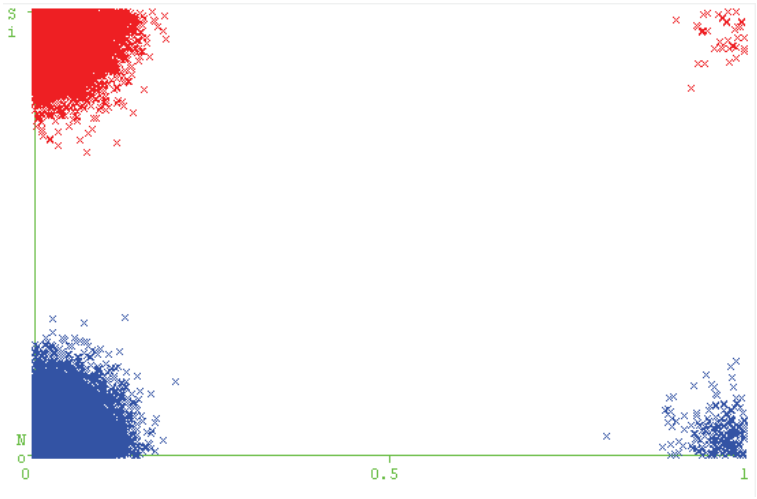
Figura 5: A la izquierda la distribución total de casos ganados y perdidos, a la derecha la distribución de casos ganados y perdidos en Antioquia. Los puntos en la parte superior corresponden a pretensiones otorgadas al demandante y aquellos en la parte inferior de la gráfica son aquellos en los que el Estado gana la demanda.



Otras entidades presentan un comportamiento favorable para el Estado, aunque en ningún caso el impacto es tan drástico como en los anteriores. Cuando el demandado es Cali, por ejemplo, el coeficiente es $-0{,}74015$, lo que implica que la probabilidad de que la pretensión se otorgue disminuye a 33,446 %. Los fallos de los procesos donde el demandado es Cali, en comparación con el total de fallos, se presenta en la figura 6.

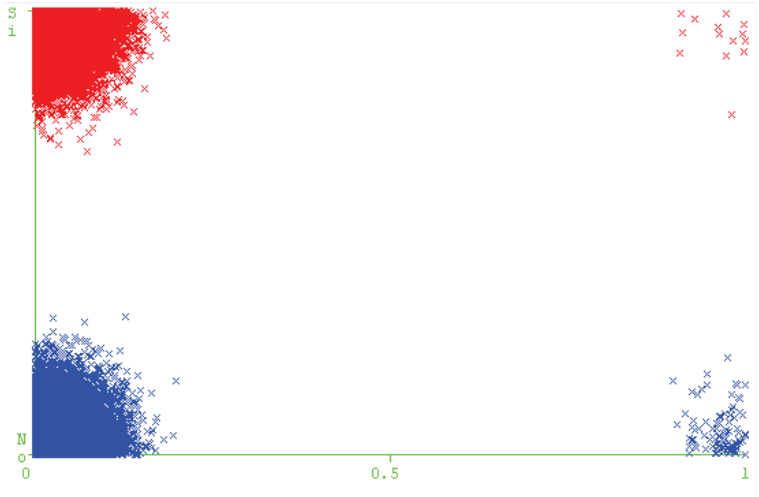
Otra característica del proceso que resulta muy útil para determinar la probabilidad de éxito del demandante es la causa de la demanda, tanto el subtipo 1 como

Figura 6: Distribución de casos ganados y perdidos vs. casos ganados y perdidos en Cali.



el 2 (variables que describen los hechos y los desagregan en dos niveles). En el caso del subtipo 1, cuando éste es por Servicios Públicos, por ejemplo, la probabilidad disminuye a 8,651 %. De nuevo, esto se debe a una distribución anómala entre los procesos fallados a favor y en contra, como se ve en la figura 7.

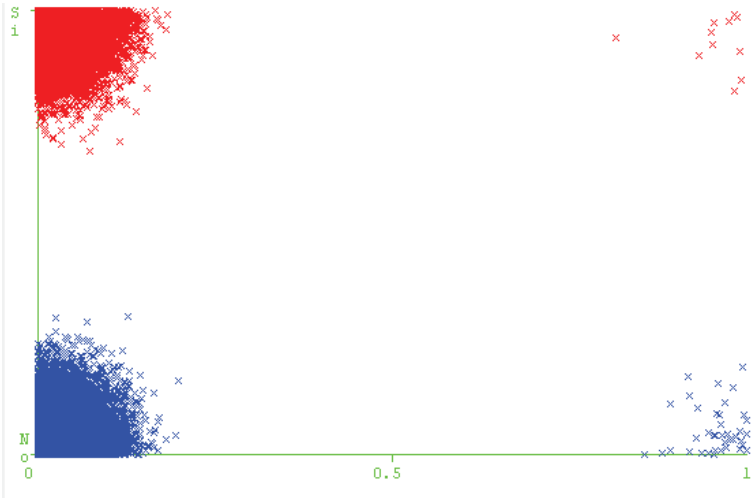
Figura 7: Distribución de casos ganados y pérdidos vs. casos ganados y perdidos por servicios públicos.



La misma tendencia se da cuando el subtipo 1 es el Sistema de Seguridad Social en Salud, como lo muestra la figura 8. En este caso, un coeficiente de $-1{,}8096$

disminuye la probabilidad de que se otorgue la pretensión a 14,71 %.

Figura 8: Distribución de casos ganados y pérdidos vs. casos ganados y perdidos para el subtipo 1 *Sistema de Seguridad Social de Salud*.



En el caso del subtipo 2, hay muchos que son relevantes para el modelo de predicción. Llaman la atención, por ejemplo, *Salarial* y *Pensión*, que presentan un comportamiento opuesto (aunque con magnitudes menores a las observadas en ejemplos anteriores). El primero tiene coeficiente negativo, y una probabilidad base de que el demandante gane de 27,41 %; el segundo, con un coeficiente positivo, aumenta la probabilidad de que el demandante gane a 65,51 %.

El documento EXCEL acompañante, y el anexo, muestran los coeficientes (y el impacto en la probabilidad en el caso del archivo de EXCEL) de las variables restantes.

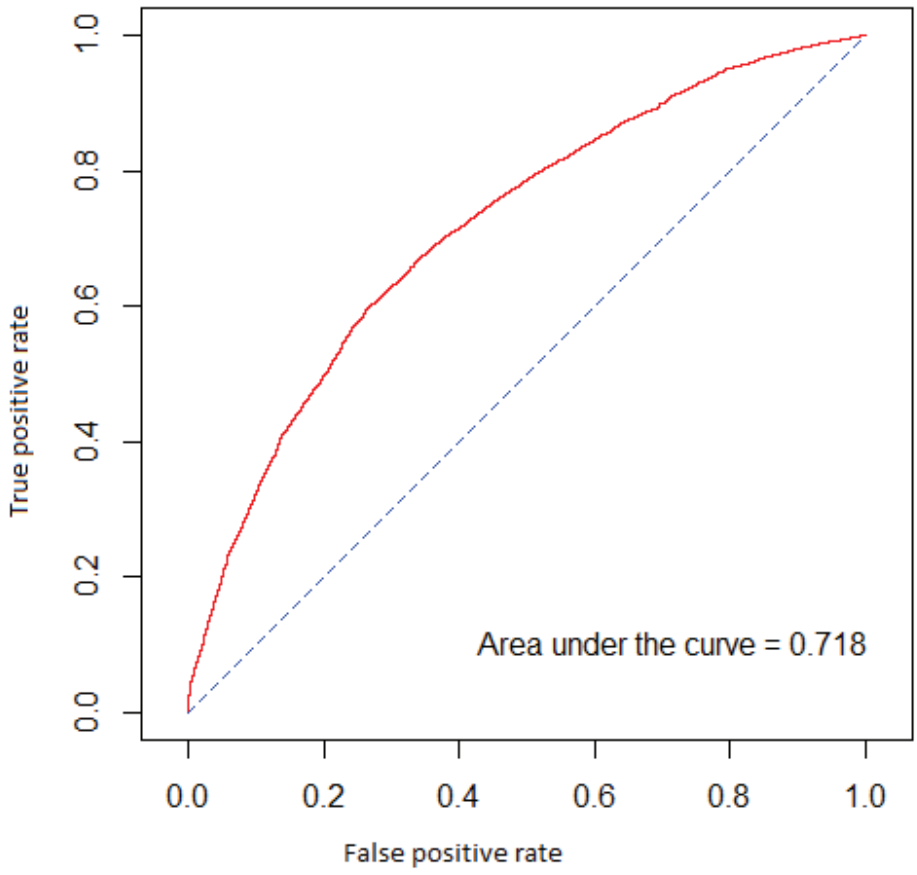
Evaluación del modelo

La bondad del modelo se mide con la curva ROC y los valores p de los coeficientes[8]. La curva ROC grafica el éxito de calificación positiva (número de positivos pronosticados acertadamente dividido por el número total de positivos) contra el error de calificación positiva (número de falsos positivos dividido por el número total de negativos); es decir, para cada umbral dado del error de calificación positiva se encuentra el nivel de acierto en pronósticos positivos. En esencia se buscan modelos capaces de pronosticar acertadamente los casos positivos, sin incurrir considerablemente en el error de falsos positivos, que producirían curvas sesgadas a la esquina

[8]Para una lectura sobre la interpretación de la curva ROC en aprendizaje de máquinas, ver [4].

superior izquierda de la gráfica. La medición del nivel de aserción de un modelo se refleja en el área bajo la curva ROC, que indica la eficacia del modelo en la clasificación de datos. El área bajo la curva ROC para el modelo logístico desarrollado es 71,8 %.

Figura 9: Curva ROC para regresión logística general.



Adicionalmente, a cada variable del modelo le corresponde un valor p, el cual cuantifica la significancia de esta variable en la predicción. En general, una variable se considera significativa si tiene un valor p menor a 0.05. En el anexo V se muestran estos valores.

En el documento de Excel que se adjunta a esta entrega, el usuario puede calcular la probabilidad de éxito que el modelo le asigna a una demanda contra el Estado, de modo que pueda ser aplicado en demandas que no forman parte de la base de

datos del Banco Mundial utilizada para la creación del modelo logístico.

5.2. Regresión para procesos de orden nacional

Dado que la ANDJE trabaja principalmente con procesos de orden nacional, se realizó una segunda regresión para obtener un modelo de probabilidad de éxito para estos procesos. Esta base contiene 15.945 casos marcados. El modelo que se presenta a continuación difiere del presentado en la primera entrega, pues en el primero no se incluyen alrededor de 2.000 casos de la Caja Nacional de Previsión Social. Sin embargo, las conclusiones presentadas en ese primer informe son igualmente válidas. Los coeficientes difieren y, como se verá más adelante, se aumenta ligeramente la precisión del modelo, pero la forma como las variables inciden en la probabilidad de que se otorgue una demanda pueden ser utilizados por la ANDJE como indicios para sus análisis e investigaciones.

En cuanto a las variables incluidas, aquí se decidió excluir la variable de *acción*, pues la forma de construir *subtipo1* y *subtipo2* hacen que estas incluyan la información de la primera. Las variables incluidas de acuerdo con la influencia relativa en la probabilidad de éxito del proceso son las que se contruyeron en el capítulo cuatro y que aparecen en el anexo IV.

Los coeficientes, junto con su valor p, y el efecto marginal, es decir, la probabilidad de que se otorgue la pretensión cuando sólo está esa variable presente, se incluyen también en el anexo VI. Este último valor indica la probabilidad de éxito cuando solo esta variable es tomada en cuenta.[9] Es importante notar que la probabilidad no corresponde a un caso real, pues ningún proceso es descrito por una única variable. Su objetivo es dar una noción más clara de la sensibilidad de la probabilidad final a las distintas variables.[10] La distribución de los casos ganados y perdidos se presenta en el histograma de la figura 10.

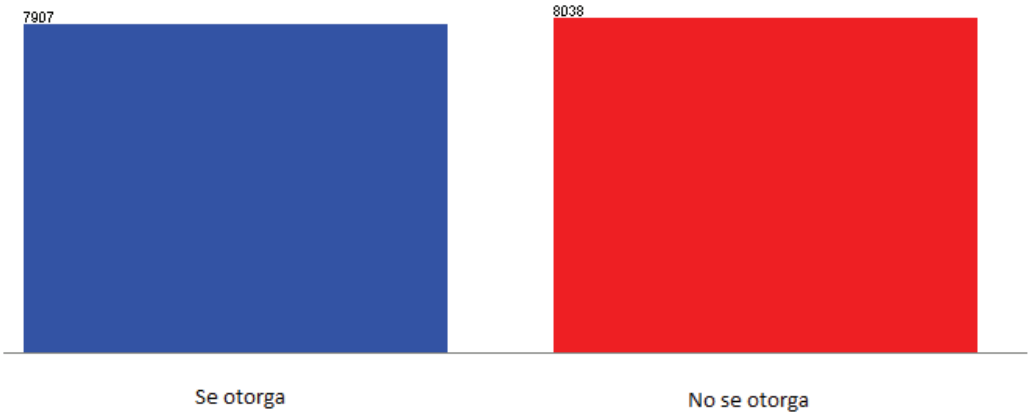
Al analizar los coeficientes y el comportamiento de las variables, se pueden extraer muchas conclusiones. A continuación, se hace una descripción de los resultados para cada una de las variables incluidas en el modelo.

En este punto, vale la pena notar que el criterio para determinar las variables que se incluyen en el modelo logístico está dado por la influencia relativa de éstas, de

[9] En el caso de las variables continuas se toma un valor igual a 1 para este análisis de sensbilidad, igual que en el caso anterior. Es decir, se mide el efecto marginal al aumentar en una unidad el valor de la variable.

[10] El valor es la razón *log-odds*, $\frac{e^{x+c}}{1+e^{x+c}}$, donde x es el intercepto y c es una unidad de la variable que se está considerando.

Figura 10: Fallos para los procesos de orden nacional.



acuerdo con *Generalized Boosted Models*. Por lo tanto, también se incluyen aquéllas para las cuales el coeficiente de la regresión logística tiene un valor p elevado. Esto no perjudica el modelo ya que los coeficientes de éstas, así como su efecto marginal, es bajo, por lo que el hecho de que tengan un márgen de error alto no compromete la precisión del modelo. Adicionalmente, el signo de los coeficientes sí permite entender en qué dirección impactan la probabilidad del modelo, ya que éste responde a la proporción de casos ganados y perdidos asociados a la variable en cuestión.

Duración: El coeficiente para la duración es muy similar al de la regresión logística con todas las variables seleccionadas (ver explicación en sección 5.1).

Abogado: Hay quince abogados de los demandantes que el modelo considera como relevantes para determinar la probabilidad de éxito de un proceso; adicionalmente, el hecho de que el caso lo tenga un abogado con menos de veinte casos en contra del Estado también incide. Siete de estos afectan la probabilidad de modo que el Estado se ve favorecido, mientras ocho se relacionan con una mayor probabilidad de que se otorgue la pretensión. El cuadro 1 muestra las tarjetas profesionales de los abogados, junto con el número de casos de cada uno y la probabilidad de que la pretensión se otorgue considerando sólo esta variable. La línea divide la tabla en dos, agrupando arriba de la raya a los abogados que influencian la probabilidad de éxito a favor del Estado, y debajo de la raya aquellos que aumentan la probabilidad de que se otorgue la pretensión al demantante. Esta convención se mantendrá para todas las tablas del informe.

Cuadro 1: Abogados del demandante que influyen en la probabilidad de éxito.

Tarjeta profesional	Número de casos	P_0
83363	103	17.01 %
31614	200	24.75 %
56834	108	25.09 %
31571	326	26.13 %
158718	171	28.67 %
59415	165	34.64 %
45113	237	37.51 %
otro	9137	52.03 %
91183	103	65.74 %
95908	126	65.97 %
41146	125	68.87 %
170560	324	74.23 %
90682	113	74.97 %
98987	80	75.73 %
109557	124	87.25 %
112907	247	96.09 %

Apelación: La ausencia total de apelaciones aumentan la probabilidad de que el demandante gane. Con un coeficiente de 0,89 esta variable tiene un P_0 de 71 %.

Es interesante notar que en 12.540 procesos contra entidades nacionales hay apelación de la sentencia. El promedio de la duración de procesos con apelaciones de este tipo es de 893 días, mientras los 3.104 procesos que no tienen nigún tipo de apelación tardan en promedio 818 días en ser fallados; es decir, la duración aumenta en un 8 % cuando se presenta una apelación a la sentencia.

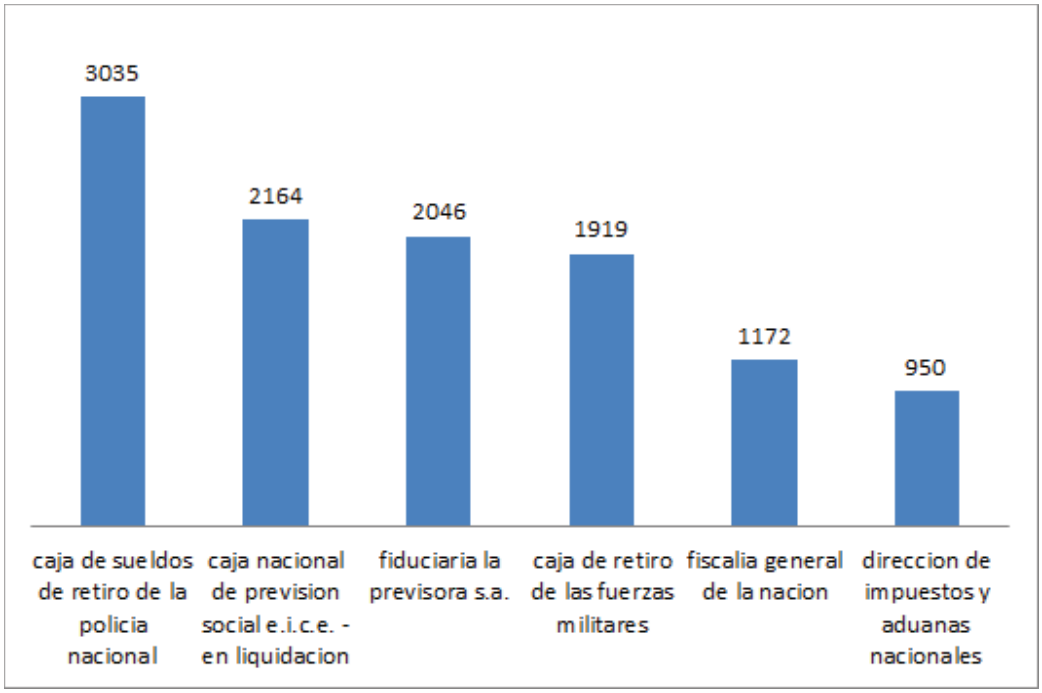
Alegatos: La presentación de alegatos por parte del demandante se relaciona con una mayor probabilidad de que le sea otorgada la pretensión. Si bien esta variable había sido identificada como irrelevante por el algoritmo *InfoGain* para el primer modelo, en el caso de procesos de orden nacional sí tiene influencia. Cabe notar que para el modelo adecuado a entidades de orden nacional no fue necesario aplicar el algoritmo *InfoGain* ya que la dimensión de los datos es menor; sin embargo, esto no altera los resultados pues una variable identificada como no relevante por ese algoritmo también será descartada en la selección realizada con *Boosting* de árboles.

Por otro lado, si es el demandado quien presenta alegatos, la probabilidad de que el Estado pierda aumenta. Mientras en el primer caso P_0 es de 54 %, en este segundo alcanza 57 %. Esto no necesariamente implica que presentar alegatos perjudique al

Estado; es posible que esta relación se deba a que generalmente la presentación de alegatos por parte del demandado se da en casos en los que éste va perdiendo la demanda.

Entidad demandada: La base incluye demandas en contra de 60 entidades. El número de demandas en contra de cada una va desde cinco -como es el caso del Fondo Nacional del Ahorro, el DANE, y otras- hasta más de 3.000 en el caso de la Caja de Sueldos de Retiro de la Policía Nacional. La figura 11 muestra las entidades con un mayor número de demandas en su contra.

Figura 11: Entidades con un mayor número de demandas en su contra.



Sin embargo, la influencia en el modelo de probabilidad no depende del número de procesos, sino de la distribución de los procesos ganados y perdidos. De acuerdo al algoritmo de selección de variables utilizado, ocho entidades de orden nacional afectan la probabilidad de que el Estado pierda o gane una demanda. De estas, cinco tienden a perder las demandas y tres las ganan más que el promedio, como se muestra en el cuadro 2.

La Fiduciaria La Previsora S.A. es la que presenta un peor desempeño, con un coeficiente de 0.5335 y un P_0 de 63,03 %, seguida por el Ejercito Nacional de Co-

Cuadro 2: Entidades demandadas que influyen en la probabilidad de éxito

Entidad	Número de casos	P_0
Superintendencia de Servicios Públicos Domiciliarios	191	24.31 %
Ministerio de la Protección Social	218	30.61 %
Caja de Sueldos de Retiro de la Policía Nacional	3035	37.29 %
Caja Nacional de Prevision Social e.i.c.e. - en liquidación	2164	52.31 %
Fiscalía General de la Nación	1172	56.03 %
Caja de Retiro de las Fuerzas Militares	1919	61.62 %
Ejercito Nacional de Colombia	495	62.44 %
Fiduciaria la Previsora S.A.	2046	63.03 %

lombia, el cual tiene un coeficiente de 0.5081 y una probabilidad P_0 de 62,44 %.

Otras entidades cuyo comportamiento histórico perjudica al Estado son: Caja Nacional de Previsión Social e.i.c.e. - en liquidación, la Fiscalía General de la Nación y la Caja de Retiro de las Fuerzas Militares.

Por otro lado, la Caja de Sueldos de Retiro de la Policía Nacional, el Ministerio de la Protección Social y la Superintendencia de Servicios Públicos Domiciliarios tienen coeficientes que impactan la probabilidad de éxito de una demanda de forma que beneficia al Estado. El primero tiene un P_0 de 37,29 %, mientras el segundo tiene uno de 30,61 % y el último presenta el P_0 más bajo, de 24,31 %.

Tipo de demandante: Cuando quien demanda son varias personas naturales, la pretensión tiene las mayores probabilidades de ser otorgada, con un P_0 de 57 %, seguida por el caso en el que el demandante es una persona jurídica, el cual tiene un P_0 de 52 %. Los demás tipos de demandantes, como la persona natural o varias jurídicas, parecen no marcar una tendencia hacia el otorgamiento de la pretensión.

Presentación de excepciones: La presentación de excepciones de alguna de las partes influye levemente la probabilidad de éxito a favor del demandante, con P_0 de 53 %.

Monto pretendido: El valor promedio del monto pretendido en los procesos contra el Estado tiene una variación muy alta. Así como existen algunos procesos por debajo de un millón de pesos, hay otros que alcanzan en orden de 10^{10}. A medida que el monto pretendido por el demandante aumenta, esto se relaciona con una menor probabilidad de que se otorgue la pretensión. Sin embargo, el coeficiente es muy pequeño, del orden de 10^{-10}, por lo que no impacta significativamente la probabilidad de éxito.

Ciudad: En la regresión logística hay procesos de 33 ciudades diferentes. De éstas, nueve son seleccionadas por el algoritmo de *Generalized Boosted Models*, al considerar que tienen influencia en la probabilidad de éxito de un proceso. La ciudad en la que hay una mayor probabilidad de éxito para la Nación es Pereira, con un P_0 de 24 %, mientras en Bucaramanga los demandantes tienen una mayor probabilidad de que se otorgue su pretensión; allí el P_0 es de 73 %. La información detallada se incluye en el cuadro 3.

Cuadro 3: Ciudades que influyen en la probabilidad de éxito

Ciudad	Número de casos	P_0
Pereira	150	23.65 %
Villavicencio	675	38.16 %
Tunja	966	42.96 %
Cali	1131	52.84 %
Bogota	7244	54.96 %
Armenia	523	61.61 %
Santa Marta	204	62.64 %
Ibague	302	63.67 %
Bucaramanga	555	73.73 %

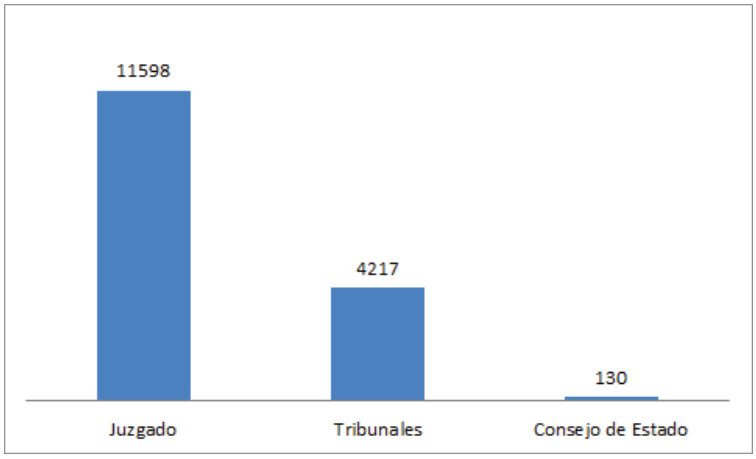
Corporación que preside: El proceso puede puede ser presidido por el Consejo de Estado, Tribunales o por un Juzgado. La figura 12 muestra el número de casos para cada una de éstas.

Para el modelo, el hecho de que se presente ante un tribunal o ante un juzgado tiene incidencia en la probabilidad. Cuando un juzgado lleva un proceso, el Estado tiene más probabilidades de perderlo. El coeficiente de esta variable es de 0.0114, por lo que tiene un P_0 de 50 %, mientras los tribunales disminuyen el P_0 a 40 %.

Magistrado/Juez: Como muestra el cuadro 4, hay cinco magistrados con influencia relativa en el modelo. Tres disminuyen la probabilidad de que se otorgue la pretensión mientras dos la incrementan. El impacto que tienen es bastante grande en algunos casos; en un extremo, Carlos Alberto Vargas Bautista tiene un P_0 de 12 %, mientras en el otro José Rodrigo Romero Romero eleva la probabilidad de que se otorgue la demanda a 70,84 %.

Cambio de representante del demandado: El número de veces que cambia de abogado el demandado tiene un impacto a favor de éste, haciendo que sus probabilidades de ganar la demanda aumenten. El coeficiente de esta variable es de 0.1244, y su P_0 es 53,11 %; esta última probabilidad corresponde a considerar un

Figura 12: Número de casos presididos por cada corporación.



Cuadro 4: Magistrados y jueces que influyen en la probabilidad de éxito

Magistrado/Juez	Número de casos	P_0
Carlos Alberto Vargas Bautista	103	12,07 %
Victor Manuel Buitago Gonzalez	81	20,73 %
Melva Giraldo Londono	103	33,05 %
Rafael Dario Restrepo Quijano	297	52,99 %
Jose Rodrigo Romero Romero	127	70,84 %

cambio de abogado y no tomar en cuenta las otras variables.

Número de pruebas: El número de pruebas tiene un coeficiente negativo (-0.0014), lo que significa que un alto número de pruebas se relacionan con una menor posibilidad de que se otorgue la pretensión. Sin embargo, el impacto en la probabilidad de éxito es muy bajo.

Motivo de la demanda: El motivo de la demanda parece tener un peso muy alto en la porbabilidad de que ésta se otorgue o no. Este motivo se entiende como la combinación entre la acción, el subtipo 1 y el subtipo 2. En algunos casos, sin embargo, la acción no tiene subtipo 2, o no tiene ningún subtipo. De los once motivos que influencian la probabilidad de éxito de un proceso, seis disminuyen la probabilidad de que se otorgue la pretensión y cinco la aumentan.

A continuación se hace una descripción del comportamiento de los motivos relevantes para el modelo, los cuales se agrupan bajo dos tipos de acción.

- **Reparación directa:** En general, aquellas demandas que tienen como subtipo 1 daños tienen una tendencia a no otorgarse, con un P_0 de 45,6 %. Sin embargo, dentro de los subtipos dos incluidos en este motivo que el modelo considera relevantes, solo uno incide en la tendencia en esta misma dirección: error jurisdiccional con un P_0 de 28 %. Por su parte, tanto los causados por la fuerza pública como los sufridos por conscripto tienen una mayor probabilidad de ser otorgados, al punto que el segundo caso tiene un P_0 % de 67 %.

  Por su parte, las demandas por una falla en prestación de servicio público, también incluidas en esta acción, tienen una tendencia a no ser otorgadas, con un P_0 de 34 %.

- **Nulidad y restablecimiento del derecho:** Esta es la otra acción que agrupa varios de los motivos de demanda considerados como relevantes para el modelo, algunos lo hacen de forma positiva para el Estado y otros de forma negativa.

  Las demandas por *impuestos nacionales*, de tipo *tributario*, tienen coeficiente positivo, por lo que favorece al demandante. Las demandas de tipo laboral por pensión y prestacional también inciden la probabilidad de éxito en contra del Estado, la primera con un P_0 de 68 % y la segunda con uno de 61 %. Sin embargo, estas no son las únicas demandas de tipo laboral influencian el modelo; aquellas por motivo salarial y las que se presentan por retiro del servicio tiende a ganarlas el Estado. La segunda es la que presenta un mejor panorama para el demandado, con un P_0 de 37 %.

  Por último, aquellas demandas por sanciones de subtipo 2 órganos de inspección y vigilancia tienen la tendencia más fuerte a ser ganadas por el Estado, con un P_0 de 16 %.

5.3. Regresión adecuada para el comienzo del proceso

Varias de las variables presentes en la base de datos solo se conocen al final del proceso o durante el transcurso de éste (como *duración* o *Apelación*, entre otras). Por lo tanto, se construyó un nuevo modelo incluyendo únicamente aquellas variables que se conocen desde el inicio del proceso, de modo que se pueda calcular una probabilidad de éxito para una demanda en el momento en el que ésta se recibe. Las características del proceso que se incluyeron son:

- Entidad demandada.
- Ciudad.

- Corporación que preside.
- Magistrado/juez.
- Tipo de demandante.
- Género del demandante.
- Abogado del demandante.
- Número de demandados.
- Subtipo 2.
- Monto que pretende el demandante.

La lista de las variables, con su influencia relativa, sus coeficientes y sus valores p y P_0 correspondientes están en el anexo VI.

El análisis de variables realizado para la regresión logística anterior sigue siendo válido ahora, pues si bien los coeficientes varían (al tener menos variables el modelo asigna diferentes pesos a cada elemento), la forma como cada variable impacta la probabilidad de éxito es similar (una variable que influencia el modelo anterior favoreciendo al Estado, también aumenta la probabilidad de que el Estado gane en este modelo, y lo mismo ocurre para aquellas que incrementan la probabilidad de que se otorgue la pretensión al demandante). Por lo tanto, resultaría redundante volver a analizar todas las variables, y solo vale la pena hacer una descripción del comportamiento de aquellas que no están presentes en la regresión anterior. Esto quiere decir que solo es necesario analizar el comportamiento de algunos abogados, unas entidades demandadas y un magistrado.

Los abogados que aparecen en este segundo modelo son exactamente los mismos que aparecen en el primero. Las entidades que resultan relevantes aquí también son las mismas, con excepción del Consejo Superior de la Judicatura que no aparecía antes. Esta entidad tiende a perder las demandas más que el promedio, con un coeficiente de 0.338 y un P_0 de 58 %.

Entre las ciudades, de nuevo tenemos las mismas del modelo general y se adicionan Popayán y Santa Rosa de Viterbo, ambas con una incidencia en la probabilidad que perjudica al Estado. La primera lo hace con un P_0 de 52 %, mientras la segunda eleva este valor hasta 67 %.

Entre los jueces, Rafael Darío Restrepo Quijano ya no aparece, y entra a formar parte del modelo Edda del Pilar Estrada Álvarez y Melva Giraldo Londoño. La primera con una influencia relativa que perjudica al Estado, con un coeficiente de 0.5,

mientras la segunda lo beneficia con uno de -0.9.

Por último, en el motivo de la demanda lo único que varía es que desaparece el subtipo 1 daños como variable influyente y lo reemplaza laboral, también un subtipo 1, con una incidencia que perjudica ligeramente al Estado, elevando P_0 a 52 %.

5.4. Entidades relevantes para la ANDJE que no tienen incidencia en el modelo

Hay ciertas entidades que son de particular interés para la agencia, pero que no tienen ninguna influencia relativa en el modelo. Entre ellas están el Ministerio de Educación Nacional, el Ministerio de Transporte, el Ministerio de Agricultura y Desarrollo Rural, el INPEC y la Dirección Ejecutiva de Administración Judicial. Es de interés de la ANDJE conocer el comportamiento de los procesos en contra de estas entidades, por lo que a continuación se presenta una breve descripción.

En el caso de los ministerios de Educación y de Agricultura, el número de procesos en el conjunto de datos utilizado para crear el modelo es muy bajo: ninguno supera los 30 casos. Por lo tanto, el modelo no puede extraer información concluyente del comportamiento de las demandas a estas entidades. De todas formas, cabe notar que la distribución de los fallos del Ministerio de Educación no presenta ninguna anomalía con respecto a la distribución general, pues 15 de los procesos se fallaron a favor del demandante. Sin embargo, de 28 procesos en contra del Ministerio de Agricultura, solo tres fueron fallados a favor del demandante.

Para todos los demás, las demandas tienen una ligera tendencia a fallarse a favor del Estado. De las 257 demandas en contra del INPEC, el 40,08 % se otorgaron al demandante. Un comportamiento similar presenta la Dirección Ejecutiva de Administración Judicial: de los 248 procesos en su contra, el 40,32 % fueron fallados en contra del Estado. Por último, el 38,01 % de las demandas en contra del Ministerio de Transporte las gana el demandante.

6. Árboles, Redes Neuronales y Comparación de Modelos

La regresión logística es un modelo de clasificación que tiende a dar buenos resultados. Sin embargo, hay muchos otros métodos de clasificación. Entre todos, no se puede decir cuál es mejor, solo cúal es mejor para cada caso. Mirar los datos y analizar el tipo de problema da indicios del tipo de modelos que pueden ofrecer mejores resultados, pero es necesario probar varios de ellos para determinar el mejor. En este capítulo se presentan dos métodos adicionales de clasificación y se hace una comparación cuantitativa entre todos.

En los tres métodos, la selección de variables con influencia relativa se realiza a partir del *boosting* de árboles, por lo que las variables son siempre las mismas, y corresponden a las expuestas en el capitulo cuatro.

El *boosting* de árboles, *Generalized Boosted models (gbm)*, utilizado para la selección de variables, también se puede aplicar como modelo de clasificación. Es decir, este método sirve para escoger los elementos que más pueden influenciar el resultado de un proceso, pero también sirve para clasificar los procesos. Éste se basa en una combinación de árboles de clasificación básicos. Partiendo de éstos métodos de clasificación débiles, logra construir un modelo fuerte al utilizarlos de forma iterativa (en este caso se construyen más de 1000 árboles).

El tercer método de predicción es el de redes neuronales. Esto se basa en una combinación no lineal de diferentes modelos logísticos para construir el método de clasificación.

A diferencia de la regresión logística, el *boosting* de árboles y las redes neuronales tienen un comportamiento de 'caja negra', en el sentido en que los resultados no se prestan para una interpretación de la sensibilidad de la probabilidad de éxito a cada variable. Por esta razón, para estos dos casos no se presenta un análisis de los resultados como en el primer informe.

6.1. Resultados

Existen diferentes estrategias para medir la precisión de un modelo. En este trabajo, se utiliza una base de entrenamiento y una de validación. Para esto, se particiona en dos (de forma aleatoria) la base que está siendo utilizada, la cual contiene 15,945 procesos; un segmento contiene el 70 % de los datos y el otro el 30 %.

La primera es llamada base de entrenamiento, y sirve para entrenar al modelo.

La segunda es la base de validación, la cual se utiliza para probar el modelo que resulta del aprendizaje de máquinas con la base de entrenamiento. De esta forma, se evalúa el comportamiento del modelo con datos que nunca ha "visto". Es importante notar que el modelo final se construirá utilizando la totalidad de los datos, por lo que su desempeño tendrá a una ligera mejora, pues habrá aprendido con más información.

Para medir la precisión de los tres métodos se utiliza la curva ROC, en particular, el área bajo la curva *(auc)*, la cual permite comparar la bondad de selección de los modelos. Ésta se obtiene para la base de entrenamiento y la de validación; la segunda ofrece una noción real del comportamiento del modelo cuando se enfrenta a nuevos procesos, tal como tendrá que hacerlo cuando sea utilizado por la ANDJE. Comparando con los resultados obtenidos para los datos de entrenamiento, se puede establecer el *overfitting* que hace el modelo; es decir, se determina qué tanto se influencia por particularidades de los datos con los que se construye el modelo.

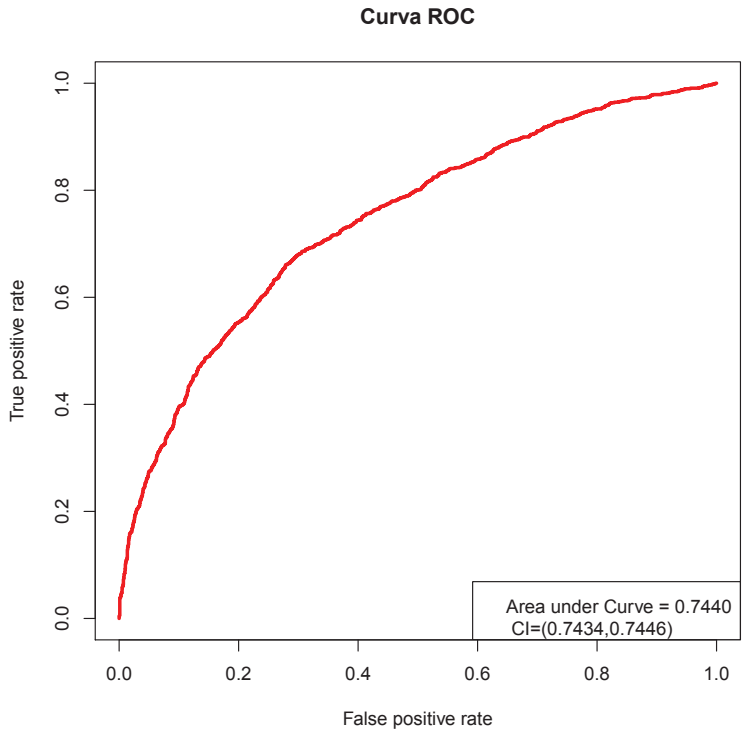
La gráfica 13 muestra la curva ROC de la regresión logística sobre la base de validación, y las figuras 14 y 15 muestran la curva ROC de *boosting* de árboles y redes neuronales, respectivamente. Adicionalmente, la tabla 5 compara los resultados de los diferentes métodos. La primera columna muestra el área bajo la curva ROC al evaluar el modelo sobre los datos de entrenamiento, y la segunda columna muestra esta área cuando se efectúan las predicciones sobre los datos de validación; este valor corresponde al desempeño del método cuando se enfrenta a nuevos procesos. Estas cifras resultan útiles para hacer la comparación. De esta tabla se concluye que el método *Generalized Boosted models* tiene el mejor desempeño para predecir el otorgamiento de la pretensión.

Cuadro 5: Comparación de *auc* para los tres métodos de clasificación.

Modelo	área ROC - Entrenamiento	área ROC - Validación
Logit	0.755	0.744
Boosting	0.8054	0.7626
Redes neuronales	0.7691	0.741

Dado que el *boosting* de árboles es el mejor modelo de probabilidad para los datos de la ANDJE, se construye un segundo modelo predictivo adecuado al comienzo del proceso, tal como se hizo con la regresión logística. En este caso, los resultados son consistentes con los presentados para el caso nacional general. En la figura 16 se muestra la curva ROC para el modelo de árboles y en la figura 17 la misma curva para el modelo logístico.

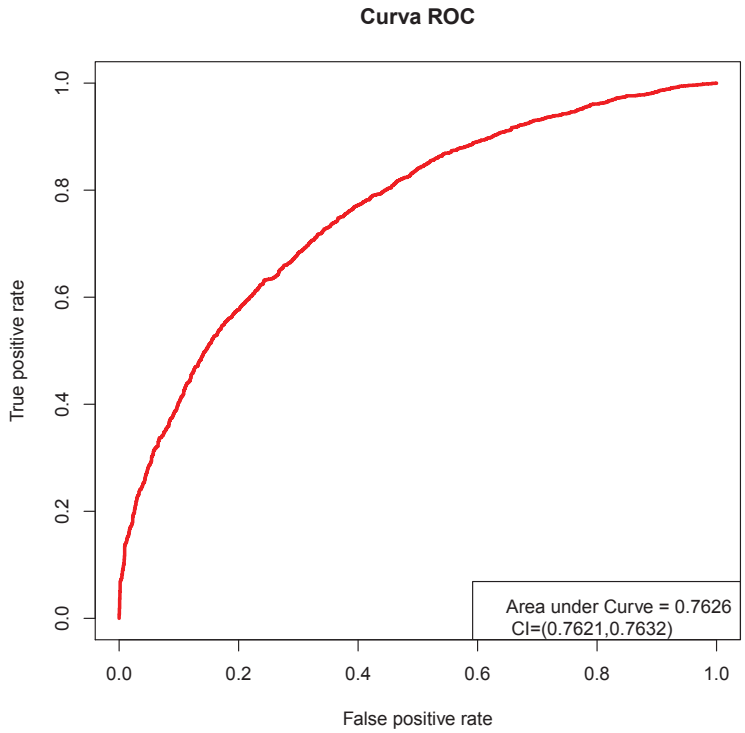
Figura 13: Curva ROC para la regresión logística nacional, calculada sobre los datos de validación.



7. Aprendizaje semi-supervisado

Hasta este momento, solo se han utilizado los datos marcados, por lo que hay alrededor de 73.000 procesos contra entidades de orden nacional que no se han incluido en ningún modelo por no tener información sobre si se otorgó o no la pretensión. Los modelos de aprendizaje semi-supervisado permiten aprovechar la información que éstos aportan esperando mejorar clasificadores supervisados. La idea en la que se fundamentan estos métodos es la de utilizar los datos marcados para predecir la decisión judicial de los no marcados, y luego utilizar estos datos para construir el modelo predictivo. No siempre mejoran los resultados; esto depende de varios factores como la similitud entre unos datos y otros y la cantidad de datos marcados y no marcados que se tengan. En ocasiones en que no se cuenta con un alto número de datos marcados, es posible que no se logren predecir con mucha precisión las etiquetas de los no marcados, por lo que el modelo final incluirá muchos errores. Sin embargo, lo contrario también puede pasar. El objetivo del análisis semi-supervisado es encontrar nuevos patrones e información que permitan construir un modelo más

Figura 14: Curva ROC para *boosting* de árboles, calculada sobre los datos de validación.

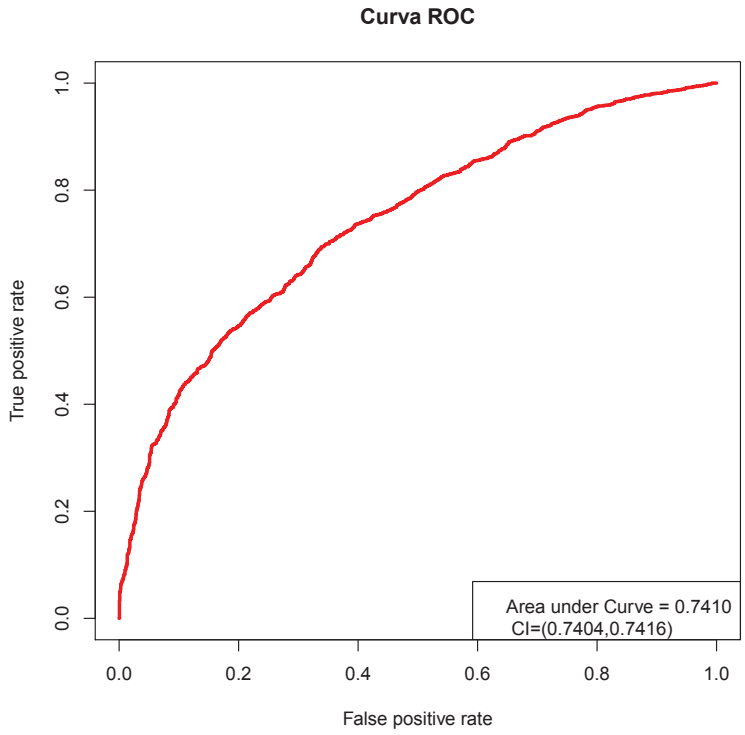


sólido; sin embargo si la base marcada ya tiene suficiente información, y no es considerable lo que se pueda aprender de los nuevos datos, los métodos semi-supervisados pueden ingresar ruido al modelo, en lugar de hacerlo más robusto. Como se verá, en el presente trabajo los métodos semi-supervisados no mejoran el desempeño del mejor método supervisado.

Entre los métodos utilizados se encuentra *Chopper*, técnica que se basa en utilizar los datos marcados para predecir las etiquetas de los no marcados, asignándoles una probabilidad de otorgamiento de la pretensión. Se fija un nivel de corte, y la porción con los pronósticos más contundentes (probabilidades cerca a 100 % o 0 %) se suma a los datos marcados. Este proceso se realiza de forma iterativa, hasta haber marcado todos los datos.

Esta técnica semi-supervisada se combina con un método de clasificación por medio de árboles basado en el algoritmo $C4{,}5$ de Ross Quinlan. En cada rama del árbol, este método escoge las variables que le permiten predecir mejor la variable

Figura 15: Curva ROC para redes neuronales, calculada sobre los datos de validación.

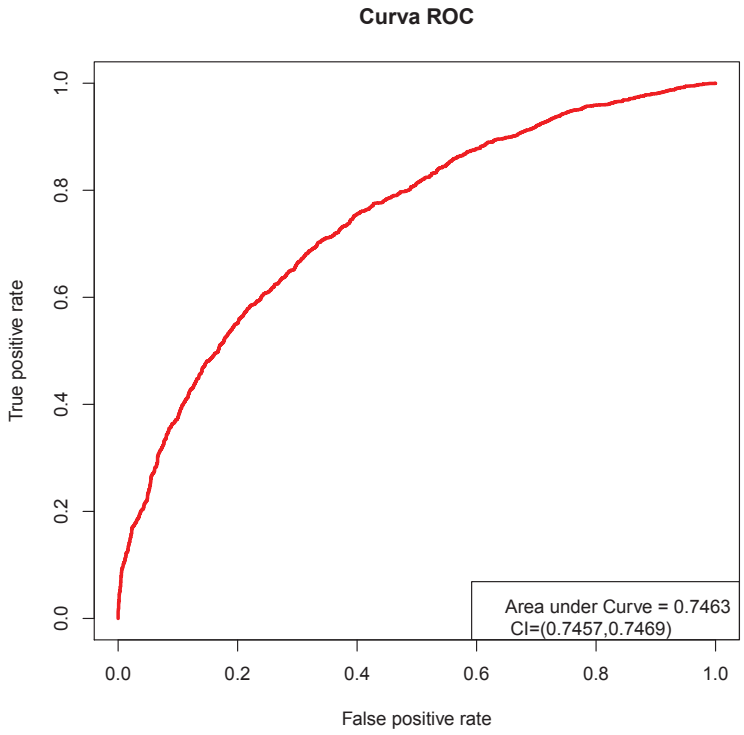


objetivo (en este caso, si se otorga o no la pretensión). Para calificar el poder de predicción de las variables se toma la entropía de información, una medida de ganancia de información que analiza la diferencia del comportamiento de cada una de las variables cuando se toma una clase o la otra. Esta es la misma entropía utilizada en el algoritmo *InfoGain* mencionado en el capítulo cuatro. Este modelo de clasificación también se combina con otra técnica semi-supervisada conocida como *Collective EM*, la cual realiza procesos iterativos basados en la distribución de los datos.

Adicionalmente, el algoritmo de Quinlan se combina con la técnica semi-supervisada *Two Stage Collective*. Ésta predice las etiquetas de los datos no marcados entrenando sobre los marcados, y luego vuelve a aplicar el modelo de clasificación de árboles utilizando todos los datos.

Estas tres técnicas semi-supervisadas también se combinan con el algoritmo supervisado *Random Forest*, el cual crea varios árboles que predicen la variable objetivo

Figura 16: Curva ROC para *boosting* de árboles adecuado al comienzo del proceso, calculada sobre los datos de validación.

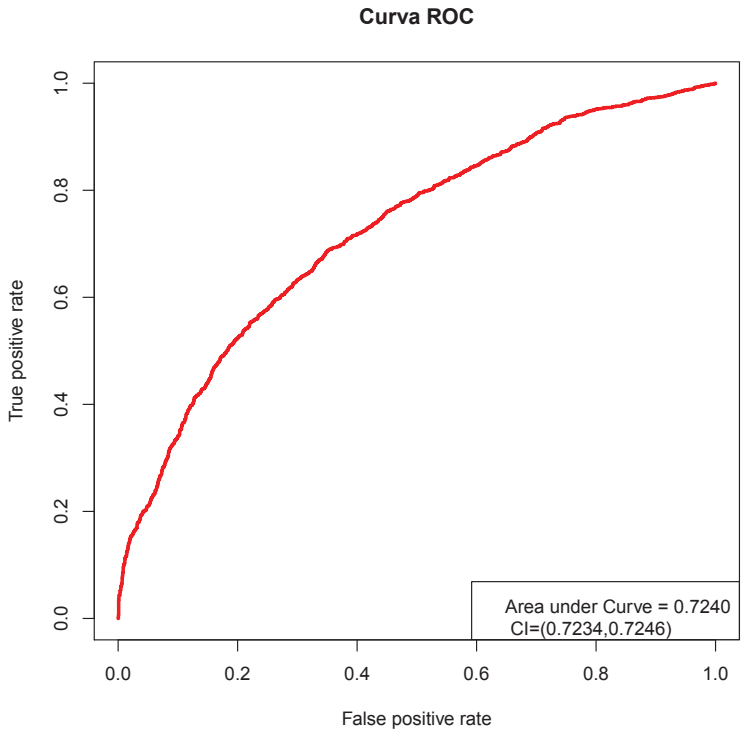


para un proceso, y luego realiza un promedio entre los resultados de todos los árboles. Este promedio puede ser entendido como una "votación" entre todos los árboles, después de la cual se decide la clase a la que pertenece el proceso.

Otra forma de hacer aprendizaje semi-supervisado es basado en métodos no supervisados. Para esto se aplica *Collective Tree*. Este algoritmo busca reglas que dividan los datos en dos partes de tamaño semejante, sin tener en cuenta las marcas de éstos. Este proceso lo realiza de forma iterativa, hasta que llega a ramas en las que sólo hay datos marcados de una clase; es decir, que todos los procesos en esa rama hayan sido otorgados o ninguno lo haya sido. En este punto, a todos los demás procesos en esa rama se les asigna esa misma marca. También puede ocurrir que llegue a ramas en las que todos los procesos ya están marcados, información de la que extrae una probabilidad. Asimismo, puede llegar a una rama final en la que todos los datos son no marcados, caso en el cual se toma en cuenta el nodo anterior.

En la tabla 6 se muestran los resultados de los modelos, tomando como variable

Figura 17: Curva ROC para regresión logística adecuada al comienzo del proceso, calculada sobre los datos de validación.



objetivo que no se otorgue la pretensión al demandante. La comparación se hace igual a como se hizo en el caso de aprendizaje supervisado, midiendo el área bajo la curva ROC sobre los datos de validación.

Cuadro 6: Comparación de *auc* para los métodos semi-supervisados

Modelo	área ROC - Validación
Chopper + C4.5	0.655
Chopper + Random Forest	0.747
Collective EM + C4.5	0.665
Collective EM + Random Forest	0.722
Two Stage Collective + C4.5	0.684
Two Stage Collective + Random Forest	0.746
Collective Tree	0.731

Como se puede ver, el modelo que da el mejor resultado es el que combina el método de clasificación *Random Forest* con la técnica *Chopper*. Sin embargo, su desempeño es similar al de las redes neuronales y la regresión logística, sin lograr superar el *boosting* de árboles.

8. Herramientas Computacionales

Esta entrega incluye dos herramientas que permiten predecir la probabilidad de éxito de otros procesos en contra del Estado. La primera, entregada en Excel, corresponde al modelo de regresión logística, y la segunda permite predecir con el modelo de *boosting* de árboles por medio de una interfaz para R Project.
La herramienta de Excel incluye tres regresiones. La primera corresponde al modelo general, construido con todos los casos marcados de la base; la segunda se enfoca en procesos en contra de entidades de orden nacional, por lo que tiene una mayor precisión para este tipo de procesos. Por último, un modelo permite asignar una probabilidad de éxito a un proceso en el momento en el que éste comienza.
Quantil sugiere utilizar el primer modelo únicamente en los casos en los que la entidad demandada no es de orden nacional. Cuando la entidad sí es de este orden, se recomienda utilizar el modelo ajustado al comienzo del proceso para conocer una probabilidad inicial, y hacer uso de la segunda regresión a medida que éste avanza, para poder determinar cómo los acontecimientos han influido en las probabilidades de que se otorgue la pretensión al demandante.
Dado que el *boosting* de árboles presentó el mejor desempeño, se entrega una segunda herramienta que permite hacer uso de este modelo. Dada la complejidad del algoritmo, no es posible implementarlo en Excel, por lo que se construyó una interfaz que facilita el uso de la herramienta por parte de los analistas de la ANDJE. El único requerimiento técnico para que se pueda utilizar en un computador es que éste tenga instalado R Project, un software de distribución libre.
La interfaz permite que el usuario introduzca los mismos datos que pide el modelo logístico de Excel, y le da la opción de ver la probabilidad correspondiente a un proceso que está iniciando o a uno que ya está avanzado.

9. Instalación de la interfaz para utilizar modelo predictivo basado en *GBM*

La interfaz entregada a la ANDJE para ser utilizada por la Agencia se puede accesar directamente desde el escritorio del computador del usuario, haciendo doble click sobre el ícono de Quantil. El proceso previo para instalarla en el computador se describe a continuación:

1. Se debe descargar el programa gratuito R Project 3.0.2 (o versión más reciente).

2. La carpeta *Interfaz* entregada por Quantil se debe guardar en el computador en el que se quiere instalar la herramienta. Para este ejemplo se supondrá que se guarda en el directorio "C:\\Users\\Documents".

3. En el script *run*, ubicado dentro de la carpeta *Interfaz* se debe abrir el script, reemplazar el valor de folder_address por la ruta de la carpeta *Interfaz*, y guardar. En el caso del ejemplo sería "C:\\Users\\Documents\\Interfaz".

4. En el script *server*, ubicado dentro de la carpeta *Interfaz* se debe abrir el script, reemplazar el valor de setwd por la ruta de la carpeta en la cual está contenida *Interfaz*, y guardar. En el caso del ejemplo sería "C:\\Users\\Documents\\".

5. El archivo de tipo .bat llamado *Quantil-Modelo predictivo* se debe editar. Para esto, se hace click derecho sobre el ícono y se selecciona la opción de editar. La primera ruta debe corresponder al ejecutable de R Project. Con frecuencia esta ruta es "C:\Program Files\R\R-3.0.2\bin\x64\R.exe"

Nota: En caso de que haya un problema con la instalación de los paquetes (incluida dentro del código), se debe abrir R Project desde *Inicio* o desde el acceso directo en el escritorio, y ejecutar los siguientes comandos en la ventana principal de R Project:

- install.packages("shiny")
- install.packages("caret")
- install.packages("gbm")

10. Conclusiones

Este documento expone el trabajo realizado para analizar la base de datos de los procesos de demanda contra el Estado del año 2012, compilada por el Banco Mundial, que contiene cerca de 150.000 datos en total, 27.000 de los cuales estaban marcados como exitosos o no exitosos para el Estado. El objetivo principal de buscar un modelo probabílistico para pronosticar la probabilidad de éxito de una demanda particular contra el Estado se atacó planteando un conjunto amplio de modelos de clasificación, para escoger así el más adecuado, siguiendo criterios de comparación basados en la curva ROC, que exhibe, para cada tasa de falsos positivos, la tasa de verdaderos positivos del modelo.

Los modelos desarrollados consisten en regresión logística, *boosting* de árboles, y redes neuronales, para el caso de análisis supervisado (que reduce el análisis a los casos marcados), y un conjunto amplio para el caso de análisis semi-supervisado (que iterativamente va amrcando los casos no marcados para aumentar la base marcada). El análisis se realiza con dos consideraciones: primero, incluyendo solo los procesos ante instituciones de nivel nacional, caso de especial interés para la ANDJE; segundo, considerando solo las variables conocidas en el momento en que comienza un proceso, para permitir al usuario calcular la probabilidad de éxito de un proceso que está en sus estados iniciales.

Consistentemente el modelo de *boosting* de árboles presenta los mejores resultados, usando el área bajo la curva ROC como la fuente de comparación entre modelos. En el caso de variables conocidas al comienzo, esta área es de 74,6 %, y para el caso general es de 76,2 %. Estas cantidades evidencian buenos ajustes del modelo, y otorgan un buen nivel de confianza a un usuario del modelo.

Tanto el modelo de regresión logística (de fácil implementación e interpretación) como el modelo de *boosting* de árboles se presentan como parte integral de esta entrega, en la forma de herramientas computacionales que permiten a un usuario calcular la probabilidad de éxito de un proceso dadas sus características.

Finalmente, el trabajo incluye un análisis de reglas de asociación, identificando vlaores de distintas variables que tienden a aparecer en conjunto más de lo que un análisis de independencia sugeriría. Esto puede permitir a analistas de la ANDJE reconocer patrones de difícil reconocimento con una simple observación de los datos. En adición, este tipo de análisis permite identificar situaciones posiblemente anómalas, ofreciendo así una herramienta eficiente para priorizar investigaciones más profundas de situaciones que llamen la atención.

Referencias

[1] Ridgeway, Greg. y H.D. Sherali, *Generalized Boosted Models: A guide to the gbm package.*, Update 1 (2007): 1.

[2] Friedman, Jerome H. y Greedy function approximation: a gradient boosting machine., Annals of Statistics (2001): 1189-1232.

[3] Freund, Yoav, and Robert E. Schapire. y Special invited paper. Additive logistic regression: A statistical view of boosting: Discussion., The Annals of Statistics 28.2 (2000): 391-393.

[4] Bradley, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 30.7 (1997): 1145-1159.

[5] Quinlan, Ross. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. (1993).

[6] Breiman, Leo. Random Forests. Machine Learning. 45(1):5-32. (2001).

[7] Pfahringer; Bernhard, Driessens, Kurt, et al. Collective Classifiers. 45(1):5-32. (2001).

11. Anexos

Anexo I: Definiciones

- **Aprendizaje de máquinas:** Hace referencia a la programación de una máquina para que ésta pueda aprender de forma automática, de modo que sea capaz de hacer predicciones correctas a partir de una serie de observaciones.

- **Apredizaje supervisado:** Tipo de aprendizaje de máquinas en el cual se conoce la variable objetivo, en este caso la decisión de si se otorga o no la pretensión.

- **Generalized Boosted Models (GBM):** En este caso, este algoritmo determina la importancia de cada variable dentro del modelo. A cada una le asigna una influencia relativa (que en muchos casos es cero), la cual determina qué tanta influencia tiene en el resultado del proceso cada una de las variables. Una influencia relativa alta indica que la presencia de esa variable (puede ser un abogado, una ciudad, un motivo de demanda, etc.) tiene un gran impacto en la probabilidad de éxito del modelo. En general, GBM es un método basado en el uso de árboles, que construye un modelo de predicción fuerte basado en la combinación de muchos métodos de predicción débiles.

- ***InfoGainAttributeEval:*** El objetivo de este algoritmo es determinar la ganancia de información para cada variable. La entropía, utilizada para realizar esta medición, se puede entender como el desorden -o viéndolo en sentido inverso, como el nivel de predictibilidad-. Así, al medir la diferencia entre la entropía de una clase, y la entropía de la clase dado determinado atributo, se está cuantificando cuánta información aporta ese atributo para poder determinar el resultado de la incognita que queremos resolver. En este caso, la clase hace referencia al resultado del fallo y el atributo a la variable con la cual intentamos predecir ese resultado.

$$InfoGain(Clase, Atributo) = H(Clase) - H(Clase|Atributo),$$

Anexo II: Preselección de acuerdo al número de casos

Cuadro 7: Preselección de acuerdo a número de casos

$>$ 25,000 marcados $\bigcap >$ 120,000 no marcados	$>$ 20,000 marcados $\bigcup >$ 90,000 no marcados
departamento	c_gen_dmte
ciudad	subtipo2
c_cons_corporacion	c_conci
cons_idcorporacion	gastos_proc
c_municipio_pres	f_rad_cont_dmdo
corp_pres	f_elab_pruebas
Magistrado	f_not_pruebas
c_dmte	n_documentales
abogado_tarjeta1	c_alegatos_dmte
c_dmdo	c_alegatos_dmdo
c_demdo_1	c_ap_1_tipo
n_demdos	f_rad_ap_1
n_cambiorep_dmte	f_info_sec_reparto1
n_cambiorep_dmdo	f_elab_admision_mand
c_tipo_proc	f_not_admision_mand
accion	f_notif_dmdo
subtipo1	
c_monto_salario_pret	
f_rad_esc_dem	
f_reparto1	
c_impedimento	
c_litisconsorcio	
c_suspension	
c_aclaracion	
c_llamamiento	
c_excepciones	
f_ult_act	
c_tipo_ult_act	
c_parte_ult_act	

Anexo III: Ranking por ganancia de información para el modelo con las observaciones de procesos contra el Estado

Cuadro 8: Ranking por ganancia de información

Rank	Atributo
0.1249025	entidad demandada
0.0923374	abogado dmte
0.0810997	subtipo2
0.0772757	abogado
0.0622423	ciudad
0.0564831	magistrado
0.0548992	municipio
0.0533935	departamento
0.0317285	duracion
0.0240224	subtipo1
0.0203457	apelacion
0.0173255	accion
0.0074724	monto pretendido
0.0066112	corporación que preside
0.0056158	suspension
0.0050242	número de pruebas
0.0042021	llamamiento
0.0032048	litisconsorcio
0.0026841	Contestacion
0.0023032	número cambio representante dmte
0.0018739	aclaracion
0.0008435	demdos
0.0005964	alegatos dmdo
0.0004266	dmte
0.0002851	excepciones
0.0000528	género dmte
0	alegatos dmte
0	número cambio representante dmdo

Anexo IV: Influencia relativa en los procesos contra entidades de orden nacional, de acuerdo al *boosting* de árboles

Variable	rel.inf (GBM)
Duracion	11.0178627
Apelacion_No	10.9110995
subtipo2_Pension (Nulidad y restablecimiento derecho)	9.66233233
Alegatos demandado	6.93553469
entidad_demandada_caja_de_sueldos_de_retiro_de_la_policia_nacional	6.01068248
Monto pretendido	5.95119752
entidad_demandada_caja_de_retiro_de_las_fuerzas_militares	5.1695054
entidad_demandada_fiduciaria_la_previsora_sa	4.16358768
ciudad_Bogota	2.43129258
abogado_112907	2.11634887
numero_pruebas	1.93879706
ciudad_Bucaramanga	1.76810427
subtipo2_Retiro_del_servicio (Nulidad y restablecimiento derecho)	1.69331222
abogado_31571	1.47499629
abogado_170560	1.42065
corporacion_Juzgado	1.40029416
subtipo2_organos_de_inspeccion_y_vigilancia (Nulidad y restablecimiento derecho)	1.31406471
subtipo2_Salarial (Nulidad y restablecimiento derecho)	1.2203316
abogado_109557	1.11515053
ciudad_Tunja	1.06002704
ciudad_Pereira	1.0130112
abogado_83363	0.90281871
cambio_representante_demandado	0.8821744
corporacion_Tribunales	0.86104177
Magistrado_Rafael_Dario_Restrepo_Quijano	0.81202084
ciudad_Ibague	0.78468251
abogado_31614	0.7118794
subtipo2_Error_jurisdiccional (Reparacion directa)	0.70738874
entidad_demandada_ejercito_nacional_de_colombia	0.58540871
subtipo1_Daños (Reparacion directa)	0.57958096

Variable	rel.inf (GBM)
ciudad_Villavicencio	0.54167937
Magistrado_Carlos_Alberto_Vargas_Bautista	0.52880923
Magistrado_0	0.51846887
abogado_56834	0.46695568
abogado_59415	0.46301359
alegatos_demandante	0.4611484
Apelacion_Sentencia	0.42991127
subtipo2_Prestacional (Nulidad y restablecimiento derecho)	0.41719056
ciudad_Armenia	0.40881332
entidad_demandada_superintendencia_de_servicios_publicos_domiciliarios	0.37380518
entidad_demandada_caja_nacional_de_prevision_social_eice___en_liquidacion	0.37181928
aclaracion	0.34862054
entidad_demandada_ministerio_de_la_proteccion_social	0.34683169
abogado_158718	0.3278574
subtipo2_Causados_por_la_fuerza_publica (Reparacion directa)	0.3085813
abogado_90682	0.3021746
subtipo2_Impuestos_nacionales (Nulidad y restablecimiento derecho)	0.2991559
subtipo2_Sufridos_por_conscripto (Reparación directa)	0.28890076
Magistrado_Jose_Rodrigo_Romero_Romero	0.28385408
abogado_41146	0.23553065
ciudad_Cali	0.23412538
subtipo2_Prestacion_de_servicios_publicos (Reparación directa)	0.23225852
demandante_Persona_juridica	0.22944018
excepciones	0.22130579
abogado_95908	0.21731833
Magistrado_Melva_Giraldo_Londono	0.21493879
entidad_demandada_fiscalia_general_de_la_nacion	0.20766018
abogado_98987	0.20278654
Magistrado_Victor_Manuel_Buitago_Gonzalez	0.16105374
abogado_45113	0.1552666
ciudad_Santa_Marta	0.15292928
abogado_91183	0.150615
abogado_otro	0.14438609
demandante_Varias_naturales	0.14023029

Cuadro 9: Anexo V - Regresión para todas las variables seleccionadas en los procesos contra el Estado

Variable	GBM-rel.inf	logit (coeficientes)	Pr($>\|z\|$)	P_0
(Intercept)		0.05205513	0.513011	
Duracion	20.52874981	-0.000212262	< 2,00E-16	0.512958
subtipo2_Salarial (Nul. y res. del derecho)	13.3340833	-1.025973627	< 2,00E-16	0.274100
subtipo2_Pension (Nul. y res. del derecho)	12.61648991	0.589626115	< 2,00E-16	0.655133
Apelacion_No	8.8291263	0.328727165	0.029628	0.594062
accion_Ejecutivos	7.97060398	1.177457656	1.52E-11	0.773733
subtipo2_Retiro_del_servicio (Nul. y rest. del derecho)	7.14917593	-0.821156344	< 2,00E-16	0.316674
Apelacion_Sentencia	6.0239332	-0.316582732	0.030583	0.434251
alegatos_demandado	3.14627121	0.40642813	< 2,00E-16	0.612654
monto_pretendido	2.74049146	-5.83E-11	0.000574	0.513011
entidad_demandada_departamento_antioquia	1.79911018	2.916298466	1.85E-10	0.951124
subtipo2_organos_de_inspeccion_y_vigilancia (Nul. y res. del derecho)	1.40624573	-1.523945974	6.24E-10	0.186655
entidad_demandada_5departamento_atlantico	1.34446714	3.114359178	1.20E-07	0.959551
numero_cambios_representante_dmte	1.28235543	0.204908571	8.75E-07	0.563890
suspension	1.0645162	-0.244063713	5.64E-07	0.452145
subtipo2_Error_jurisdiccional (Rep. directa)	1.04137673	-1.300540198	4.24E-12	0.222962
subtipo1_Servicios_publicos (Nul. y res. del derecho)	0.80955025	-2.409055071	1.53E-09	0.086511
entidad_demandada_direccion_de_impuestos_y_aduanas_nacionales	0.70902341	0.348552963	1.34E-05	0.598834
entidad_demandada_cauca_otro	0.62698058	0.884842776	6.97E-05	0.718473
subtipo2_Sufridos_por_conscripto (Rep. directa)	0.59025623	0.519232021	0.007107	0.639060
entidad_demandada_boyaca_otro	0.53505411	0.87699034	3.62E-08	0.716882
entidad_demandada_departamento_valle_del_cauca	0.41082582	0.743077529	3.54E-05	0.688932
numero_pruebas	0.37008278	-0.004007746	0.001692	0.512010
subtipo2_Asuntos_urbanisticos (Nul. y res. del derecho)	0.35208747	-1.481788613	1.18E-05	0.193140
abogado_109557	0.33036944	-1.089195	5.24E-06	0.261702

Variable	GBM-rel.inf	logit (coeficientes)	Pr($>\|z\|$)	P_0
entidad_demandada_valle_del_cauca_cali	0.27692477	-0.740153526	0.000138	0.334456
entidad_demandada_superintendencia_de_industria_y_comercio	0.26135656	0.924508284	0.002922	0.726426
entidad_demandada_cundinamarca_otro	0.24275792	0.960784301	0.00021	0.733575
entidad_demandada_atlantico_barranquilla	0.23798089	1.177629469	5.24E-06	0.773763
entidad_demandada_departamento_santander	0.21954385	-0.868844308	0.000162	0.306446
abogado_170560	0.20795684	-0.328799161	0.018861	0.431252
llamamiento	0.19810198	-0.280113345	0.001315	0.443231
ciudad_Buenaventura	0.19191456	1.109961834	0.002315	0.761699
entidad_demandada_policia	0.18719929	-0.31969611	0.002514	0.433486
subtipo1_Sistema_de_seguridad_social_en_salud (Nul. y res. del derecho)	0.18165944	-1.809552931	0.000182	0.147104
abogado_122470	0.18052222	-1.087095986	0.004259	0.262108
subtipo1_Sanciones (Nul. y res. del derecho)	0.16911993	-0.141306932	0.267001	0.477702
subtipo2_Causados_por_particulares_e_imputables_al_Estado (Rep. directa)	0.14831896	-0.746515045	7.33E-05	0.333042
Magistrado_Victor_Manuel_Buitago_Gonzalez	0.12694247	-0.496918615	0.001732	0.390583
Magistrado_Alfredo_Vargas_Morales	0.12652242	-1.337558279	0.000459	0.216615
entidad_demandada_magdalena_otro	0.12344893	0.645358458	0.015086	0.667614
entidad_demandada_departamento_administrativo_de_seguridad	0.12167039	0.77574757	0.0132	0.695890
entidad_demandada_secretaria_de_educacion	0.11314211	-0.148926064	0.154805	0.475801
ciudad_Cartagena	0.10470414	0.379334136	0.000898	0.606205
entidad_demandada_superintendencia_de_servicios_publicos_domiciliarios	0.0968232	0.592950717	0.000194	0.655884
subtipo2_Prestacion_de_servicios_publicos (Rep. directa)	0.09082928	-0.45414409	0.002453	0.400811
subtipo1_Tributario (Nul. y res. del derecho)	0.09042891	0.076292714	0.356577	0.532043
subtipo2_Colectivo (Nul. y res. del derecho)	0.08826144	-0.879628336	0.007104	0.304158
subtipo2_Impuestos_distritales (Nul. y res. del derecho)	0.08824492	0.48395035	0.031508	0.630883
litisconsorcio	0.08300251	-0.157309683	0.059983	0.473711
Magistrado_Jose_Rodrigo_Romero_Romero	0.07792304	0.320614746	0.098774	0.592104

Cuadro 10: Anexo V - Regresión para procesos de orden nacional(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Variable	logit (coeficientes)	z value	Pr($>\|z\|$)	P_0
(Intercept)	-0.9519	0.003479	**	0.27850288
Duracion	-0.0001	0.025482	*	0.499979
Apelacion_No	0.8963	1.75E-06	***	0.71018856
subtipo2_Pension (Nulidad y restablecimiento del derecho)	0.7710	1.69E-13	***	0.68373717
alegatos_demandado	0.2739	2.44E-06	***	0.5680501
entidad_demandada_caja_de_sueldos_de_retiro_de_la_policia_nacional	-0.5200	3.35E-06	***	0.37285223
monto_pretendido	0.0000	0.44923		0.5
entidad_demandada_caja_de_retiro_de_las_fuerzas_militares	0.4736	1.78E-05	***	0.61623547
entidad_demandada_fiduciaria_la_previsora_sa	0.5335	4.02E-06	***	0.63029906
ciudad_Bogota	0.1992	0.008953	**	0.54963598
abogado_112907	3.2010	1.55E-05	***	0.96087189
numero_pruebas	-0.0014	0.472509		0.499654
ciudad_Bucaramanga	1.0320	1.97E-11	***	0.73730345
subtipo2_Retiro_del_servicio (Nulidad y restablecimiento del derecho)	-0.5052	0.00025	***	0.37631943
abogado_31571	-1.0390	8.61E-08	***	0.26134299
abogado_170560	1.0580	8.34E-07	***	0.74230816
corporacion_Juzgado	0.0114	0.962688		0.50285247
subtipo2_organos_de_inspeccion_y_vigilancia (Nulidad y rest. del derecho)	-1.5440	1.24E-05	***	0.17595454
subtipo2_Salarial (Nulidad y restablecimiento del derecho)	-0.3032	0.043131	*	0.4247754
abogado_109557	1.9230	3.81E-07	***	0.8724726
ciudad_Tunja	-0.2835	0.032841	*	0.42959591
ciudad_Pereira	-1.1720	7.97E-06	***	0.23649367
abogado_83363	-1.5850	0.000134	***	0.17008853
numero_cambio_rep_demandado	0.1244	0.00062	***	0.53105995
corporacion_Tribunales	-0.3976	0.093226	.	0.4018891
Magistrado_Rafael_Dario_Restrepo_Quijano	0.1197	0.503644		0.52988932
ciudad_Ibague	0.5612	0.008032	**	0.63673015
abogado_31614	-1.1120	3.93E-06	***	0.24749821
subtipo2_Error_jurisdiccional (Reparacion directa)	-0.9283	0.00051	***	0.28326973
entidad_demandada_finalejercito_nacional_de_colombia	0.5081	0.000981	***	0.62436096
subtipo1_Daños (Reparacion directa)	-0.1757	0.125843		0.45618765

Variable	logit (coeficientes)	z value	Pr($>\|z\|$)	P_0
ciudad_Villavicencio	-0.4828	0.006854	**	0.38159116
Magistrado_Carlos_Alberto_Vargas_Bautista	-1.9860	0.007573	**	0.12068069
Magistrado_0	0.0785	0.250467		0.51960245
abogado_56834	-1.0940	9.61E-05	***	0.2508658
abogado_59415	-0.6349	0.019469	*	0.34640031
alegatos_dmte	0.1560	0.012131	*	0.5389211
Apelacion_Sentencia	0.0230	0.895306		0.50574725
subtipo2_Prestacional (Nulidad y restablecimiento del derecho)	0.4591	0.000119	***	0.61280065
ciudad_Armenia	0.4731	0.010944	*	0.61611722
entidad_demandada_superintendencia_de_servicios_publicos_domiciliarios	-1.1360	0.00017	***	0.24305552
entidad_demandada_caja_nacional_de_prevision_social_eice___en_liquidacion	0.0925	0.394595		0.52310104
aclaracion	0.0114	0.908648		0.50283997
entidad_demandada_ministerio_de_la_proteccion_social	-0.8186	0.000678	***	0.30606092
abogado_158718	-0.9113	0.000434	***	0.28673389
subtipo2_Causados_por_la_fuerza_publica (Reparacion directa)	0.3563	0.035246	*	0.58814447
abogado_90682	1.0970	0.001378	**	0.74969757
subtipo2_Impuestos_nacionales (Nulidad y restablecimiento del derecho)	0.5594	0.003793	**	0.6363137
subtipo2_Sufridos_por_conscripto (Reparacion directa)	0.7264	0.008251	**	0.67401478
Magistrado_Jose_Rodrigo_Romero_Romero	0.8874	0.000615	***	0.70835333
abogado_41146	0.7939	0.010245	*	0.68866812
ciudad_Cali	0.1136	0.309545		0.5283695
subtipo2_Prestacion_de_servicios_publicos (Reparacion directa)	-0.6318	0.030241	*	0.34710251
dmte_Persona_juridica	0.0812	0.497823		0.52028885
excepciones	0.1319	0.01995	*	0.53292728
abogado_95908	0.6621	0.024613	*	0.65973197
Magistrado_Melva_Giraldo_Londono	-0.7058	0.025717	*	0.33052755
entidad_demandada_fiscalia_general_de_la_nacion	0.2425	0.034445	*	0.56032964
abogado_98987	1.1380	0.011931	*	0.75731225
Magistrado_Victor_Manuel_Buitago_Gonzalez	-1.3410	0.005158	**	0.20734566
abogado_45113	-0.5103	0.020733	*	0.3751232
ciudad_Santa_Marta	0.5166	0.014329	*	0.62635239
abogado_91183	0.6517	0.032607	*	0.65739345
abogado_otro	0.0811	0.21666		0.52025641
dmte_Varias_naturales	0.2730	0.004885	**	0.56782925

Cuadro 11: Anexo VI - Regresión para procesos de orden nacional - comienzo del caso (Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Variable	logit (coeficientes)	z value	Pr(> |z|)	$P_{($
(Intercept)	-0.523300	0.052114	.	0.372
monto_pretendido	0.000000	0.456928		0.500
subtipo2_Pension (Nulidad y restablecimiento del derecho)	0.596300	0.000000	***	0.644
entidad_demandada_caja_de_retiro_de_las_fuerzas_militares	0.457700	0.000005	***	0.612
entidad_demandada_fiduciaria_la_previsora_sa	0.510700	0.000001	***	0.624
entidad_demandada_caja_de_sueldos_de_retiro_de_la_policia_nacional	-0.449000	0.000005	***	0.389
ciudad_Bogota	0.095890	0.229327		0.523
abogado_112907	2.986000	0.000001	***	0.951
Magistrado_0	0.153200	0.016374	*	0.538
subtipo2_Retiro_del_servicio (Nulidad y restablecimiento del derecho)	-0.707200	0.000001	***	0.330
corporacion_Tribunales	-0.442800	0.053150	.	0.391
abogado_170560	1.202000	0.000000	***	0.768
ciudad_Tunja	-0.388300	0.000767	***	0.404
ciudad_Bucaramanga	1.005000	0.000000	***	0.732
abogado_109557	2.025000	0.000000	***	0.883
ciudad_Ibague	1.134000	0.000000	***	0.756
abogado_31571	-0.984900	0.000000	***	0.271
subtipo2_organos_de_inspeccion_y_vigilancia (Nulidad y restablecimiento del derecho)	-1.498000	0.000008	***	0.182
subtipo2_Salarial (Nulidad y restablecimiento del derecho)	-0.526800	0.000427	***	0.371
ciudad_Pereira	-1.492000	0.000000	***	0.183
corporacion_Juzgado	0.040030	0.864265		0.510
ciudad_Villavicencio	-0.536000	0.001738	**	0.369
abogado_31614	-0.911700	0.000017	***	0.286
subtipo2_Error_jurisdiccional (Reparacion directa)	-0.988200	0.000074	***	0.271
abogado_83363	-1.355000	0.000032	***	0.205
abogado_59415	-0.599600	0.009396	**	0.354
entidad_demandada_caja_nacional_de_prevision_social_eice___en_liquidacion	0.101300	0.291692		0.525
entidad_demandada_finalejercito_nacional_de_colombia	0.486000	0.000255	***	0.619
abogado_56834	-1.034000	0.000092	***	0.262
Magistrado_Jose_Rodrigo_Romero_Romero	0.989200	0.000087	***	0.728

Variable	logit (coeficientes)	z value	Pr(> |z|)	P_0
ciudad_Cali	0.065010	0.558861		0.516247
abogado_90682	1.144000	0.000192	***	0.758413
genero_dmte_Hombre	0.032610	0.560157		0.508152
Magistrado_Carlos_Alberto_Vargas_Bautista	-2.051000	0.005589	**	0.113951
cdmte_Persona_juridica	0.066870	0.554772		0.516711
abogado_98987	1.359000	0.000648	***	0.795597
subtipo2_Impuestos_nacionales (Nulidad y restablecimiento del derecho)	0.766300	0.000016	***	0.682720
entidad_demandada_ministerio_de_la_proteccion_social	-0.510000	0.010249	*	0.375194
abogado_final41146	0.908100	0.000886	***	0.712611
ciudad_Armenia	0.385800	0.020499	*	0.595271
abogado_otro	0.067980	0.237217		0.516988
abogado_45113	-0.609800	0.002764	**	0.352105
ciudad_Valledupar	0.492800	0.017827	*	0.620766
Magistrado_Melva_Giraldo_Londono	-0.923000	0.002594	**	0.284347
entidad_demandada_superintendencia_de_servicios_publicos_domiciliarios	-0.852600	0.002593	**	0.298888
subtipo2_Prestacional (Nulidad y restablecimiento del derecho)	0.272800	0.020943	*	0.567780
Magistrado_otro	-0.015130	0.840640		0.496218
abogado_95908	0.750700	0.003612	**	0.679331
ciudad_Medellin	0.096780	0.318850		0.524176
entidad_demandada_fiscalia_general_de_la_nacion	0.350800	0.000348	***	0.586812
subtipo2_Prestacion_de_servicios_publicos (Reparacion directa)	-0.581700	0.021740	*	0.358542
abogado_91183	0.708700	0.013191	*	0.670114
abogado_158718	-0.520800	0.010387	*	0.372665
ciudad_Santa_Rosa_de_Viterbo	0.713100	0.002607	**	0.671086
subtipo2_Sufridos_por_conscripto (Reparacion directa)	0.559000	0.008604	**	0.636221
subtipo1_Laboral (Nulidad y restablecimiento del derecho)	0.102400	0.420498		0.525578
Magistrado_Edda_Del_Pilar_Estrada_Alvarez	0.511500	0.044122	*	0.625158
numero_demandados	-0.026240	0.637749		0.493440
entidad_demandada_consejo_superior_de_la_judicatura	0.338900	0.040665	*	0.583923
Magistrado_Victor_Manuel_Buitago_Gonzalez	-0.867500	0.028659	*	0.295775
dmte_Varias_naturales	0.214900	0.019933	*	0.553519
subtipo2_Causados_por_la_fuerza_publica (Reparacion directa)	0.354200	0.006884	**	0.587636
ciudad_Popayan	0.082260	0.489233		0.520553
Magistrado_Jaime_Orlando_Santofimio_Gamboa	-0.697700	0.026936	*	0.332322