PROYECTO DATA SCIENCE

Predicción sobre la aprobación de leyes en el congreso chileno

Documentación

Mackarena Toloza ❖ Diego de la Rivera ❖ Pablo Toloza ❖ Paula Llanos

El presente documento tiene como objetivo servir como documentación que acompaña el desarrollo del proyecto final de data science. El objetivo de este documento será presentar el equipo de trabajo, el tema de investigación, la motivación, el plan de trabajo y seguimiento para el tema escogido. Para efectos de una mejor organización, dividiremos esta minuta en cuatro secciones: Equipo de trabajo, Selección del tema; Planificación y Seguimiento de la investigación. En la primera sección se presentarán los roles que asumirá cada integrante durante la realización del proyecto. Posteriormente, se reflexionará en torno a la motivación, el impacto y la importancia que representa el tema escogido. Aquí también se definirán los objetivos generales, específicos, la pregunta de investigación y la hipótesis de trabajo. Luego se definirá la planificación y los requerimientos específicos del proyecto, junto con los plazos para cada una de las tareas propuestas, para finalmente en el seguimiento comentar los avances semana a semana.

1. EQUIPO DE TRABAJO

El equipo de trabajo está constituido por cuatro integrantes, que se dividirán en torno a cinco tareas a realizar a lo largo del proyecto. Las tareas, junto con las asignaciones son las siguientes:

- **Líder del equipo:** Mackarena Toloza.

  El líder del equipo estará encargado de dirigir, gestionar y organizar las tareas en torno al proyecto final. Para este caso específico también se encargará de que todos los entregables relacionados al proyecto se realicen dentro de los plazos correctos y de la mejor manera posible. Esto incluye las documentaciones, las presentaciones y los documentos en formato jupyter notebook.

- **Analista de datos:** Pablo Toloza, Paula Llanos, Diego de la Rivera, Mackarena Toloza.

  Los analistas de datos estarán encargados de preprocesar la base de datos, hacer las recodificaciones correspondientes y realizar toda la estadística descriptiva preliminar, esto incluye tablas de frecuencias, gráficos, entre otros. Además, serán los encargados de procesar la base de datos para que quede preparada para la ejecución de los modelos correspondientes.

- **Control de calidad y validación de datos:** Paula Llanos, Diego de la Rivera, Pablo Toloza.

  Los encargados de control de calidad y validación de datos serán los responsables de verificar que todos los procesos se estén realizando de manera adecuada y correcta para asegurar resultados óptimos.

- **Ingeniero/a de modelamiento:** Paula Llanos, Diego de la Rivera, Pablo Toloza.

  Los ingenieros de modelamiento estarán a cargo de preparar, ejecutar y evaluar el rendimiento de los modelos que utilizaremos para el desarrollo del proyecto. De ser necesario, también serán los encargados del mejoramiento de los mismos.

- **Documentador:** Mackarena Toloza.

  El documentador estará a cargo de redactar y preparar todos los archivos relacionados a la documentación del proyecto y los materiales de apoyo solicitados.

2. SELECCIÓN DEL TEMA Y MOTIVACIÓN

El tema escogido para la realización de este proyecto final es la aprobación de leyes en el congreso chileno. Las motivaciones para escoger esta temática radican en dos razones principales: En primer lugar, resulta interesante abordar la productividad legislativa desde una perspectiva del aprendizaje de máquinas. Generalmente, en el ámbito de la ciencia política y de las ciencias sociales en general, más que buscar predecir acontecimientos, lo que se busca es explicar fenómenos sociales, políticos, culturales que *ya ocurrieron*. Muchas veces además, los fenómenos que se estudian son bastante específicos, poco generalizables. Dado lo anterior, resulta interesante y sobre todo novedoso para la disciplina abordar la productividad legislativa desde el *machine learning* para intentar predecir si una ley será aprobada o no durante su trámite legislativo.

En segundo lugar, y al estar hablando de instituciones tan importantes como el ejecutivo o el legislativo, es que abordar esta problemática puede desencadenar en una poderosa herramienta para la toma de decisiones y un mejor manejo de recursos. Teniendo un modelo que sea capaz de predecir si una ley puede ser aprobada o no de manera más o menos consistente, por ejemplo, el Presidente podría tomar la decisión de apoyar o centrar ciertos recursos (como las urgencias) en aquellos proyectos que tienen mayores chances de ser aprobados. A la vez, los legisladores también podrían centrar sus esfuerzos en apoyar a estos proyectos o intentar ponerlos en tabla antes que aquellos proyectos que parecen no tener chances de aprobación.

Dado lo anterior, la definición del problema se desprende de la siguiente manera:

Objetivo de Investigación General:

- Predecir si un proyecto de ley tiene chances de ser aprobado o no.

Objetivos específicos:

- Generar modelos de machine learning y evaluar sus rendimientos respecto a la problemática.
- Evaluar la literatura especializada para definir si hay algunas variables importantes que puedan ayudar a predecir la problemática de mejor forma.

Pregunta de Investigación:

- ¿Es posible predecir de manera más o menos robusta si un proyecto de ley tiene chances de ser aprobado?

Hipótesis de Investigación:

H1: Es posible predecir con al menos un 0.75 de precisión para la clase 1, si un proyecto de ley tiene chances de ser aprobado tomando en cuenta variables como el tipo, la procedencia, el origen, si están relacionados a alguna ley, si fueron presentados durante ciertos períodos específicos, entre otros.

3. PLANIFICACIÓN DE LA INVESTIGACIÓN

La base que utilizaremos para trabajar en esta investigación es una base de datos que contempla todas las leyes presentadas al congreso durante 20 años, desde el año 1990 al año 2009. Como nuestra base de datos contempla leyes, no tenemos datos sensibles, de hecho estamos trabajando con datos los cuales cualquier persona tiene acceso a través de las páginas del senado y de la cámara de diputados[1].

No aplicaremos técnicas para trabajar *missing values* dado que no tenemos datos faltantes. Trabajaremos con todas las leyes y tenemos datos para cada una de ellas. De la misma manera, tampoco aplicaremos técnicas de muestreo, dado que como nuestra "población objetivo" son las leyes presentadas, trabajaremos con todas dentro de un período específico.

Todas las variables que incluiremos en nuestros modelos tienen una razón de ser, la mayoría de ellas las incluimos porque hay un respaldo científico detrás que afirma que dichas variables pueden tener efectos sobre la productividad legislativa. Explicaremos esto con más detalle.

Los estudios sobre productividad legislativa en Chile identifican algunas variables importantes para explicar una mayor o menor productividad, entendiendo productividad como el porcentaje de aprobación de leyes en un período específico. En general, hay bastante consenso respecto a que variables relacionadas a los ciclos políticos, entendidos como períodos que se van repitiendo cada cierto tiempo tales como elecciones o primer año de gobierno, son importantes para explicar la productividad en esta materia (Visconti, 2011). También existe consenso respecto a que las variables que afectan la productividad legislativa parecieran ser variables más "estructurales", donde aquellos acontecimientos más cambiantes entre un gobierno y otro, como por ejemplo, la aprobación en mayor o menor medida de un Presidente, no afectarían significativamente la productividad de éste en materia legislativa (Alemán & Navia, 2009). Ni siquiera un cambio en la coalición de gobierno tendría efectos en la productividad debido a que la coalición que gobierna siempre ocupa las mismas estrategias para conseguir mayor productividad legislativa (Toloza y Toro, 2017).

Dicho esto, explicaremos con más detalle las variables que utilizaremos para la realización del presente proyecto. Trataremos un problema de **clasificación**, dado que nuestro vector objetivo será la variable "estado"

[1] http://www.senado.cl – http://www.camara.cl

que identifica si un proyecto fue publicado o no. Esta variable será recodificada como *dummy* donde el valor 0 representará aquellos proyectos que fueron rechazados y el valor 1 aquellos proyectos que fueron publicados.

Tenemos algunas variables informativas como "ingreso" que representa la fecha en que fue ingresado el proyecto, y el "boletín" que es un registro único que se le asigna a cada proyecto de ley que ingresa al congreso, esto como un identificador. Como tenemos valores para nuestro vector objetivo, utilizaremos modelos de aprendizaje supervisado.

Dentro de las variables explicativas que utilizaremos está el "título". El título está representado como un string, lo que haremos será hacer *text mining* para intentar identificar algunos patrones a través de un count vectorizer. Sin embargo, si el tiempo lo permite también intentaremos realizar otras técnicas para extracción de tópicos.

La variable procedencia identifica si un proyecto es "Mensaje" o "Moción". Cuando un proyecto está catalogado como Mensaje, es que es un proyecto enviado por el Presidente. Cuando está catalogado como Moción, es que fue enviado por un Legislador. Esta distinción resulta importante dado que para la literatura hay variables que, por ejemplo, pueden afectar los Mensajes pero no las Mociones. Hablaremos más detenidamente sobre esto al explicar la última variable que incluiremos en nuestra base de datos. Esta variable será recodificada como dummy.

La variable tipo identifica el tipo de ley; en este caso, si se trata de un Proyecto de Ley, una Reforma Constitucional, etc. Creemos que incorporar esta distinción es de suma relevancia dado que resulta de sentido común que no es lo mismo hablar de un proyecto de ley sobre alguna materia normal dentro de ciertos tópicos, que de un proyecto de ley que busca reformar la Constitución del país.

Luego tenemos la variable de origen. La variable de origen identifica si la cámara de origen del proyecto es la cámara de diputados o senadores. Esta será recodificada como *dummy*, y creemos que es una variable que puede aportarle información extra y relevante a nuestro modelo.

Dado que el modelo que estamos pensando construir sería un modelo que evaluaría las leyes presentadas al congreso luego de un mes y medio/dos meses aproximadamente desde que ingresan, incluimos las variables dg1t y dp1t que son variables dummies que identifican si el proyecto tuvo discusión general del primer trámite (dg1t) o discusión particular del primer trámite (dp1t).

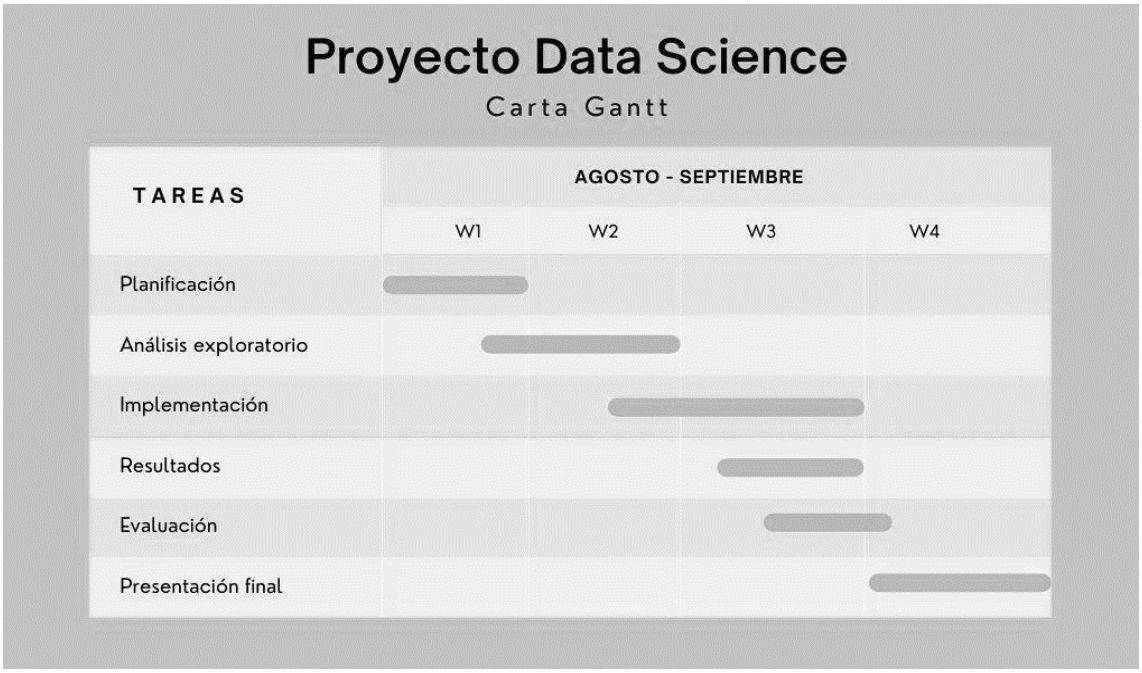
Por último, construiremos dos variables apoyadas por la literatura. Hemos visto anteriormente que los ciclos políticos, sobre todo aquellos relacionados a las elecciones, parecieran tener capacidad explicativa sobre la productividad legislativa. Es por esto, que incluiremos una variable llamada elección que identifique si hubo o no elección cuando el proyecto fue presentado, y de qué tipo (presidenciales, parlamentarias o ambas).

Por último, incluiremos una variable llamada *honeymoon*. Alemán & Navia (2009) encontraron que esta variable resulta importante para explicar la productividad legislativa **del Ejecutivo** (es decir, de los Mensajes,

no de las Mociones). Ellos afirman que los primeros seis meses de un gobierno (llamado período de "luna de miel"), impacta significativamente y positivamente en la productividad legislativa. En palabras simples, los Presidentes tienen muchas más chances que sus proyectos sean aprobados si los envían durante este período. Por tanto, incluiremos esta variable que identifique si los Mensajes fueron enviados durante períodos de *honeymoon* o no.

4. PLANIFICACIÓN

Por último, presentamos la planificación de acuerdo a las cuatro semanas que durará el proyecto final. De acuerdo a esto utilizaremos la primera semana como planificación y comenzaremos los análisis exploratorios. Dedicaremos la segunda semana al análisis exploratorio y primeras implementaciones de modelos. La semana tres estará enfocada en la implementación y la obtención de los resultados definitivos. También en esta semana se realizará la evaluación de los modelos. Por último, la última semana estará dedicada a la preparación de la presentación final.



5. SEGUIMIENTO

El objetivo de esta sección es realizar un seguimiento semana a semana e ir comentando los avances del proyecto a medida que se va desarrollando.

La primera semana estuvo dedicada casi exclusivamente a la planificación y programación de las tareas, junto con un primer acercamiento a la base con la que se trabajará a lo largo del proyecto y al vector objetivo. Respecto a la base de datos, durante esta semana se han discutido las variables que consideramos importantes incluir y aquellas que será necesario crear según la documentación científica. Frente a esto, hemos dejado definidas todas las variables a considerar, y para aquellas variables a construir, hemos especificado cómo deben ser construidas y qué aspectos se debe considerar. Además de ello, hemos realizado un par de gráficos

preliminares simplemente para acercarnos un poco a cómo están distribuidas algunas de las variables que utilizaremos. También durante esta semana se comenzaron a recodificar algunas variables que tenían problemas con los nombres de sus categorías. Por ejemplo, en la variable "tipo" la categoría "Reforma Constitucional" estaba escrita como "Reforma Constitucional", "reforma constitucional", "RC", entre otras; nos hemos preocupado de recodificar estas inconsistencias para que todo quede de la misma manera.

Durante la segunda semana, hemos puesto en marcha lo planificado durante la primera y hemos comenzado ya a trabajar de lleno con la base de datos que utilizaremos. Hemos analizado a fondo los *missings values*, hemos recodificado todas las variables incluido el vector objetivo. Se han creado gráficos y tablas para las variables relevantes. Además, se han construido las dos variables que no venían preliminarmente en la base: *elección* y *honeymoon*, y se han pasado a dummy todas aquellas variables categóricas. Teniendo todo el preprocesamiento de estos aspectos ya hechos, se ha comenzado a realizar la tokenización del texto de la variable título. Creamos una función que tiene como finalidad limpiar el texto antes de aplicar el count vectorizer para eliminar caracteres especiales y otros adicionales. Analizamos la frecuencia de las palabras más utilizadas en los títulos y finalmente hemos dejado la base lista incluyendo el procesamiento del texto para comenzar a entrenar y evaluar algunos modelos.

Respecto a los modelos, hemos discutido preliminarmente probar tres: en primer lugar, un *Lineal Discriminant Analysis*, en segundo lugar un *Gradient Boosting,* y en tercer lugar un *Random Forest*. La razón por la cual creemos que un *Lineal Discriminant Analysis* puede desempeñarse bien en este caso, es porque estamos convencidos que tenemos buenas variables explicativas y que por consecuencia, nuestro modelo podría aprender bastante bien esa "línea" de discriminación para identificar entre una clase u otra. Es por ello que nos parece interesante ver en un primer momento cómo este modelo se comporta con nuestros datos y su rendimiento.

Como segundo modelo nos interesa probar un *Gradient Boosting* básicamente dadas sus ventajas. Como se trata de un modelo que va aprendiendo de manera secuencial y de los mismos errores que va cometiendo en cada iteración, creemos que también puede aprender bastante bien a clasificar nuestro vector objetivo dadas las variables explicativas que tenemos, y por ende tener también un muy buen rendimiento.

Como tercer y último modelo probaremos un *Random Forest* también dadas sus ventajas al ser un modelo ensamblador. Lo que hacen este tipo de modelos es formarse a partir de un grupo de modelos predictivos, lo que les permite alcanzar mejor precisión y rendimiento. La forma de trabajar es que cada "árbol" da una clasificación; el resultado es la clase con mayor número de votos. Los modelos de *Random Forest* son bastante populares por desempeñarse bastante bien generalmente independiente del problema, además de que se pueden utilizar en problemas de regresión o clasificación.

Considerando lo anterior, hemos puesto a prueba estos modelos consiguiendo resultados preliminares bastante buenos. Para el caso del *Lineal Discriminant Analysis* hemos conseguido una precisión de un 0.78 para la clase 1 (aprobado), un recall de 0.82 para la misma clase, y un *accuracy* de 0.88. Para el caso del modelo *Gradient Boosting*, hemos conseguido una precisión de un 0.79 para la clase 1 (aprobado), un recall de 0.88

para la misma clase, y un accuracy de 0.90. Por último, para nuestro modelo *Random Forest* hemos conseguido un 0.81 de precisión para la clase 1, un recall de 0.85 y un accuracy de 0.90.

Hemos implementado además un gráfico de curva ROC para evaluar los tres modelos. Gráficamente observamos que los tres modelos entrenados cuentan con métricas bastante buenas y tienen buen rendimiento predictivo. Sin embargo, tanto *Gradient Boosting* como *Random Forest* están un poco por sobre el modelo *LDA*.

BIBLIOGRAFÍA

Alemán, Eduardo, and Patricio Navia. 2009. "Institutions and the Legislative Success of 'Strong' Presidents: An Analysis of Government Bills in Chile." Journal of Latin American Studies 41 (3):467-91.

Toloza, Mackarena, and Sergio Toro. 2017. "Amigos Cerca, Enemigos Más Cerca: El Gobierno de Sebastián Piñera y las Dinámicas Legislativas en Chile." Revista Uruguaya de Ciencia Política 26 (1):131-49.

Visconti, Giancarlo. 2011. "Comportamiento Diacrónico Del Congreso en Chile: ¿Crecimiento o Estancamiento de su Influencia?" Revista de Ciencia Política 31 (1):91-115.