Machine Learning para la predicción de alertas en la ejecución de Proyectos de Regalías

Edwin Torres, MSc., PhD candidate
Machine Learning
Grupo ciencia de datos DNP

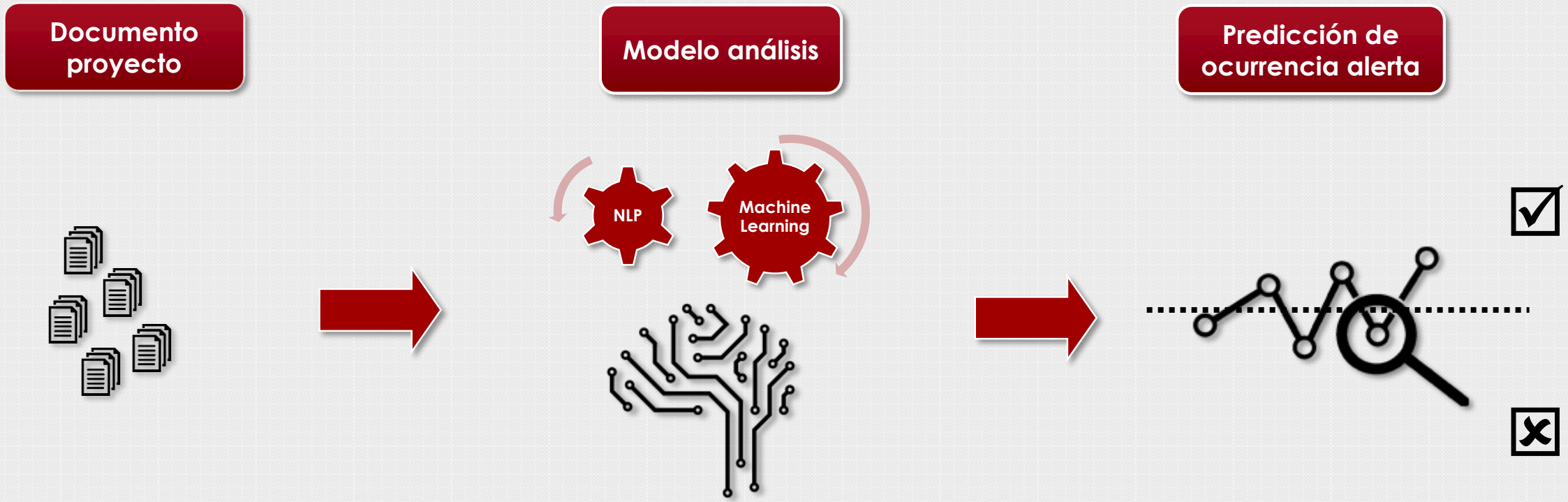
etorres@dnp.gov.co

Octubre, 2017
dnp.gov.co

¿Es posible determinar si un proyecto (SGR) tendrá dificultades en su ejecución analizando el texto de la propuesta?

Predicción en la aprobación

Predecir, antes de aprobación, cuales proyectos de regalías presentarán dificultades en su etapa de ejecución.



Flujo para la construcción del modelo

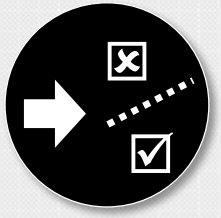
Estructuración del proyecto: ejecución por módulos

1. Recolección de los datos

- Origen de la información
- Proceso de captura
- Cantidad, Tipo y disponibilidad de los datos
- Definir dataset muestras ($\boldsymbol{X}$, $\boldsymbol{Y}$)

2. Transformación del texto

- Pipeline NLP
  - Ortografía
  - Lematizar
  - Remover SW
- Transformar (mapear - vectorizar)



3. Diseño clasificador

- Partición dataset
- Selección de modelo
- Explorar
  - NLP básico
  - Dim. Reducción (clustering)

4. Análisis resultados

- Medir
- Evaluar
- Analizar

- Evidenciar características principales

1. Recolección de datos

Ficha EBI (Estadísticas Básicas de Inversión)

- Resumen de los elementos que componen la descripción de un Proyecto
- Dividido por módulos

- Módulo de identificación
  - Problema central
  - Descripción de la situación
  - Causas
  - Efectos
  - Objetivos
  - Alternativa de solución

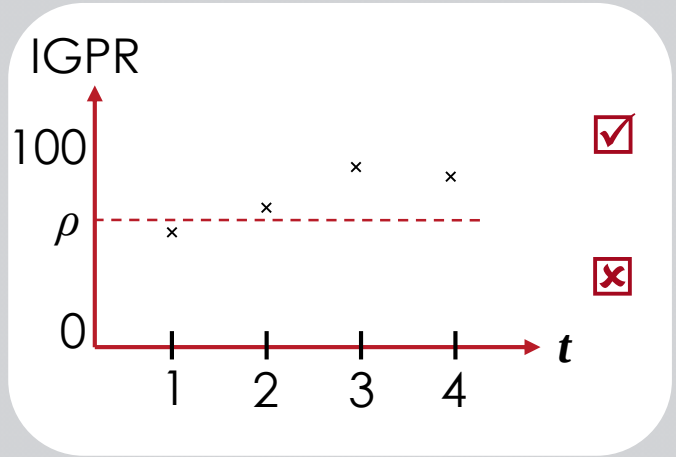
- **10.529** Proyectos
- **1'591.595** palabras*
- **68.539** palabras únicas

- Longitud promedio **151** palabras por documento
- Promedio **91** palabras únicas por documento

1. Recolección de datos

Índice de Gestión de Proyectos de Regalías - IGPR

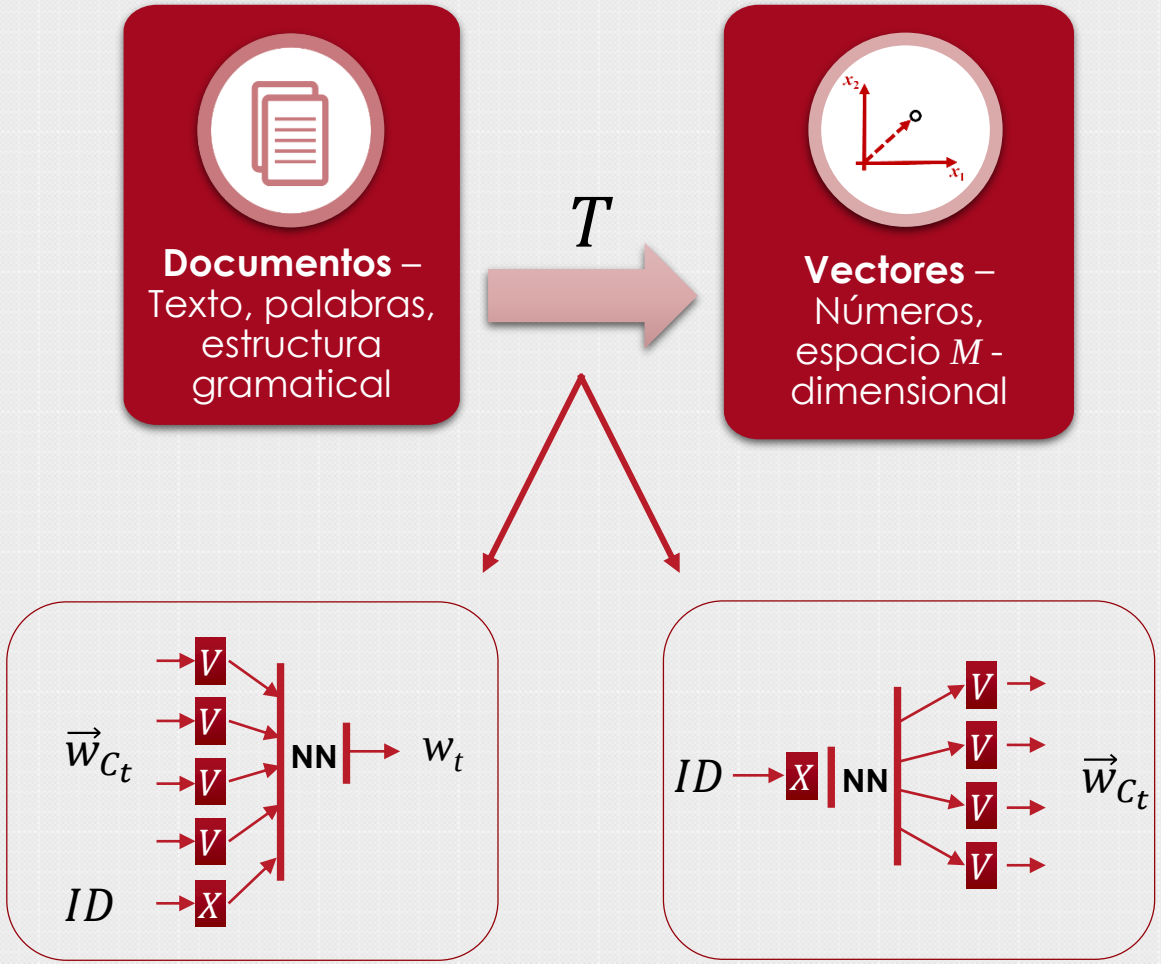
- Herramienta para medir y valorar la gestión de las entidades ejecutoras
- Visibilizar la gestión de los proyectos
- Análisis en dos dimensiones
  - Administrativa
    - Transparencia y medidas del SMSC
  - Desempeño
    - Eficiencia y eficacia
- IGPR = f (Transparencia, SMSCE, Eficiencia, Eficacia | θ)



$$y(IGPR) = \begin{cases} 1, & IGPR \leq \rho\ ,\forall\ t \\ 0, & IGPR > \rho\ ,\forall\ t \end{cases}$$

2. Transformación del texto

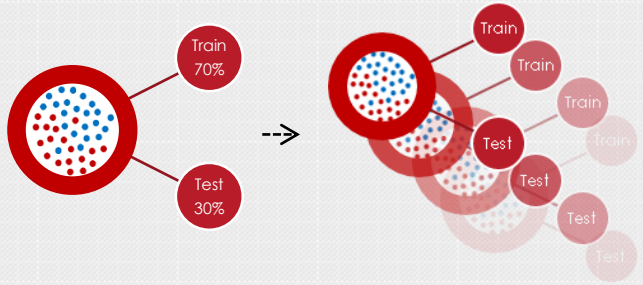
Extraer información cuantificable del texto: Vector Space Model, Neural network embedding



- Representaciones densas (no dispersas)
- Preservar la información contenida (Ej.: orden de las palabras)
- Tomar en cuenta el tipo de texto
- Permite construir una medida de similitud (Ej.: Ranking, búsqueda)

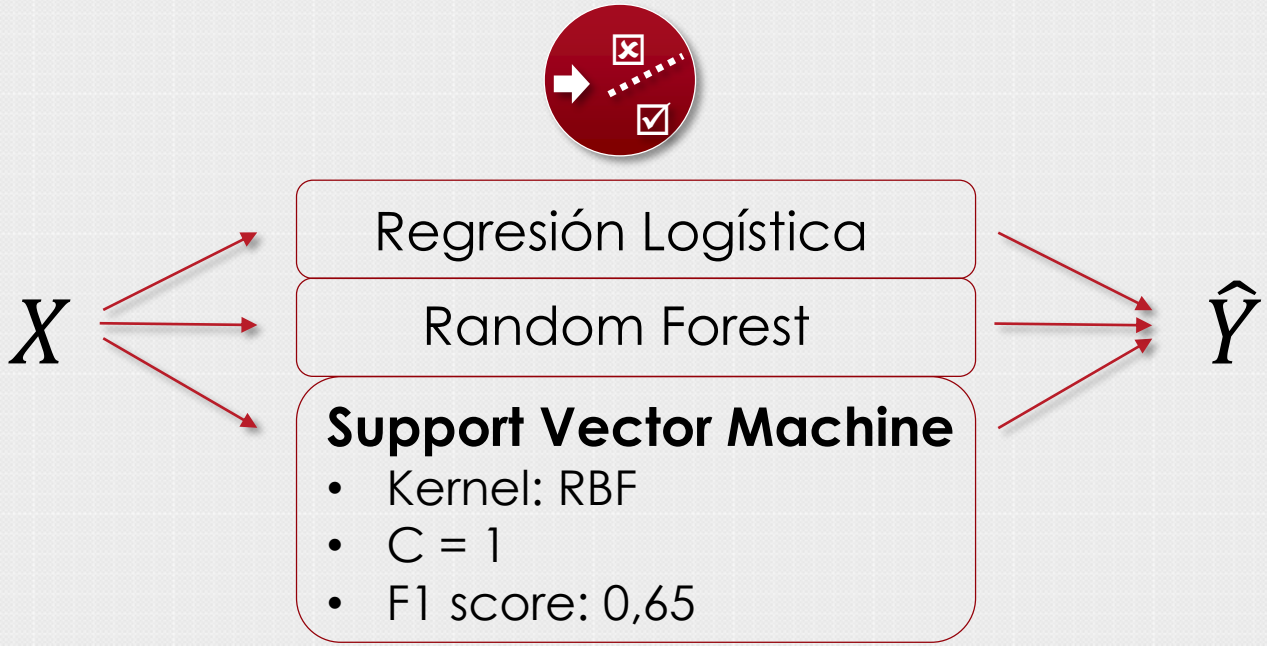
3. Diseño clasificador

Aprendizaje supervisado



- Probar diversos algoritmos de clasificación
- Complejidad del modelo
- Estimar los parámetros de cada modelo

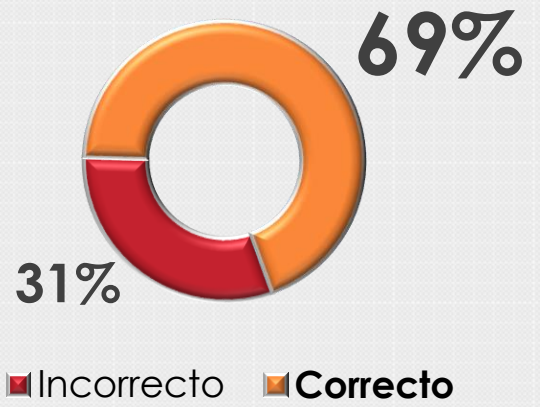
Proporción clases



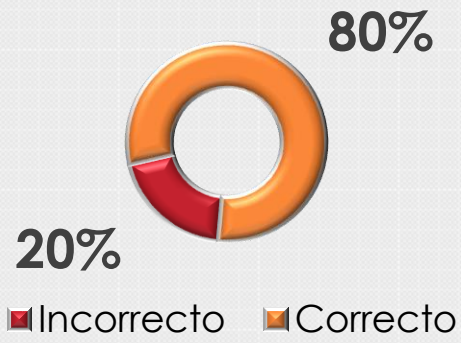
4. Análisis resultados

Precisión del clasificador

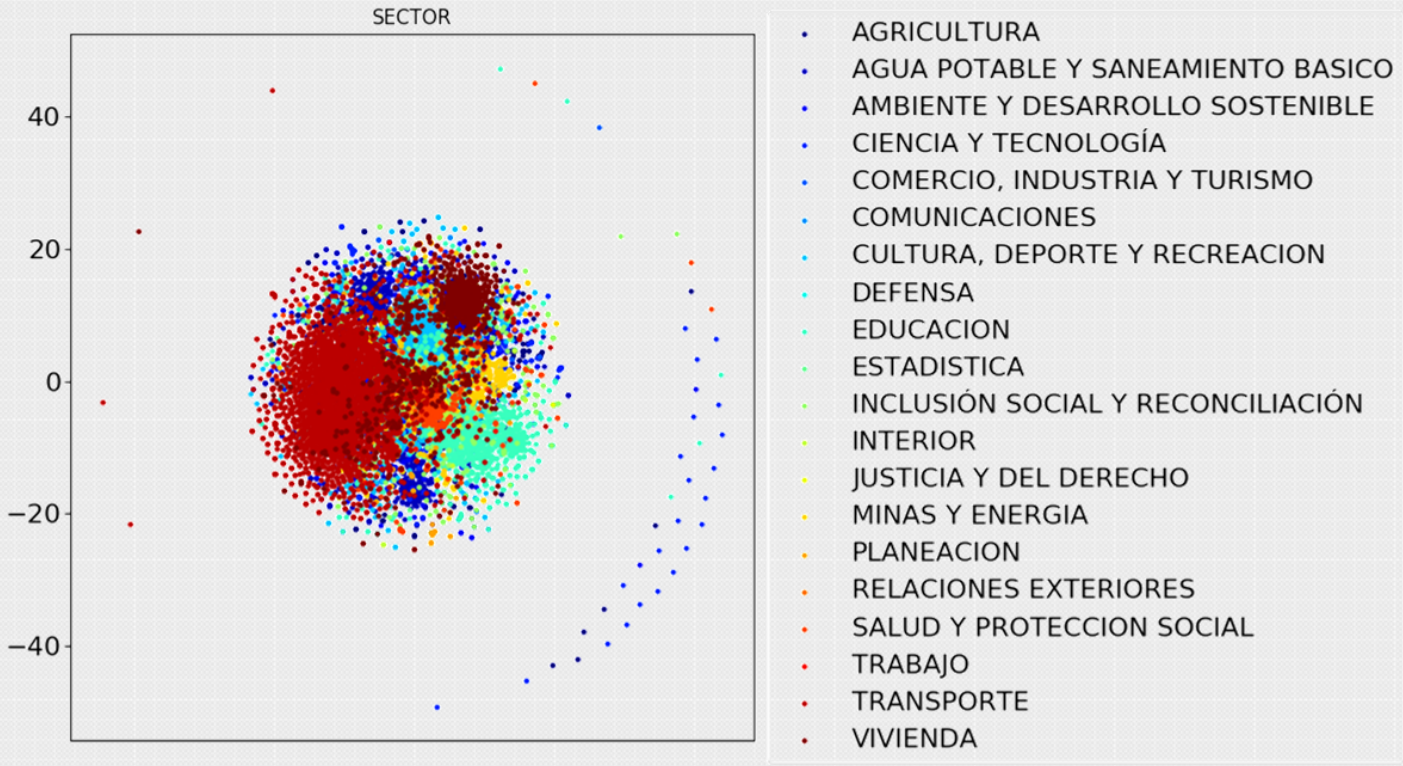
Error conjunto de prueba



Error conjunto de entrenamiento

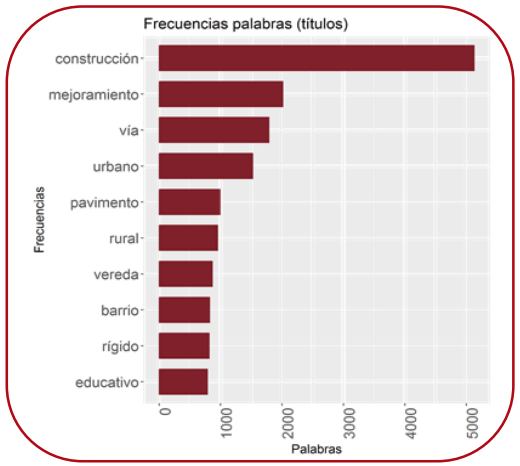


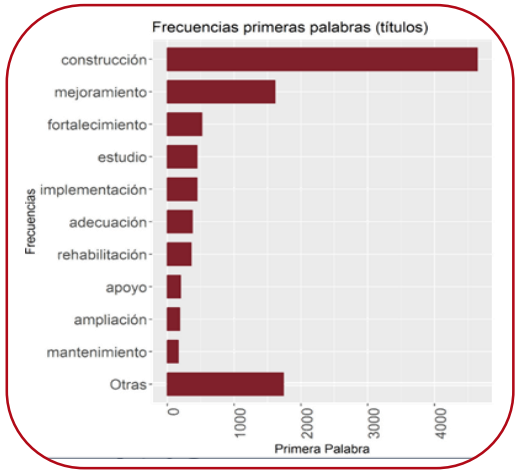
Vectorización – Clusters (t-SNE)

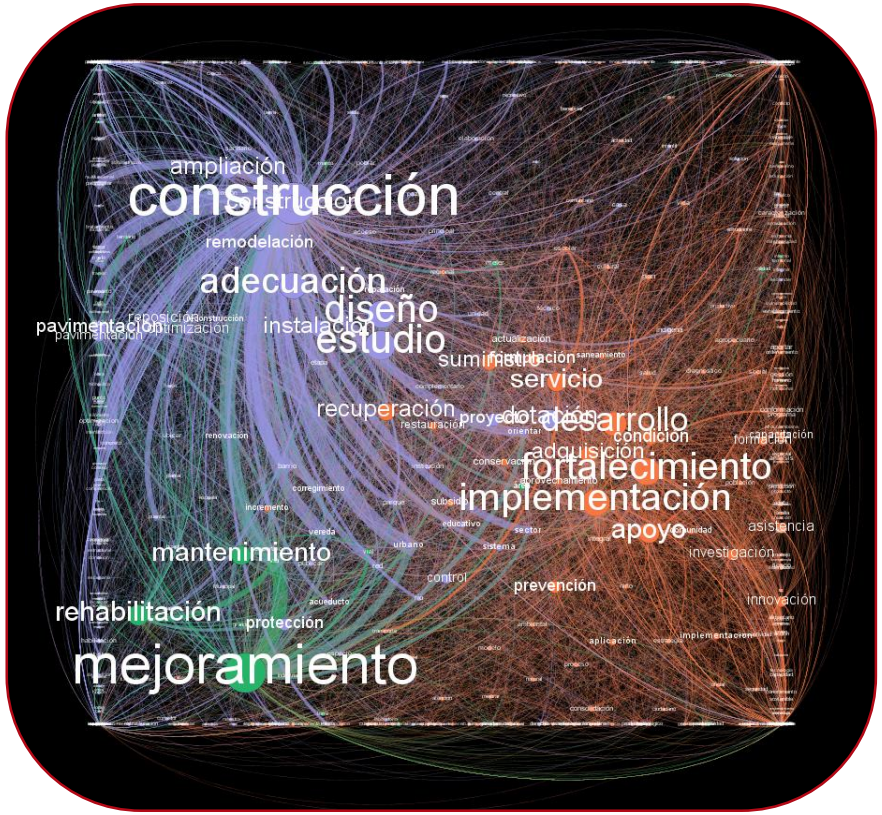


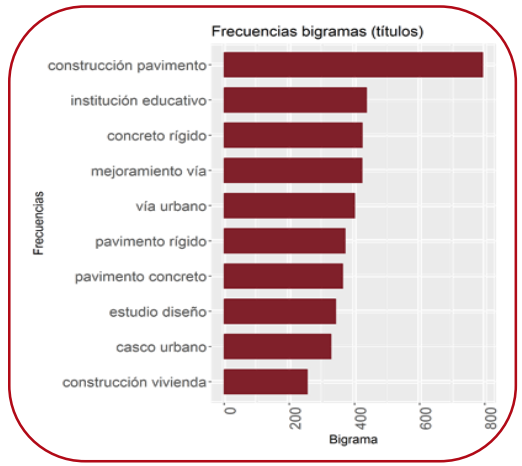
4. Análisis resultados

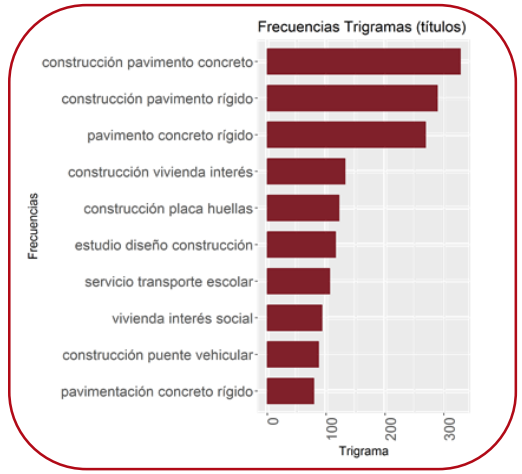
Indicadores del texto que evidencian relaciones









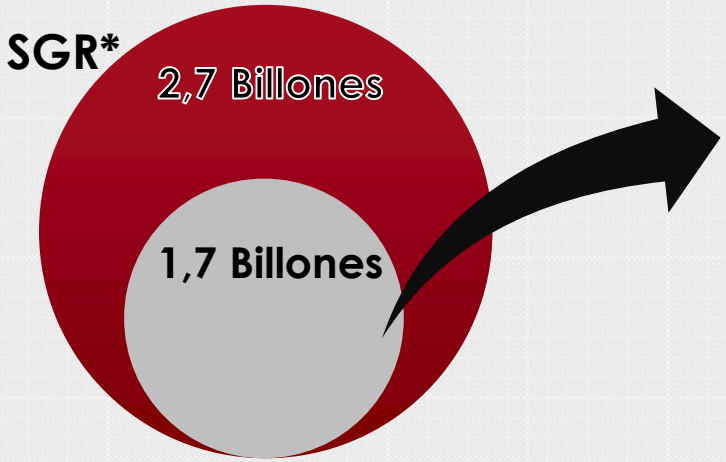


Beneficios

Análisis conjunto y medidas anticipadas: uso eficiente de recursos

Sin modelo predictivo

Detección sobre la **ejecución**

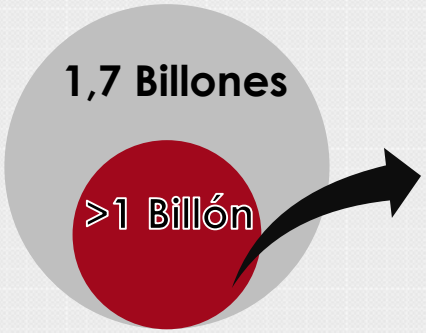


Proyectos que presentaron dificultades en su ejecución

¿Que sucedió?: retrasos, detección tardía, dificultad aplicación correctivos, ……..

Con modelo predictivo

Detección en la **pre-aprobación**



Proyectos clasificados que presentaran dificultades en su ejecución

¿Que sucedería?: minimización óptima retrasos, aplicación correctivos en la formulación, colaborar al seguimiento DNP

Conclusiones

- El análisis del texto complementa la información cuantitativa, ofrece otro punto de vista para la evaluación.
- Descubrir contexto a través de la vectorización, detectar patrones.
- Clasificador adaptable a otras métricas de desempeño SGR, o a otra tarea
- Priorización del seguimiento
  - Clasificar proyectos o entidades ejecutoras (Ranking)
- Evidenciar fallas "tempranas" en la formulación (correctivos anticipados).