El futuro es de todos

DNP
Departamento
Nacional de Planeación

Esta presentación es propiedad intelectual controlada y producida por el Departamento Nacional de Planeación

El futuro es de todos

DNP
Departamento Nacional de Planeación

Aplicación de Machine Learning al Sistema General de Regalías

Unidad de Científicos de Datos
Dirección de Desarrollo Digital

Diciembre, 2018

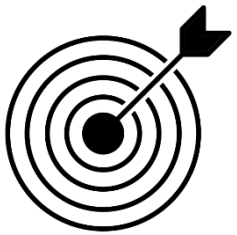
60 años 1958 2018

1. Descripción del proyecto
2. Descripción del SGR y el módulo de inteligencia artificial
3. Recolección y procesamiento de los datos
4. Entrenamiento de los modelos
5. Resultados y conclusiones
6. Cómo utilizar el modelo

1. Descripción del proyecto

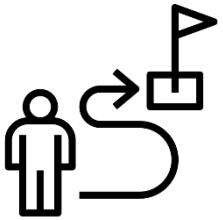
Descripción del proyecto

Desarrollo de modelos de aprendizaje automático para apoyar la evaluación y seguimiento de proyectos de regalías



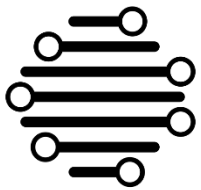
Objetivo

Predecir dificultades en la ejecución de proyectos de regalías, a partir de la información disponible al momento de su formulación



Metodología

Procesamiento de variables continuas, categóricas y de texto
Redes neuronales para representación numérica de textos
Algoritmos de agrupamiento
Algoritmos de *boosting* para clasificación



Insumos

Información de proyectos de regalías en su etapa de formulación
Mediciones IGPR de los proyectos de regalías

2. Descripción del SGR y el módulo de inteligencia artificial

Sistema General de Regalías (SGR)

Se encarga de la administración y vigilancia de los ingresos provenientes de la explotación de los recursos naturales no renovables, precisando las condiciones de participación de sus beneficiarios.

13.184

Proyectos aprobados desde el 2012

32

Billones de pesos aportados por recursos del SGR para el desarrollo de estos proyectos

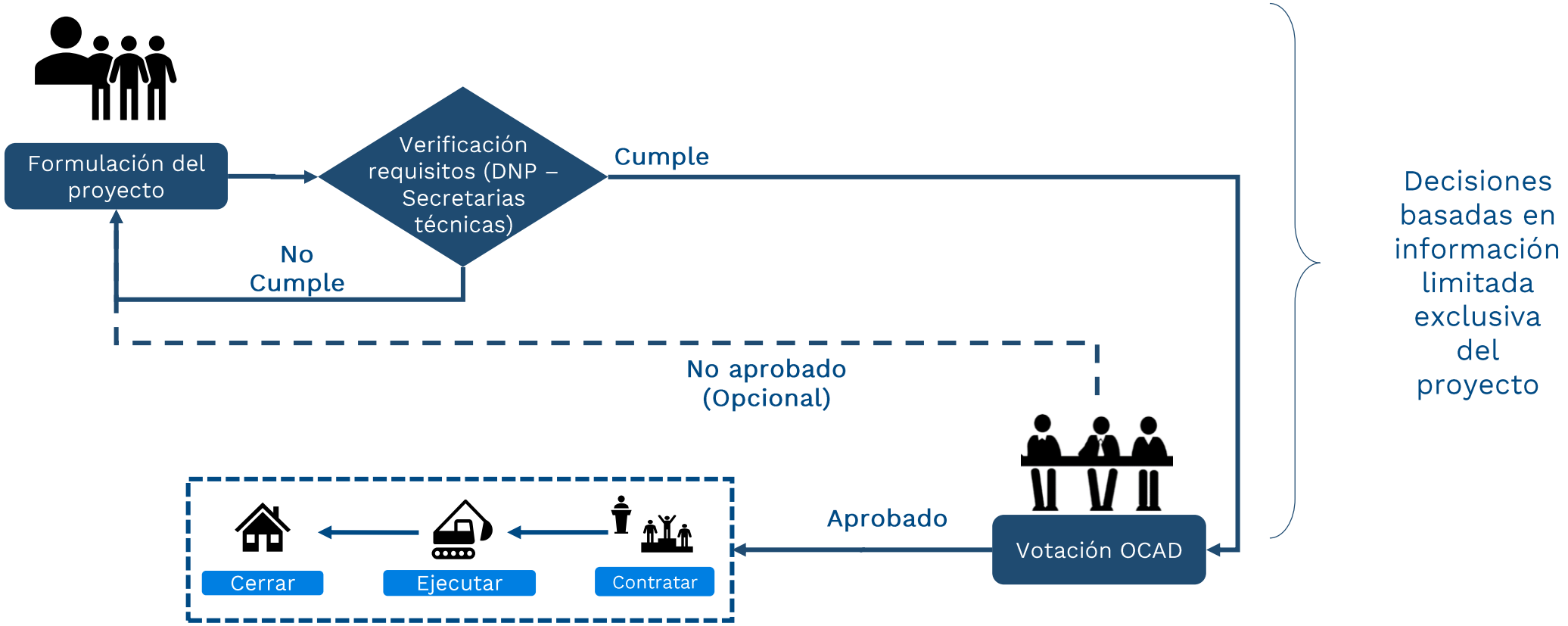
1.293

Entidades han sido designadas como ejecutoras de los proyectos

Fuente: Boletín mensual de proyectos SGR – octubre de 2018

Flujo de aprobación de proyectos del SGR

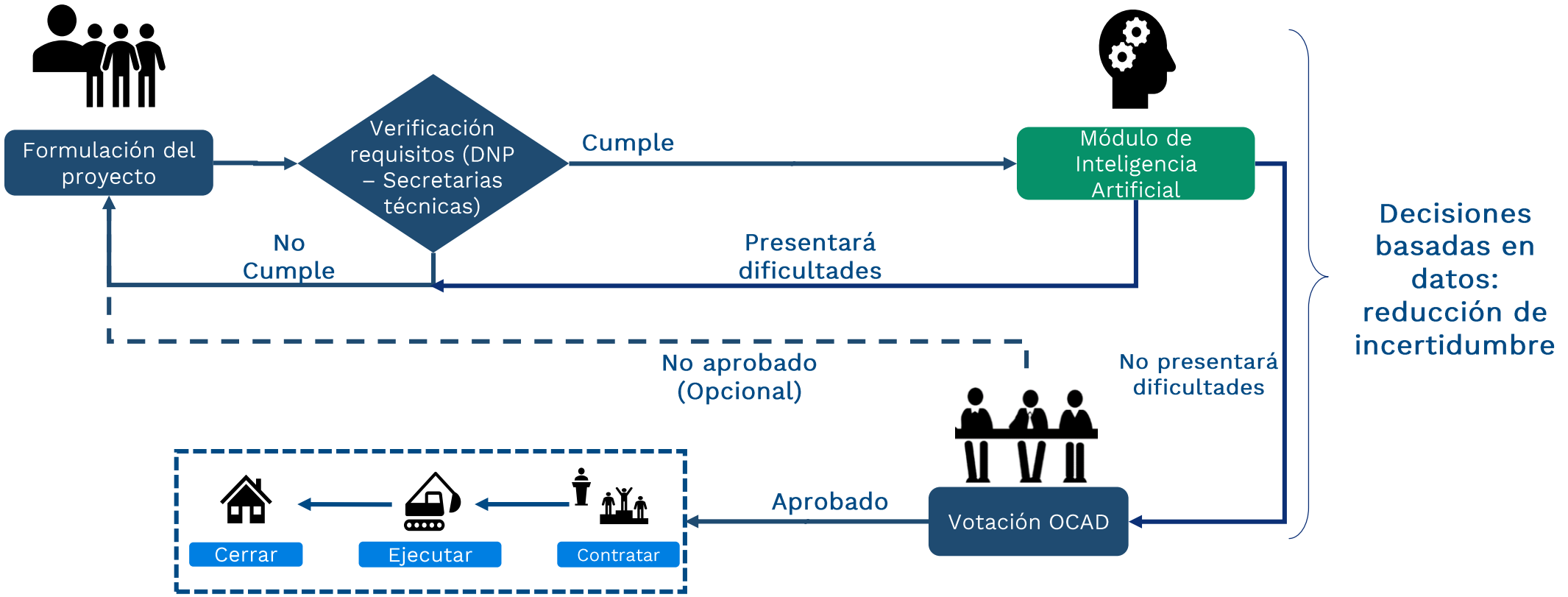
La inversión de recursos del SGR requiere un ciclo iterativo de preparación y formulación, previo a la presentación del proyecto ante el OCAD



OCAD: Órgano Colegiado de Administración y Decisión

Flujo de aprobación de proyectos SGR con IA

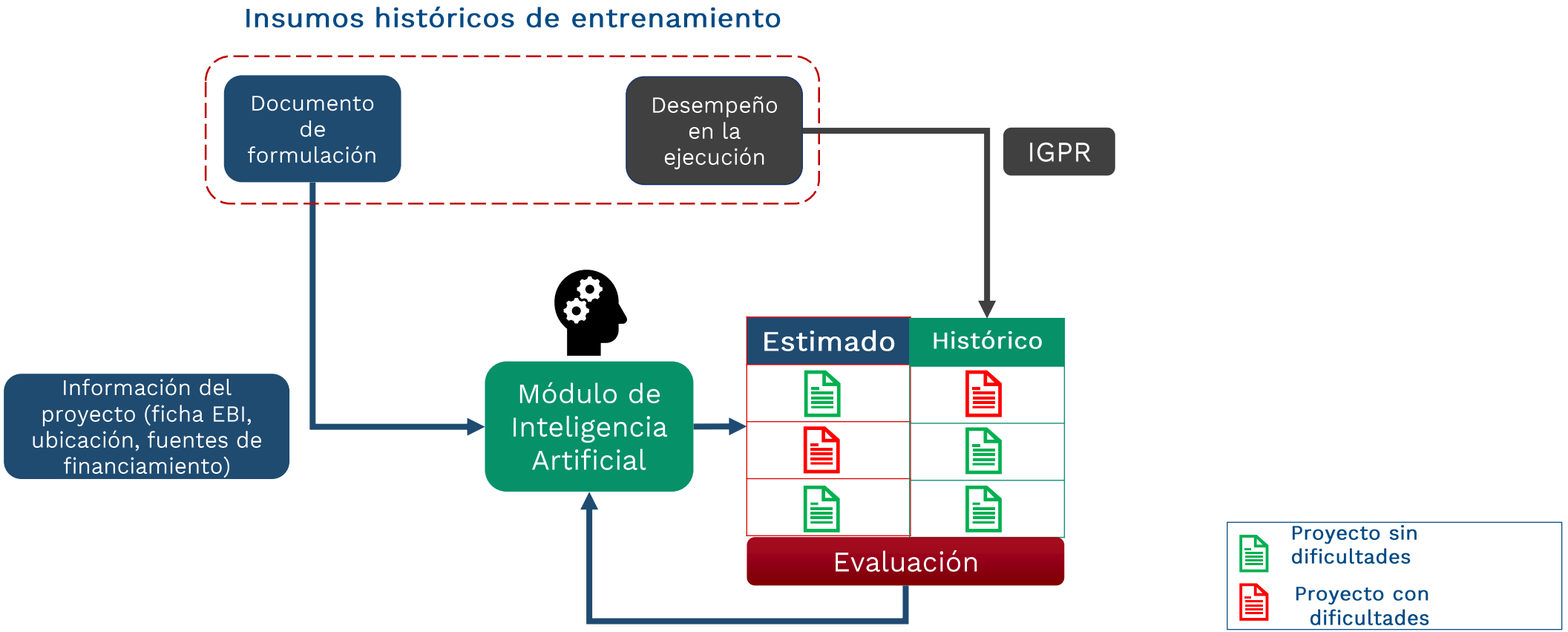
El módulo de Inteligencia Artificial (IA) analiza la información de formulación y desempeño para predecir si el proyecto formulado presentará dificultades en su ejecución



OCAD: Órgano Colegiado de Administración y Decisión

Diseño del módulo de IA

El módulo identifica los relaciones entre la información de formulación y el desempeño en ejecución de los proyectos

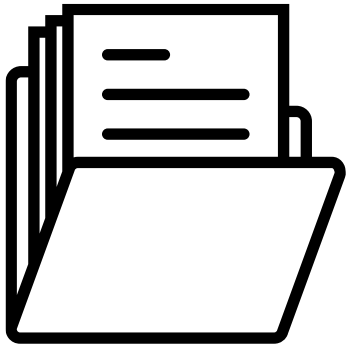


EBI: Estadísticas Básicas de Inversión
IGPR: índice de Gestión de Proyectos de Regalías

3. Recolección y procesamiento de los datos

Recolección de los datos

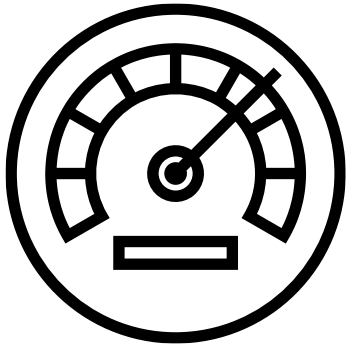
Información de los proyectos formulados es guardada en bases de datos mantenidas y actualizadas por el DNP



Fichas Estadísticas Básicas de Inversión (EBI)

Información adicional del proyecto

- Ubicación y duración del proyecto
- Fuentes de financiamiento
- Sector al que pertenece el proyecto



Resultados IGPR

Ventana de observación: Del primer trimestre del 2016 al segundo trimestre de 2018

Fichas Estadísticas Básicas de Inversión

Resumen las características centrales de un proyecto o programa. Estas fichas deben ser diligenciadas por las entidades para cada uno de los proyectos o programas que requiera financiamiento.

Permiten capturar de manera automatizada aspectos cualitativos y cuantitativos del proyecto como:

1. Nombre del proyecto
2. Descripción
3. Problema central
4. Causas directas e indirectas
5. Efectos directos e indirectos
6. Objetivo general y específicos
7. Productos

Variables disponibles para entrenar el modelo

	Variables numéricas	Variables categóricas	Variables de texto
Variables existentes	VALOR_SGR VALOR_TOTAL AnioInicio AnioFinal	REGION DEPARTAMENTO SECTOR	Causa Efecto Indicador Alternativa Riesgo Medidas_Mitigacion Nombre_Proyecto ProblemaCentral Descripcion ObjetivoGeneral
Variables construidas	Longitud_proyecto Causas_directas Causas_indirectas Efectos_directos Efectos_indirectos Numero_metas	financiamiento_externo	texto_unificado

Procesamiento de las variables

Dependiendo del tipo de variable, se realizaron distintos procesamientos de los datos para poderlos utilizar en el modelo.

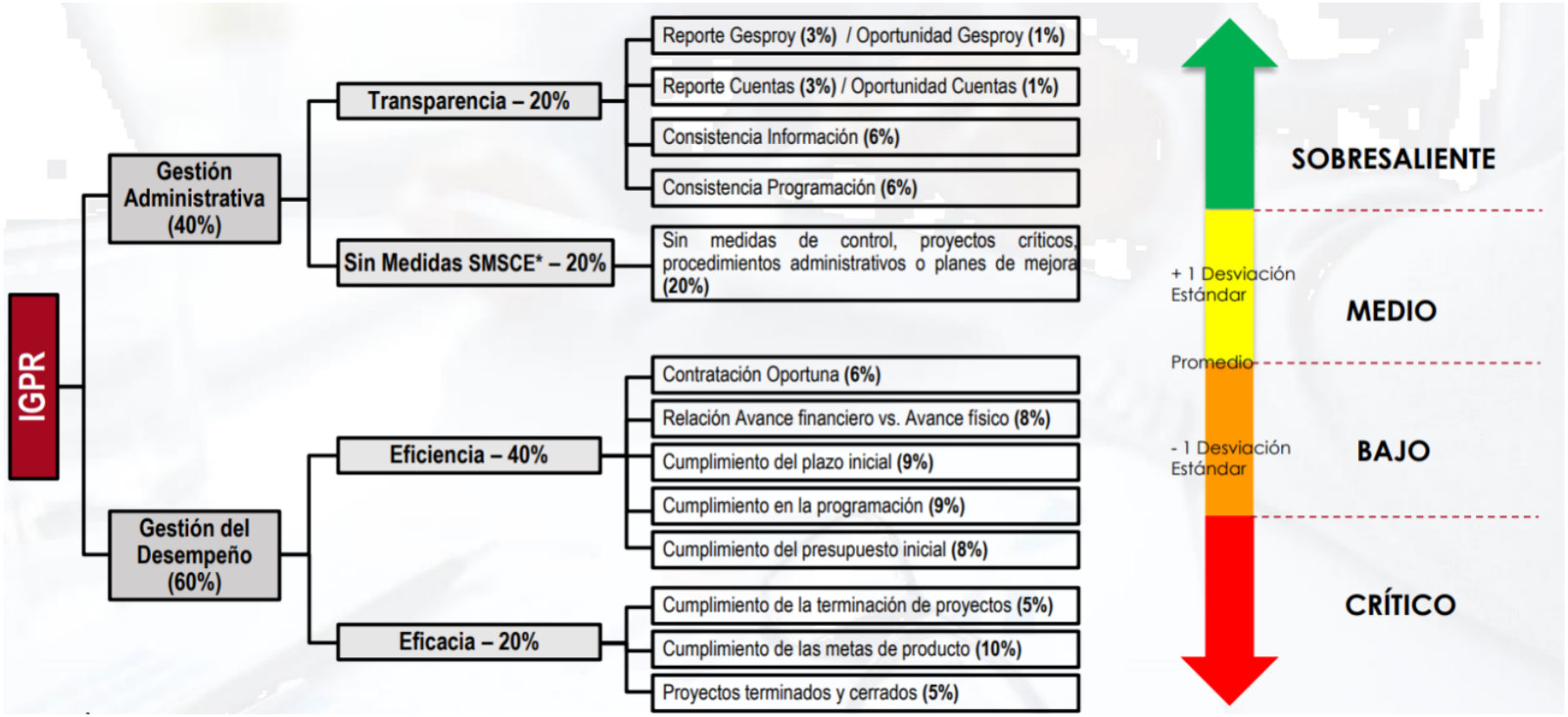
- Codificación de variables binarias a partir de las categorías de cada variable.
- Cálculo del *weight of evidence* (WOE) de las categorías, con respecto a cada variable objetivo.

- Tratamiento de valores faltantes.
- Acotamiento de cada variable para evitar la influencia negativa de valores atípicos.
- Normalización de las variables.

- Formación de un solo texto a partir de todas las variables.
- Limpieza del texto (minúsculas, eliminar puntuación, conectores y *stop words*).

Desempeño en la ejecución: Metodología actual SGR

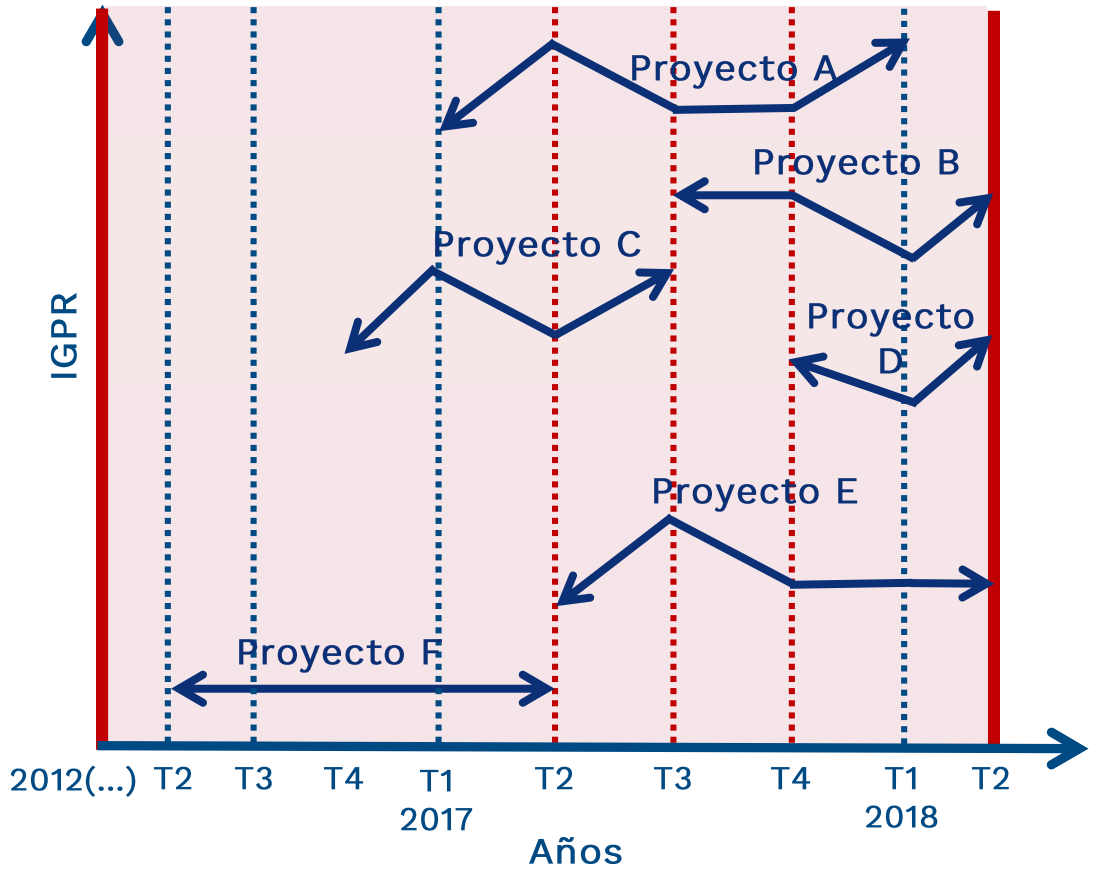
La metodología actual de evaluación del Sistema General de Regalías se basa en el cálculo del Índice de Gestión de Proyectos de Regalías



Fuente: SMSCE DNP

Cálculo de las variables objetivo

A partir de las mediciones del IGPR para los proyectos durante la ventana de observación, se calculan varias alternativas que midan el desempeño de cada proyecto.



Las alternativas para calcular la variable objetivo involucran variación en:

- Cálculos sobre el IGPR en la ventana de observación.
- Proyectos que son tenidos en cuenta para el cálculo.
- Utilización del IGPR completo o de solo algunos componentes.

Variable objetivo considerada para el desarrollo del modelo

Se define como proyecto “malo”, o que tuvo dificultades, a cualquier proyecto que haya requerido de por lo menos una medida del SMSCE durante la ventana de observación.

Intervenciones que puede realizar el SMSCE:

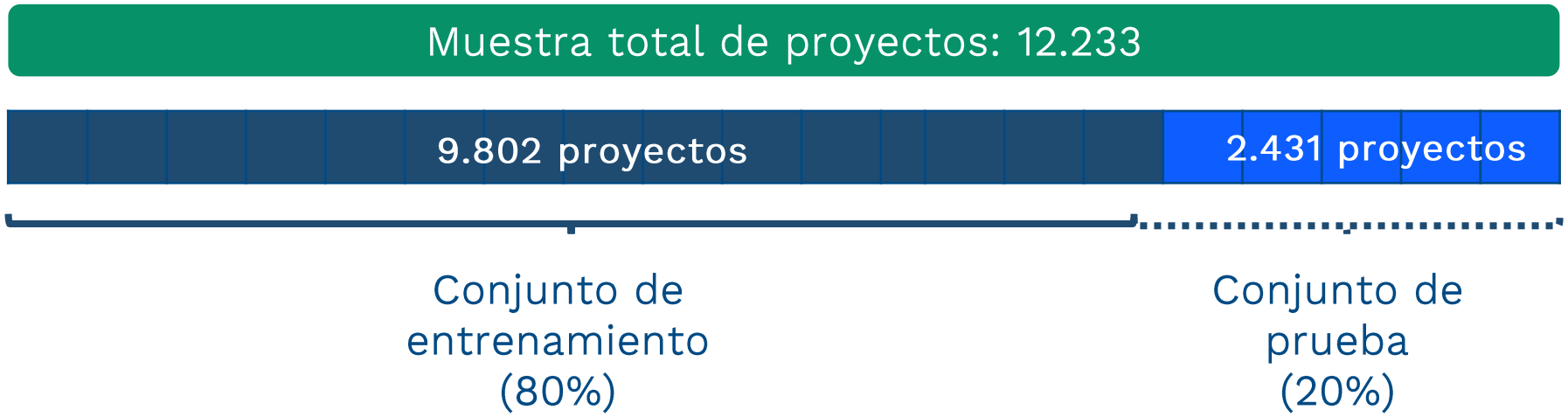
- Plan de mejora
- PAP - Procedimientos administrativos preventivo
- PACS – Procedimiento administrativo correctivo y/o sancionatorio
- Proyecto crítico
- Medidas de suspensión

SMSCE: Sistema de Monitoreo, Seguimiento, Control y Evaluación

4. Entrenamiento de los modelos

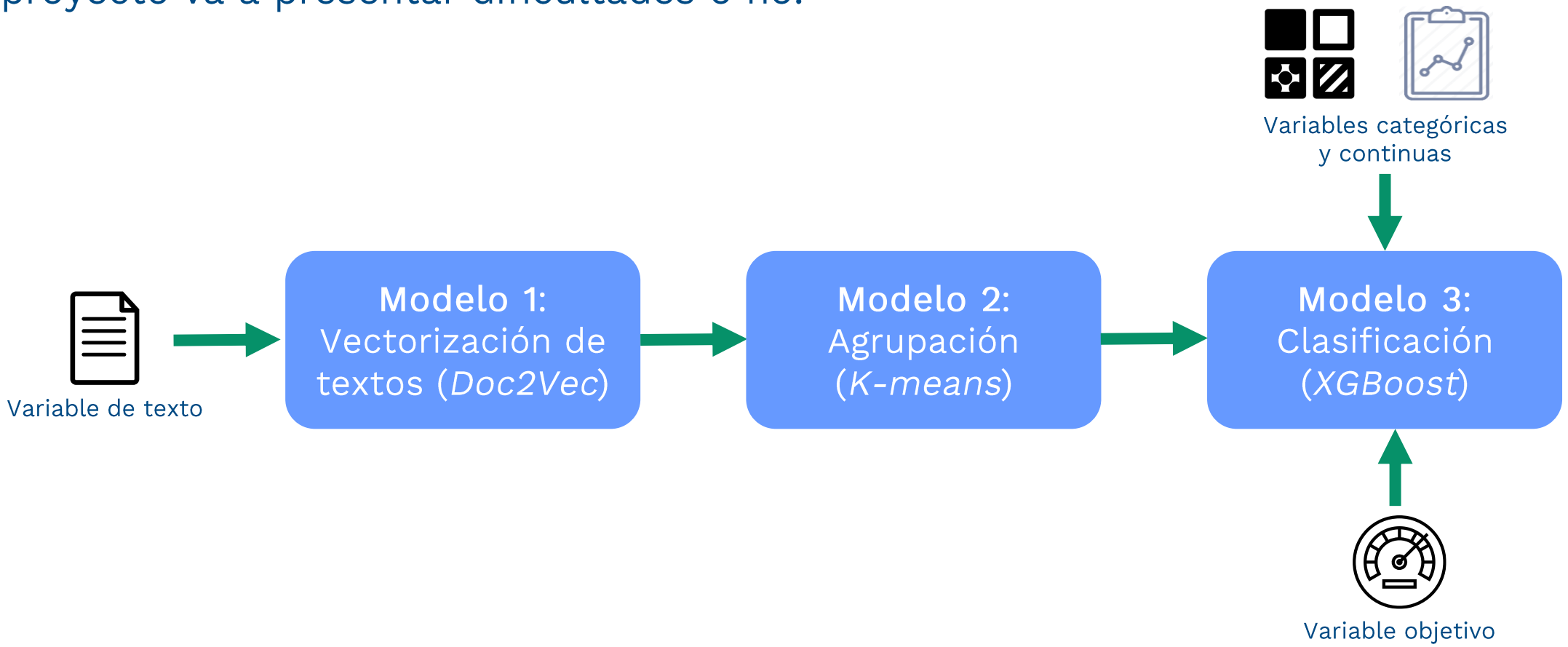
Segmentación de los datos para entrenamiento del algoritmo

El conjunto de datos se segmenta aleatoriamente para validar el aprendizaje del algoritmo. La iteración del proceso aumenta progresivamente la precisión del algoritmo



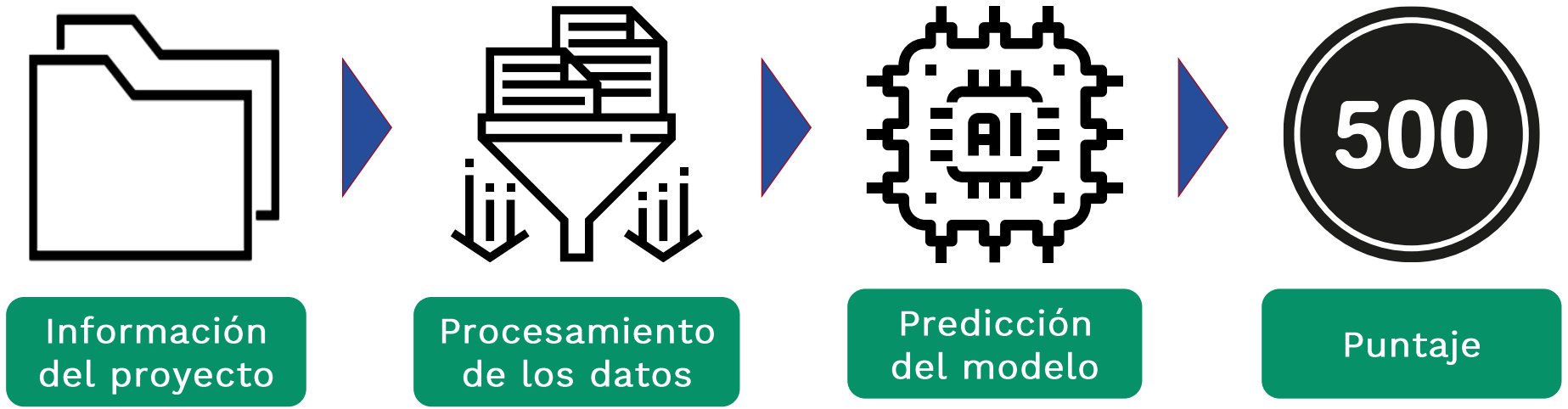
Flujo de datos en el entrenamiento de los modelos

Tres modelos (2 no supervisados y 1 supervisado) fueron desarrollados para determinar si un proyecto va a presentar dificultades o no.



Proceso de evaluación de un proyecto

El modelo de clasificación asigna un puntaje a cada proyecto, que indica la probabilidad de que sea “bueno”. Es decir, no vaya a requerir medidas de la SMSCE.



5. Resultados y conclusiones

Scorecard: Datos de entrenamiento

El indicador Kolmogorov–Smirnov (KS) indica el grado de separación entre dos distribuciones. A medida que el KS es mayor, el modelo puede discriminar mejor entre proyectos “buenos” y “malos”.

Rango de score	# de proyectos	% Total	Malos	Buenos	Tasa de malos	Tasa de buenos	KS
> 901	1093	11.05	72	1021	6.59%	93.41%	9.94%
880 - 901	1345	13.60	116	1229	8.62%	91.38%	20.55%
857 - 880	1261	12.75	144	1117	11.42%	88.58%	28.43%
819 - 857	1219	12.33	206	1013	16.90%	83.10%	32.11%
783 - 819	1247	12.61	251	996	20.13%	79.87%	33.51%
722 - 783	1247	12.61	309	938	24.78%	75.22%	31.49%
622 – 722	1229	12.43	397	832	32.30%	67.70%	24.06%
<= 622	1249	12.63	684	565	54.76%	45.24%	0.00%

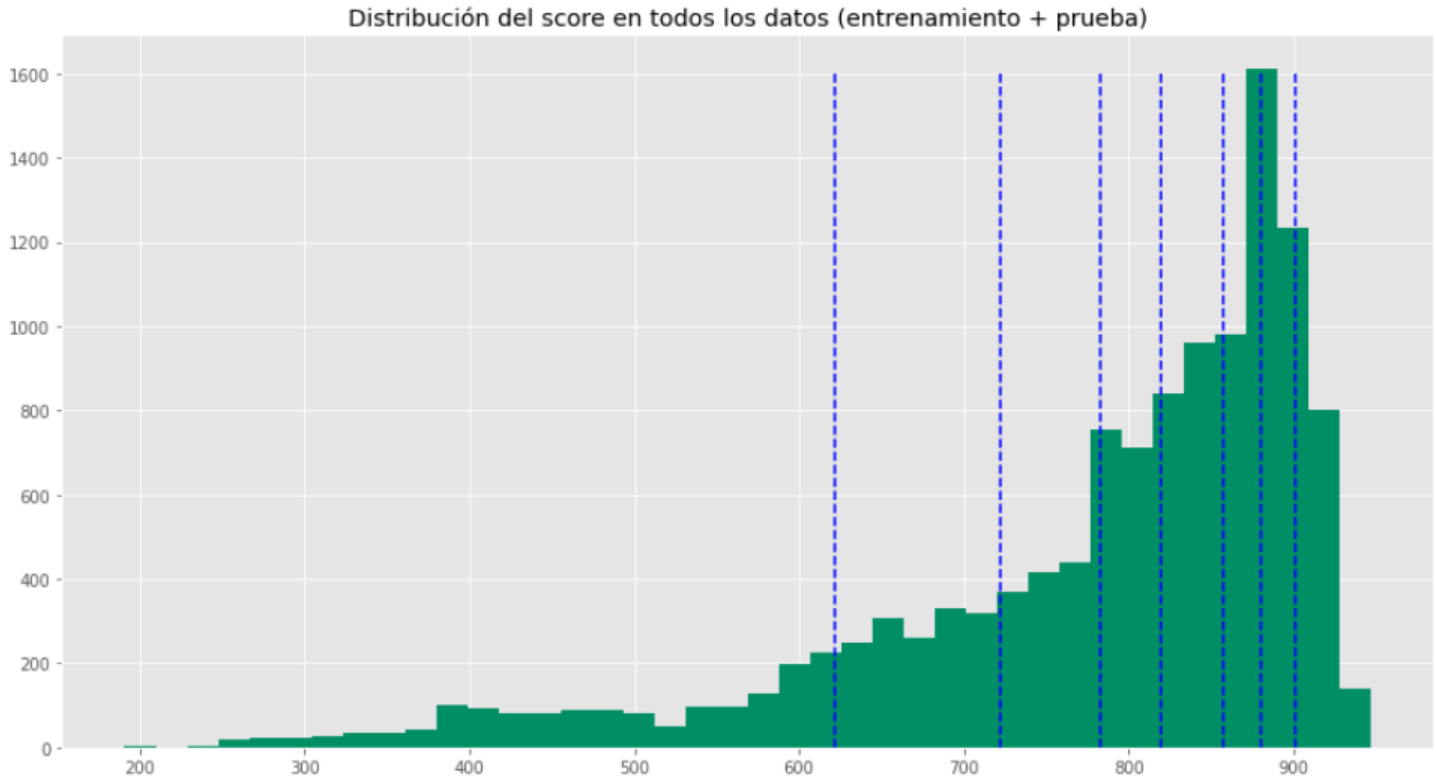
Scorecard: Datos de prueba

El desempeño del modelo se mantiene al ser utilizado en datos que no ha visto anteriormente.

Rango de score	# de proyectos	% Total	Malos	Buenos	Tasa de malos	Tasa de buenos	KS
> 901	274	11.17	28	246	10.22%	89.78%	7.74%
880 - 901	305	12.43	33	272	10.82%	89.18%	15.92%
857 - 880	338	13.78	41	297	12.13%	87.87%	23.94%
819 - 857	323	13.17	55	268	17.03%	82.97%	27.87%
783 - 819	265	10.80	42	223	15.85%	84.15%	31.83%
722 - 783	313	12.76	82	231	26.20%	73.80%	28.86%
622 – 722	300	12.23	93	207	31.00%	69.00%	22.61%
<= 622	335	13.66	170	165	50.75%	49.25%	0.00%

Distribución de los puntajes

Se busca que los puntajes tengan una distribución normal. En este caso los puntajes muestran un sesgo hacia los valores más altos, que puede ser corregido con una alineación.



Conclusiones

- Mediante la aplicación de técnicas de inteligencia artificial es posible el aprovechamiento de los datos disponibles de cada proyecto para generar un indicador distinto a los ya existentes.

- Los modelos logran **reducir la incertidumbre en la evaluación de un proyecto**, a partir de la predicción de la ocurrencia de dificultades que actualmente no pueden inferirse en su formulación.

- El **modelo de clasificación es adaptable** a otro umbral o criterio de clasificación diferente al IG

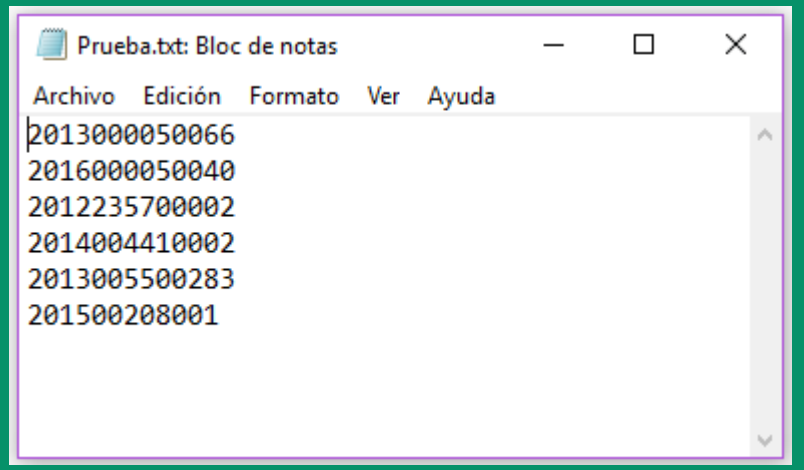
A partir de los modelos desarrollados en este proyecto es posible:

- Combinar los puntajes con conocimiento experto para definir nuevos criterios en la evaluación y seguimiento de los proyectos del SGR.
- Identificar proyectos que tengan textos muy similares en su

6. Cómo utilizar el modelo

Cómo utilizar el modelo

1. Cree un archivo de texto con los códigos BPIN de los proyectos que desea evaluar.

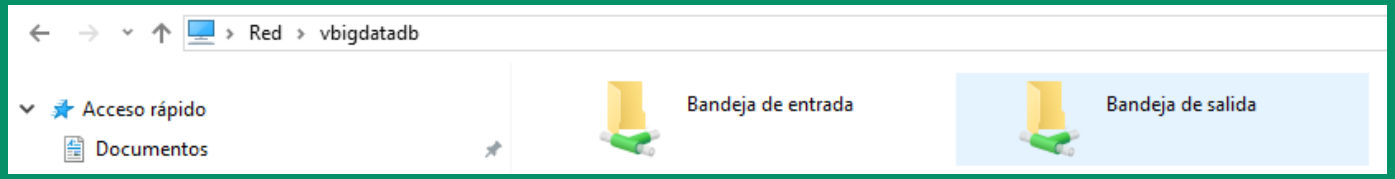


Nota: 1 BPIN por renglón. Sin comas, espacios ni ningún otro carácter.

Cómo utilizar el modelo

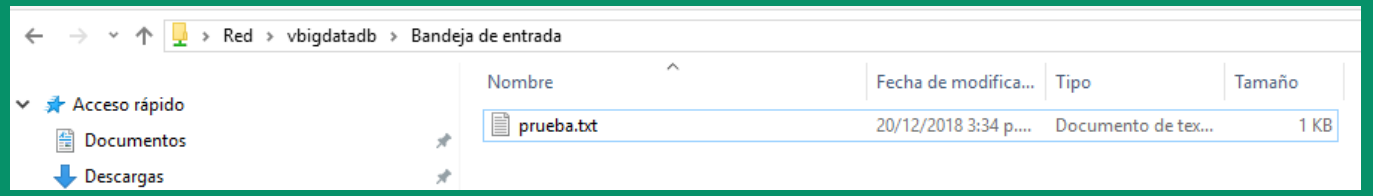
2. Ingrese desde su computador del DNP a la dirección \\vbigdatadb

Allí encontrará dos carpetas.



Cómo utilizar el modelo

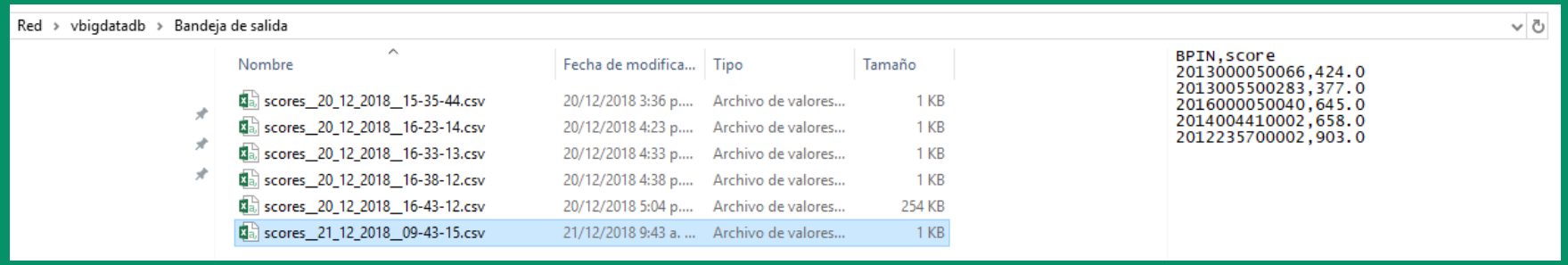
3. Ingrese a la carpeta “Bandeja de entrada”, y deje ahí el archivo de texto que creó en el paso 1.



Cómo utilizar el modelo

4. En un tiempo de máximo 5 minutos, el servidor detectará los archivos (.txt) que estén en la bandeja de entrada, y los calificará.

Una vez se termine el proceso, un nuevo archivo aparecerá en la carpeta “Bandeja de salida”.



Nota: Los archivos que se ponen en la bandeja de entrada son borrados luego de ser calificados.

Cómo utilizar el modelo

5. Lleve el archivo recién generado a la ubicación que desee. En este archivo viene cada BPIN (que haya sido encontrado en las bases de datos) con su respectivo puntaje.

Con esta información se pueden realizar luego los análisis pertinentes.

Nota: Los archivos que quedan en la bandeja de salida se borrarán cada cierto tiempo, por lo que se recomienda moverlos de ahí para evitar pérdida de la información.

El futuro es de todos

DNP
Departamento
Nacional de Planeación

Esta presentación es propiedad intelectual controlada y producida por el Departamento Nacional de Planeación