Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

ANALÍTICA PREDICTIVA DE PROYECTOS DE INVERSIÓN FINANCIADOS CON RECURSOS SGR

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Desarrollo Digital - Unidad de Científicos de Datos • Dirección de Vigilancia de las Regalías
Sector	Territorial
Tecnologías utilizadas	Python
Fuentes de datos	• GESPROY • IGPR • Mapas De inversiones

Contenido

1. Presentación

En este proyecto se desarrolló un modelo para predecir qué proyectos de inversión financiados con recursos de regalías podrían tener dificultades durante su ejecución, ya sea en el cumplimiento de los alcances en tiempos y en costos (cambios en su valor). El documento presenta la búsqueda, tratamiento de bases de datos y las variables escogidas que serán insumo para el modelo de predicción. Además presenta los resultados de los modelos de clasificación utilizando distintas variables dependientes y con el modelo XgBoost, con el cual se obtuvieron los mejores resultados. Por último, se presenta cómo utilizar la herramienta de visualización de resultados para nuevos proyectos de inversión.

This project developed a model to predict which investment projects financed with royalties could have difficulties during their execution due to excessive extensions in its original dates or surges in its monetary values. This document presents which datasets and variables where used to build the machine learning model and the results with the XgBoost model, the one with the best results. Additionally, it shows how to use a program to generate results with new investment projects.

2. Objetivos del proyecto

General

Desarrollar un modelo para predecir qué proyectos de inversión financiados con recursos de regalías podrían tener dificultades durante su ejecución y en la medición del desempeño para el cumplimiento de los alcances en tiempos y en costos (cambios en su valor).

Específicos

1. Definir las bases de datos y variables que serán insumo para el modelo de predicción
2. Desarrollar un modelo de predicción de los proyectos de inversión que podrían tener dificultades en tiempos de ejecución O cambios de valor monetario significativos
3. Crear una herramienta de visualización de resultados que sirva para calcular la predicción de nuevos proyectos de inversión

3. Metodología

Adecuación de bases de datos

El primer paso consiste en trabajar con las fuentes de datos para seleccionar el número de registros, que en este caso son los proyectos de inversión, de cada base de datos para facilitar la unificación de las variables seleccionadas para el modelo de predicción. Las tres fuentes de información son IGPR, GESPROY y Mapas de Inversión. Tanto IGPR como GESTRPOY fueron entregados por los miembros de la DVR en formato de Excel y fue necesario conectarse al servidor de Microsoft SQL Server de DNP para acceder a Mapas de Inversión.

La unidad de análisis se seleccionó de la base de datos de IGPR del último trimestre de 2020, ya que cuenta con la información más actualizada del índice IGPR y se puede pegar fácilmente con la base de GESPROY de enero de 2021. Esta base contiene 9.611 registros y, a partir del BPIN (el código único de cada proyecto), se filtraron las bases de GESPROY y Mapas de Inversión para que tuvieran los mismos registros de IGPR.

Dado que se escogió el cuarto trimestre de 2020 de la base de IGPR, se seleccionó la base de datos de GESPROY del mes de enero de 2021. Adicionalmente, las variables de Mapas de Inversión seleccionadas son de tipo texto y descripciones de los proyectos, así que pueden ser utilizadas para cualquier fecha.

Selección de variables

Para poder desarrollar un modelo de predicción, es necesario primero identificar cuáles son las variables que podrían utilizarse. Por ello se revisaron las bases de datos de IGPR, GESPROY y Mapas de Inversión y se extrajeron los datos de interés. En términos generales, los modelos de aprendizaje de máquina supervisados se construyen con variables independientes y dependientes. Las dependientes son aquellas que uno busca predecir con el modelo. En el caso de este proyecto, se quiere predecir si un proyecto de inversión puede o no tener algún problema durante su ejecución, por lo cual se utilizarían como dependientes el índice IGPR o si un proyecto tuvo o no retrasos en tiempos o cambios muy grandes en su valor inicial. Por el otro lado, las independientes son todas aquellas que nos pueden explicar por qué el proyecto pudo tener un problema. Por ejemplo, la región, el monto total de la inversión, el origen de los recursos, la entidad ejecutora, el sector SUIFP, entre otros.

En las bases de datos de IGPR, GESPROY y Mapas de inversión, se encontraron variables de tipos categóricos, numéricos, fecha y texto, que contienen la información que se consideró relevante para el análisis. Estas variables no se utilizarán todas como se encuentran en su formato original y serán adecuadas según el modelo de predicción seleccionado. Los tipos de las variables en su formato original son los siguientes:

- **Categórico**: contiene valores finitos que describen los proyectos. Sobre estas variables no se pueden hacer cálculos matemáticos como promedios, sumas o restas. Por ejemplo, variables de región o sector de un proyecto. Dependiendo del modelo de predicción utilizado, es posible que sea necesario convertir cada categoría de estas variables a variables binarias, donde 1 represente si sucede la categoría y 0 de lo contrario.
- **Numérico discretao o continua**: estas son variables que representan con valores numéricos algún aspecto del proyecto. Sobre ellas se pueden hacer cálculos matemáticos como promedios, sumas o restas. Por ejemplo, el valor SGR o el IGPR. Es posible que sea necesario normalizar estas variables para evitar fluctuaciones demasiado altas de sus valores.
- **Fecha**: este tipo de variables son aquellas que representan la fecha con información del año, mes y día y que pueden ser transformadas para calcular tiempos de ejecución de proyectos y diferencias entre distintos momentos del tiempo.
- **Texto**: las variables de texto escogidas son aquellas que describen características de los proyectos, por lo cual cada valor de estas variables es único para cada proyecto. Se utilizaron con el fin de agrupar los proyectos con textos similares y crear variables categóricas de la similitud. Para agrupar los proyectos a partir del texto se unificaron los textos de las variables *NombreProyecto, ProblemasCentral, Descripcion* y *ObjetivoGeneral*, se vectorizaron con la metodología Doc2Vec y luego se agruparon con el algoritmo de KMeans.

Adecuación de las variables independientes y dependientes

El primer paso consiste en convertir las variables a un formato adecuado para ser utilizadas en los modelos de predicción. En el caso de las variables independientes, las variables numéricas se estandarizaron con la resta de la media y división por su desviación estándar[1]. Las variables de fecha, que representan la diferencia entre dos fechas en días, se transformaron a numéricas y también se estandarizaron. Para los modelos que requerían transformar las variables categóricas a binarias (*dummies*), se decidió utilizar la metodología de reducción de dimensionalidad Análisis

[1] La estandarización no hace falta para el modelo XgBoost, pero no afecta los resultados. La estandarización se hizo para probar otros modelos que sí requieren este paso

de Correspondencia Múltiple (MCA, en inglés) con el fin de reducir el número de variables categóricas que se ingresarían al modelo de predicción. De esta manera, se redujeron las variables binarias a 30, las cuales explican el 90% de la variabilidad de las variables originales.

La Tabla 1 muestra las variables independientes que se utilizaron en los modelos de predicción.

Tabla 1. Variables independientes

Variable	Descripción	Tipo actual
Valor SGR	Valor de la inversión con recursos de regalías	Numérica
Total proyecto	Valor total de la inversión, incluyendo regalías	Numérica
Valor SGR / Total proyecto	Valor SGR de la inversión como proporción del valor total	Numérica
Longitud inicial del proyecto	Diferencia en días de la fecha final y la fecha de inicio del proyecto según la programación inicial	Numérica
Retraso en el inicio del proyecto	Diferencia en días entre la fecha de inicio actual y la fecha de inicio original del proyecto	Numérica
Demora en el inicio del proyecto	Diferencia en días entre la fecha de inicio actual y la fecha de aprobación del proyecto	Numérica
Valor SGR / Longitud inicial	Cociente entre el valor del proyecto con recursos de regalías y la longitud inicial del proyecto	Numérica
Sede	Indica el lugar geográfico donde se encuentra el proyecto de inversión	Categórica
Entidad ejecutora	Entidad que ejecuta el proyecto. Las entidades se redujeron a las 25 más frecuentes y el resto se catalogó como "otra"	Categórica
Sector-SUIFP	Sector SUIFP al que pertenece el proyecto de inversión	Categórica
PGN	Indica si el proyecto de inversión tiene recursos pertenecientes al PGN	Categórica
Valor no SUIFP	Indica si el proyecto de inversión tiene recursos no pertenecientes al SUIFP	Categórica
SGP	Indica si el proyecto de inversión tiene recursos pertenecientes al SGP	Categórica
Enfoque diferencial	Indica si el proyecto de inversión tiene un enfoque sobre una población en particular o no	Categórica

Por el otro lado, las variables dependientes representan las dificultades que pudo haber tenido un proyecto de inversión por cuenta de demoras en tiempos de ejecución, incrementos significativos en el valor monetario inicial o según el índice IGPR. La *Tabla 2* contiene las variables dependientes utilizadas en los modelos de predicción.

Tabla 2. Variables dependientes

Variable	Descripción	Comentarios
Cambios significativos en tiempos de ejecución iniciales	Muestra si el proyecto de inversión tuvo incrementos en X por ciento en su tiempo de ejecución inicial	Se creó una variable por cada valor de X: 20%, 50% y 80%
Cambios significativos en el valor monetario original de la inversión	Muestra si el proyecto de inversión tuvo incrementos en X por ciento en valor monetario inicial	Se creó una variable por cada valor de X: 20%, 50% y 80%
Cambios significativos en tiempos y en valores monetarios	Muestra si el proyecto tuvo cambios en tiempos de ejecución y valores monetarios de acuerdo a las definiciones de arriba	Se utilizan los valores de 20%, 50% y 80%
Cambios significativos en tiempos o en valores monetarios	Muestra si el proyecto tuvo cambios en tiempos de ejecución o valores monetarios de acuerdo a las definiciones de arriba	Se utilizan los valores de 20%, 50% y 80%
IGPR - rangos	Muestra la calificación más baja del IGPR de un proyecto según los siguientes rangos: alto (82 a 100), medio (62 a 81), bajo (32 a 61) e insuficiente (0 a 31)	
IGPR - binaria	Muestra si un proyecto de inversión tuvo un IGPR menor a un valor específico	Se utilizaron los valores: 32% y 62%

Modelos de predicción

Los modelos de predicción utilizados fueron XgBoost, *Random Forest*, Support Vector Machine, KNN (vecino más cercano) y logístico. Dado que el modelo XgBoost obtuvo los mejores resultados, se escogió como el modelo principal y los resultados se presentan con este modelo. Estos modelos se pueden utilizar en el caso de los problemas de clasificación, donde se busca predecir los valores de variables categóricas como las dependientes presentadas anteriormente.

Como los modelos de predicción son supervisados, contienen variables independientes y dependientes, se dividió la base de datos en secciones de entrenamiento y prueba. Con el fin de crear varias divisiones de entrenamiento y

prueba, se utilizó la metodología *K-Fold Cross Validation*, que consiste en partir la base en K segmentos y realizar K modelos, siempre utilizando K-1 segmentos para entrenar y 1 segmento para validar. Se escogió que K sea 5, es decir, que para cada modelo haya 5 grupos de entrenamiento y prueba distintos. Luego de correr cada modelo K veces, se promedian las métricas con los resultados.

Interpretación de resultados

Si bien se utilizan distintos modelos de predicción, todos los resultados se interpretan de la misma manera, con matrices de confusión y las métricas de *recall, precision* y *accuracy*. Como este es un ejercicio de predicción binomial[2] (se predice si un proyecto de inversión tiene o no una dificultad), la matriz de confusión tiene el tamaño 2X2 y muestra los resultados de la siguiente manera, suponiendo que los datos que se quieren predecir tienen los valores "positivo" y "negativo" (Tabla 3):

Tabla 3. Matriz de confusión general

		Categoría predicha	
		Negativo	**Positivo**
Categoría Original	**Negativo**	Verdaderos negativos	Falsos positivos
	Positivo	Falsos negativos	Verdaderos positivos

La Tabla 4 muestra la matriz de confusión en el caso de la predicción de proyectos de inversión con potenciales problemas.

Tabla 4. Matriz de confusión para el caso de predicción de rendimiento de proyecto de inversión

		Categoría predicha	
		No problema	**Problema**
Categoría Original	**No problema**	Predicción correcta: proyecto no tiene dificultad	Predicción incorrecta: proyecto tiene dificultad
	Problema	Predicción incorrecta: proyecto no tiene dificultad	Predicción correcta: proyecto tiene dificultad

Vale la pena recordar que la dificultad de un proyecto de inversión se define según las variables dependientes de la Sección 0.Las tres métricas que se utilizaron para validar los resultados de las predicciones son las de *accuracy*, *recall* y *precision.* Estas se calculan de la siguiente manera, a partir de la información en la matriz de confusión (Tabla 3).

Tabla 5. Métricas de rendimiento de modelos

Métrica	Cálculo
Acuracy	$\frac{verdaderos\ positivos + verdaderos\ negativos}{Total}$
Recall	$\frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos}$
Precision	$\frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos}$

[2] El modelo que busca predecir los rangos del IGPR es multinomial y no binomial. Sin embargo, como los resultados no fueron satisfactorios no se explica su metodología de interpretación de resultados

El *accuracy* mide el porcentaje de proyectos de inversión que se calcularon correctamente como "sin dificultad" y "con dificultad". Sin embargo, en caso de que las muestra sea desbalanceada (por ejemplo, que los proyectos problemáticos sean muy inferiores a los no problemáticos), esta métrica podría mostrar un modelo exitoso de forma errónea al predecir muchos proyectos sin dificultades pero sin predecir aquellos que sí tuvieron difcultades. Por esta razón se utiliza la métrica de *recall*, que calcula el porcentaje de proyectos con dificultades (verdaderos positivos) que sí se lograron predecir entre todos los proyectos con dificutlades. Por ejemplo, si se predicen 300 proyectos con dificultades y el total de proyectos con dificultades son 1.000, entonces el *recall* sería 30%. Por último, *precision* mide el porcentaje de proyectos con dificultades predichos que en realidad tuvieron dificultades. Es decir, si un modelo predice 1.000 proyectos con dificultades, pero 500 no tuvieron, es decir, se predijeron incorrectamente como problemáticos (son falsos positivos), entonces el *precision* sería 50%.

Saber cuál es el mejor modelo estimado depende entonces del *recall* y *precision*. El mejor modelo sería uno con *recall* de 100% y *precision* de 100%. Sin embargo, estas dos métricas casi siempre entran en conflicto. Si se logra predecir una gran cantidad de proyectos problemáticos (se tendría un *recall* alto), es muy probable que se tenga una cantidad alta de falsos positivos, por lo cual disminuiría el *precision*. Por esta razón la decisión final del modelo utilizado para la predicción depende si se tolera tener una cantidad alta de predicción de proyectos problemáticos con muchos falsos positivos o si se desea tener mejor una cantidad baja de proyectos problemáticos predichos con más exactitud.

Umbral de éxito

Un aspecto importante de señalar es el del umbral de éxito de los modelos de clasificación. El siguiente ejemplo se usará para explicar cómo funciona el umbral. La Tabla 6 muestra un resultado hipotético con los resultados de la predicción de proyectos de inversión con dificultades.

Tabla 6. Resultado hipotético de predicción proyectos con problemas

Proyecto	Probabilidad de proyecto con dificultad
1	0.674702
2	0.0193333
3	0.283917
4	0.005
5	0.562079

La columna "Probabilidad de proyecto con dificultad" contiene las probabilidades que haya un problema en un proyecto de inversión. Como se trata de un modelo de clasificación, se tiene que definir cuándo se considera que haya un problema o no en un proyecto de inversión. Por ello se supone una probabilidad mínima de éxito, o de que haya un problema. Usualmente la probabilidad mínima es 0.5 y en este caso, con esta probabilidad, se predeciría que los proyectos 1 y 5 serían problemáticos. Sin embargo, este umbral puede aumentar para considerar únicamente como delitos los casos con las probabilidades más altas. Por ejemplo, si el umbral mínimo es de 0.6, entonces solo el proyecto 1 sería predicho como problemático.

Es importante mencionar el umbral de éxito para la predicción porque al aumentarlo se consideran como problemáticos únicamente los proyectos con mayor probabilidad, lo cual incrementa la métrica *precision* pero disminuye el *recall*. Es

decir, al aumentar el umbral de éxito se predicen menos proyectos con dificultades correctamente, pero los proyectos predichos tienen menos falsos positivos en sus resultados. Los resultados se presentan con distintos valores del umbral de éxito para verificar la variación en las métricas *recall* y *precision*.

4. Resultados

En esta sección se presentan las variables encontradas para el análisis. Se dividen por cada tipo de variable: categórico, numérico, fecha y texto. También se presentan las variables independientes que se buscan predecir con el modelo.

El criterio para escoger las variables independientes se basó en qué tanto podrían explicar las dificultades que podría tener un proyecto de inversión. Por eso se escogieron variables relacionadas con la zona geográfica, la entidad ejecutora, el sector SUIFP, los valores y tipos de inversión, la duración de los proyectos y también la agrupación de los textos que describen los proyectos (explicado en la sección 0).

Las variables escogidas y creadas presentadas en esta sección tienen que ser adaptadas dependiendo del modelo de predicción escogido y no todas las presentadas aquí tienen que ser utilizadas, lo cual se evaluará en el transcurso del proyecto. Adicionalmente, es posible también que se descubran nuevas variables o maneras de crear variables que podrían ingresarse al modelo de predicción.

Es importante mencionar que se tuvieron que eliminar 917 proyectos de inversión que no contaban con información en las variables de fecha de GESPROY, las cuales son esenciales para los modelos de predicción. Adicionalmente, se estimaron también los modelos sin los proyectos que iniciaron en 2019 y 2020, dado que muy probablemente no se detectarían tantos problemas en los proyectos más recientes justo por el hecho de ser los más nuevos. Si bien con este filtro de ambos años mejoran los resultados, se pierde información valiosa que podría servir para las predicciones de los proyectos que iniciaron luego de 2020.

Los resultados de los modelos de predicción del IGPR se encuentran en los anexos, ya que no son tan buenos como el resto. Se muestran los resultados cuando la variable dependiente es de tipo binario únicamente, dado que la predicción por los rangos del IGPR no dio resultados satisfactorios. En todo caso, no se recomienda utilizar los modelos de predicción de IGPR.

Variables categóricas

La Tabla 7 presenta las variables categóricas encontradas en las bases de IGPR y GESPROY. Vale la pena mencionar que hay variables que contienen la misma información y será necesaria una sola para el modelo. Todas las variables de esta tabla se utilizarían como independientes. En general, las variables de la *Tabla 7* contienen información del proyecto sobre la región, la entidad ejecutora, el sector SUIFP, el destino de los recursos y el enfoque diferencial.

La variable con la mayor cantidad de categorías es “Entidad ejecutora”, tiene 1.269, y tendrían que reducirse a unas 30 e incluir el valor “otros” en caso de que se utilice. El resto de las variables contiene una cantidad mucho más reducida y pueden convertirse a binarias en los modelos de predicción.

Tabla 7. Variables categóricas

Base de datos	Nombre de variable	Valores y frecuencias	Operación sobre variable
IGPR	Región	Caribe: 3045 Pacifico: 1688 Centro Oriente: 1386 Centro Sur: 1237 Del Llano: 1138 Eje Cafetero: 881 Nacional: 220 Cormagdalena: 2	Convertir en variables binarias
IGPR	Sede	Pacífico :1688 Caribe I: 1548 Caribe Ii: 1497 Centro Oriente: 1386 Centro Sur 1237 Llanos: 1138 Eje Cafetero: 881 Entidades De Orden Nacional: 222	Convertir en variables binarias
IGPR	Departamento	32 departamentos, Bogotá y orden nacional	Convertir en variables binarias
IGPR	Entidad ejecutora	1269 entidades	Reducir las categorías y convertir a binarias
IGPR	Tipo ejecutor	Municipio: 6537 Departamento: 1851 Otros: 1114 CAR: 95	Convertir en variables binarias
IGPR	Tipo ejecutor detallado	Departamento: 1851 G5- Nivel Bajo: 1710 G1- Nivel Alto: 1404 G4- Nivel Medio Bajo: 1262 G3- Nivel Medio: 1077 G2- Nivel Medio Alto: 1000 Otros: 530 Institución De Educación Superior: 281 E.S.P.: 268 CAR: 95 Ciudades: 85 E.S.E.: 34	Convertir en variables binarias

Base de datos	Nombre de variable	Valores y frecuencias	Operación sobre variable
GESPROY	SECTOR-SUIFP	Transporte: 3293 Vivienda, Ciudad Y Territorio: 1397 Deporte Y Recreación: 972 Educación: 881 Agricultura Y Desarrollo Rural: 707 Ciencia, Tecnología E Innovación: 520 Ambiente Y Desarrollo Sostenible: 437 Minas Y Energía : 361 Salud Y Protección Social: 289 Inclusión Social Y Reconciliación: 194 Cultura: 150 Gobierno Territorial: 142 Comercio, Industria Y Turismo: 70 Interior : 42 Presidencia De La República: 34 Tecnologías De La Información Y Las Comunicaciones: 33 Justicia Y Del Derecho: 25 Planeación: 20 Trabajo: 12 Información Estadística: 8 Defensa: 8 Empleo Público: 1 Fiscalía: 1	Convertir en variables binarias
GESPROY	DESTINO RECURSOS	EJECUCION: 6018 ESTUDIOS: 385	Convertir en variables binarias
GESPROY	ENFOQUE DIFERENCIAL	Sin Enfoque Diferencial: 9144 Población Afrocolombiana: 150 Pueblos indígenas: 148 Población Indígena: 118 Población Raizal: 33 Pueblo Rrom: 2 Población ROM: 2	Convertir en variables binarias

Variables numéricas

Las variables numéricas se consiguieron de la base de datos de GESPROY y se presentan en la *Tabla 8*, junto con algunas estadísticas descriptivas. Estas variables representan los valores monetarios de la inversión asignada al proyecto en distintas asignaciones y el total. Varias variables contienen pocos valores mayores a 0, así que tendrán un promedio muy cercano a 0. Es posible que varias de estas variables se transformen a binarias, con el fin de mostrar si un proyecto tuvo o no una inversión de algún tipo de asignación. Aquellas que se utilicen como numéricas en el modelo muy probablemente tendrán que ser normalizadas.

Estas variables se utilizarán como independientes, ya que podrían explicar el éxito final de un proyecto de inversión, ya sea por el destino de los recursos o por el monto de los recursos.

Adicional a estas variables, se construyó la variable $\frac{VALOR\ SGR}{tiempo\ de\ ejecución}$, el cociente entre el valor SGR y el tiempo de ejecución en meses, para identificar si el proyecto tiene discrepancias muy altas entre el valor asignado y los tiempos de ejecución programados. Sin embargo, es necesario definir con la DVR los tiempos de ejecución "normales" de un proyecto según el sector SUIFP.

Tabla 8. Variables numéricas

Variable	Media	Desviación estándar	Mínimo	Mediana	Máximo
VALOR SGR	3.901.808.489	8.411.748.356	160.000	1.349.226.903	200.000.000.000
VALOR NACIÓN	508.795.731	34.420.862.539	0	0	3.357.443.454.052
VALOR OTROS	475.531.356	4.482.735.548	0	0	172.610.955.963
TOTAL PROYECTO	4.886.135.576	37.651.917.702	160.000	1.484.394.103	3.527.443.454.052
DIRECTAS	1.137.780.307	4.369.951.708	0	0	155.393.200.915
FCR_40	405.746.133	975.967.444	0	0	35.409.231.424
FCR_60	843.972.273	4.426.976.870	0	0	149.655.760.164
FDR	623.890.268	3.715.320.357	0	0	126.975.384.332
FCTEI	400.317.146	2.495.573.401	0	0	64.541.936.186
CORMAGDALENA	33.992.518	1.079.213.655	0	0	90.000.000.000
INCENTIVO A LA PRODUCCIÓN	27.738.389	364.667.796	0	0	20.286.705.899
ASIGNACION PAZ - AP50	304.426.920	2.503.190.157	0	0	86.183.934.966
ASIGNACION PAZ APFDR50	60.789.386	1.037.554.134	0	0	45.287.111.018
ASIGNACION PAZ FONPET	3.909.710	72.786.441	0	0	3.172.980.952

Asignaciones Directas - Gestión del riesgo, adaptación al cambio climático o situaciones de emergencia	51.875.156	1.924.213.603	0	0	185.093.578.907
Incentivo a la Producción - 30% Rendimientos Financieros	7.370.283	164.671.051	0	0	9.732.146.040

Variables de fecha

Las variables de fecha se consiguieron de la base de datos de GESPROY y se encuentran en el formato AÑO-MES-DÍA-HORA. No se utilizarán en su formato original en el modelo de predicción y con ellas se crearon las siguientes variables:

- **Longitud programada del proyecto**: variable independiente
- **Diferencia entre la fecha final programada y la fecha final actual**: variable dependiente
- **Diferencia entre la fecha inicial programada y la fecha inicial actual**: variable independiente
- **Diferencia entre la fecha de aprobación y la fecha inicial ejecutada**): variable independiente

La Tabla 9 muestra las variables de fecha y algunas estadísticas. Acá se comprueba que las fechas se encuentran en el momento de tiempo adecuado, entre enero de 2012 y enero de 2021. Adicionalmente, la variable "Conteo" muestra que no hay fechas para el total de los 9.611 proyectos para cada variable, por lo cual será necesario eliminar algunos registros.

Tabla 9. Variables de fecha

Variable	Conteo	No. únicos	Fecha más frecuente	Fecha más antigua	Fecha más reciente
FECHA INICIO PROGRAMACIÓN INICIAL	8683	119	2019-12-01 00:00:00	2012-01-01 00:00:00	2021-12-01 00:00:00
FECHA FINAL PROGRAMACIÓN INICIAL	8683	181	2019-12-31 00:00:00	2012-10-31 00:00:00	2029-12-31 00:00:00
FECHA INICIAL PROGRAMACIÓN ACTUAL	8680	110	2019-12-01 00:00:00	2012-09-01 00:00:00	2109-12-01 00:00:00
FECHA FINAL PROGRAMACIÓN ACTUAL	8680	174	2020-12-31 00:00:00	2012-11-30 00:00:00	2110-04-30 00:00:00
FECHA APROBACIÓN	9597	1648	2018-12-21 00:00:00	2012-02-19 00:00:00	2020-12-30 00:00:00
MAXIMO PERIODO APROBADO	9475	10	2020-12-01 00:00:00	2018-10-01 00:00:00	2020-12-01 00:00:00

Variables de texto

Las variables de texto fueron obtenidas de las bases de datos de Mapas de Inversión y se utilizaron para agrupar los proyectos con textos similares. Los textos se consiguieron para todos los 9.611 proyectos. Si bien su formato original es texto, se juntaron los textos de cada proyecto, se vectorizaron y se agruparon para crear variables categóricas que representen el grupo de cada proyecto según el texto escrito. Aquí se presentan las variables originales como se consiguieron en las bases de datos de Mapas de Inversión.

La Tabla 10 muestra unos ejemplos de cómo se ven los textos de las variables de texto de Mapas de Inversión.

Tabla 10. Variables de texto y ejemplos

BPIN	NombreProyecto	ProblemaCentral	Descripcion	ObjetivoGeneral
2.013.413.960.003	construcción pavimento en concreto rigido de la carr 1 entre cll 6 y 8 del municipio de la plata, huila, centro oriente	deficiente movilidad del transporte en la zona urbana	pavimentacin de 1.272,60 metros cuadrados de vias urbanas en concreto rigido de 3500 psi con un espesor de 15 cms sobre una base granular de 15 cms y acero de transferencia pdr 60. se hara el cajeo mecanico	pavimentar vias para mejorar la movilidad
2.019.681.470.001	mejoramiento de las vias rurales mendiante la construccion de estructuras en concreto reforzado tipo alcantarillas para el municipio de capitanejo, santander	dificultad en la intercomunicación terrestre de una parte de la población rural de la entidad territorial	construcción de 26 unidades de alcantarillas en concreto reforzado tipo invias	mejorar la intercomunicación terrestre de una parte de la población rural de la entidad territorial
2.012.501.500.010	construcción de vivienda de interes prioritario la pradera orinoquia, meta, castilla la nueva	deficit de vivienda de para personas en condiciones de vulnerabilidad	construir 270 soluciones de vivienda de interes prioritario en el casco urbano del municipio de castilla la nueva meta	disminuir el deficit de vivienda en el casco urbano del municipio
2.018.157.740.002	pavimentación en adoquin de la calle 7 desde carrera 3 a la salida del municipio de susacón- boyacá	bajos niveles de movilidad en el tránsito vehicular, en la zona urbana del municipio.	construcción de 102 metros lineales de adoquin vehicular de 8 cm, con una estructura de 30 cm de mejoramiento; adicional	mejorar el tránsito vehicular en la zona urbana del municipio.

BPIN	NombreProyecto	ProblemaCentral	Descripcion	ObjetivoGeneral
			a ello la construccion de adoquin peatonal confinado con bordillos prefabricados	

Variables dependientes

Las variables dependientes son aquellas que muestran si un proyecto tuvo cambios grandes en sus tiempos de ejecución originales o en sus valores monetarios o si tuvo un índice IGPR insuficiente. De esta manera se mide si un proyecto de inversión tuvo un problema o no en algún momento de su ejecución. La información de los tiempos de ejecución y los valores monetarios se consiguió de GESPROY y el índice IGPR de la base de datos de IGPR. Vale la pena decir que este índice solo está disponible para los últimos tres trimestres de 2020. El objetivo es elaborar varios modelos de predicción con cada una de estas variables para observar con cuál se obtienen los mejores resultados.

Es importante mencionar que hay que definir cuáles son los valores en tiempos y monetarios que representan si un proyecto tuvo dificultades en su ejecución. Es decir, no es lo mismo si un proyecto tuvo un retraso de un mes o de cinco años, o de un millón de pesos o de mil millones de pesos. Por ello, si bien se puede ir definiendo un umbral temporal, es necesario definir estos valores junto con la DVR.

Tabla 11. Variables independientes

Variable	Descripción
IGPR - original	La variable se obtiene del índice IGPR del último trimestre de 2020 y tiene los siguientes valores 1: el proyecto no tiene un IGPR satisfactorio 0: de lo contrario Es necesario definir cuál es el umbral
IGPR - mínimo	La variable se obtiene a partir del índice IGPR más bajo de los trimestres disponibles de 2020 y tiene los siguientes valores: 1: el proyecto no tiene un IGPR satisfactorio 0: de lo contrario Es necesario definir cuál es el umbral

Variable	Descripción
Cambios significativos en tiempos de ejecución	Se crea a partir de las variables 'FECHA FINAL PROGRAMACIÓN INICIAL' y 'FECHA FINAL PROGRAMACIÓN ACTUAL'. Tiene los siguientes valores: 1: el proyecto tuvo dificultades en su ejecución 0: de lo contrario Es necesario definir qué diferencia entre las fechas inicial y actual de finalización del proyecto se considera perjudicial para un proyecto
Cambios significativos en el valor monetario original de la inversión	Se crea a partir de las variables 'VALOR SGR' históricas de los conjuntos de datos de proyectos de GESPROY. Tiene los siguientes valores: 1: el proyecto tuvo dificultades en su ejecución 0: de lo contrario Es necesario definir qué diferencia entre el valor monetario original de un proyecto y el valor actual se considera perjudicial para un proyecto

Resultados con todos los años de la muestra

A continuación, se muestran 4 tablas que contienen los resultados para los modelos de predicción con distintas variables dependientes. Cada tabla contiene los resultados de 3 modelos, los cuales contienen variables dependientes similares que varían en los porcentajes de los cambios ya sea en tiempos de ejecución o valores monetarios. Estos resultados contienen todos los proyectos de inversión, pertenecientes a todos los años de la muestra, desde 2012 a 2020.

La *Tabla 12* contiene los resultados con las variables dependientes que miden si el proyecto de inversión tuvo demoras en tiempos de ejecución. Se entrenaron 3 modelos distintos, uno para medir los proyectos con demoras en más del 20% en los tiempos de ejecución, otro con demoras del 50% y tercero con 80%. En general, el *recall* es mayor para predecir proyectos con demoras más pequeñas. El *precision* se encuentra siempre por encima del 74% y es mayor a 90% para los casos de predicción más restrictivos, donde el umbral de éxito es mayor a 90%. *Accuracy* mide qué tan bien se predijeron los proyectos con y sin problemas en conjunto y este valor es siempre mayor a 70%.

Tabla 12. Aumentos significativos en tiempos de ejecución

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo retrasos de más del 20% en tiempos	0.5	0,785	0,816	0,794	1 (proyecto con dificultad): 4657 0 (proyecto sin dificultad): 4211
	0.6	0,745	0,843	0,793	
	0.7	0,702	0,869	0,788	
	0.8	0,647	0,895	0,775	
	0.9	0,550	0,930	0,742	

Proyecto tuvo retrasos de más del 50% en tiempos	0.5	0,719	0,770	0,783	1 (proyecto con dificultad): 3893 0 (proyecto sin dificultad): 4975
	0.6	0,667	0,796	0,779	
	0.7	0,612	0,832	0,775	
	0.8	0,533	0,865	0,759	
	0.9	0,414	0,917	0,727	
Proyecto tuvo retrasos de más del 80% en tiempos	0.5	0,652	0,745	0,794	1 (proyecto con dificultad): 3194 0 (proyecto sin dificultad): 5674
	0.6	0,592	0,787	0,795	
	0.7	0,529	0,828	0,790	
	0.8	0,454	0,869	0,778	
	0.9	0,343	0,922	0,753	

Fuente: elaboración propia

La Tabla 13 muestra los resultados para las variables dependientes que miden si los proyectos de inversión tuvieron incrementos significativos en sus valores monetarios originales provenientes de regalías. Los porcentajes en los incrementos son 20%, 50% y 80%. A diferencia de la Tabla 12, en este caso los datos son bastante desbalanceados, es decir, los proyectos con aumentos en su valor son mucho menores que los que no tuvieron estos aumentos, por lo cual el accuracy no es recomendado para medir el rendimiento del modelo. En general, el recall no suele ser más alto que 54%, pero el precision está siempre encima del 80% y en varios casos encima del 90%.

Tabla 13. Aumentos significativos en valores monetarios iniciales

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo incrementos de más del 20% en su valor monetario inicial	0.5	0,380	0,830	0,954	1 (proyecto con dificultad): 586 0 (proyecto sin dificultad): 8282
	0.6	0,358	0,867	0,954	
	0.7	0,334	0,899	0,954	
	0.8	0,296	0,937	0,952	
	0.9	0,190	0,959	0,946	
Proyecto tuvo incrementos de más del 50% en su valor monetario inicial	0.5	0,515	0,848	0,976	1 (proyecto con dificultad): 364 0 (proyecto sin dificultad): 8504
	0.6	0,497	0,873	0,976	
	0.7	0,474	0,905	0,976	
	0.8	0,396	0,919	0,974	
	0.9	0,262	0,942	0,969	
Proyecto tuvo incrementos de más del 80% en su valor monetario inicial	0.5	0,536	0,823	0,979	1 (proyecto con dificultad): 316 0 (proyecto sin dificultad): 8552
	0.6	0,536	0,853	0,980	
	0.7	0,504	0,873	0,979	
	0.8	0,420	0,910	0,978	
	0.9	0,296	0,941	0,974	

Fuente: elaboración propia

La Tabla 14 tiene los resultados para los modelos que predicen si los proyectos de inversión tuvieron cambios significativos en tiempos de ejecución y en valores monetarios. La muestra también es bastante desbalanceada (hay muchos menos proyectos con retrasos en tiempos e incrementos en valores monetarios), por lo cual no se recomienda interpretar los resultados con el *accuracy*. Estos resultados se destacan por tener un *precision* bastante alto, siempre por encima del 88%, lo cual indica que cuando se predice un proyecto con una dificultad, según este criterio, con una probabilidad muy alta será un proyecto problemático.

Tabla 14. Cambios significativos en tiempos de ejecución y valores monetarios

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo retrasos de más del 20% en tiempos e incrementos de más del 20% en su valor monetario inicial	0.5	0,444	0,881	0,972	1 (proyecto con dificultad): 401 0 (proyecto sin dificultad): 8467
	0.6	0,419	0,905	0,972	
	0.7	0,376	0,924	0,970	
	0.8	0,343	0,950	0,969	
	0.9	0,257	0,965	0,966	
Proyecto tuvo retrasos de más del 50% en tiempos e incrementos de más del 50% en su valor monetario inicial	0.5	0,611	0,950	0,989	1 (proyecto con dificultad): 236 0 (proyecto sin dificultad): 8632
	0.6	0,588	0,961	0,988	
	0.7	0,545	0,971	0,987	
	0.8	0,509	0,977	0,986	
	0.9	0,402	0,982	0,984	
Proyecto tuvo retrasos de más del 80% en tiempos e incrementos de más del 80% en su valor monetario inicial	0.5	0,626	0,888	0,988	1 (proyecto con dificultad): 228 0 (proyecto sin dificultad): 8640
	0.6	0,622	0,912	0,988	
	0.7	0,577	0,943	0,988	
	0.8	0,524	0,944	0,987	
	0.9	0,378	0,954	0,983	

Fuente: elaboración propia

La Tabla 15 contiene los resultados para los modelos que predicen si un proyecto de inversión puede tener demoras en tiempos de ejecución o cambios significativos en sus valores monetarios originales. Los resultados son similares a aquellos de la Tabla 12 y contienen una muestra relativamente balanceada de los datos de las variables dependientes.

Tabla 15. Cambios significativos en tiempos de ejecución o valores monetarios

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo retrasos de más del 20% en tiempos o incrementos de más del 20% en su valor monetario inicial	0.5	0,781	0,806	0,778	1 (proyecto con dificultad): 4842 0 (proyecto sin dificultad): 4026
	0.6	0,738	0,834	0,777	
	0.7	0,698	0,867	0,776	
	0.8	0,635	0,892	0,759	
	0.9	0,540	0,930	0,727	
Proyecto tuvo retrasos de más del 50% en tiempos o incrementos de más del 50% en su valor monetario inicial	0.5	0,717	0,767	0,773	1 (proyecto con dificultad): 4021 0 (proyecto sin dificultad): 4847
	0.6	0,662	0,796	0,770	
	0.7	0,608	0,830	0,766	
	0.8	0,532	0,866	0,751	
	0.9	0,407	0,917	0,715	
Proyecto tuvo retrasos de más del 80% en tiempos o incrementos de más del 80% en su valor monetario inicial	0.5	0,645	0,738	0,782	1 (proyecto con dificultad): 3316 0 (proyecto sin dificultad): 5552
	0.6	0,590	0,778	0,784	
	0.7	0,530	0,821	0,781	
	0.8	0,456	0,868	0,770	
	0.9	0,329	0,911	0,737	

Fuente: elaboración propia

Resultados sin los años 2019 ni 2020

Dado que solo es posible identificar las dificultades que puede tener un proyecto de inversión si este ha tenido una existencia relativamente prolongada en el tiempo, se decidió entrenar los mismos modelos de predicción que se muestran en la Sección 0 pero sin los proyectos de inversión que iniciaron en los años 2019 y 2020. Los resultados fueron mejores (ver *recall* y *precision*) y son coherentes con los obtenidos en la sección anterior, pero se pierden varias observaciones que podrían ser importantes para utilizar en predicciones de años posteriores a 2020. Estos resultados se encuentran en los anexos, al final del documento

5. Instrucciones de uso de herramienta de visualización de resultados

En esta sección se presenta cómo utilizar la herramienta desde un computador conectado a la intranet del DNP

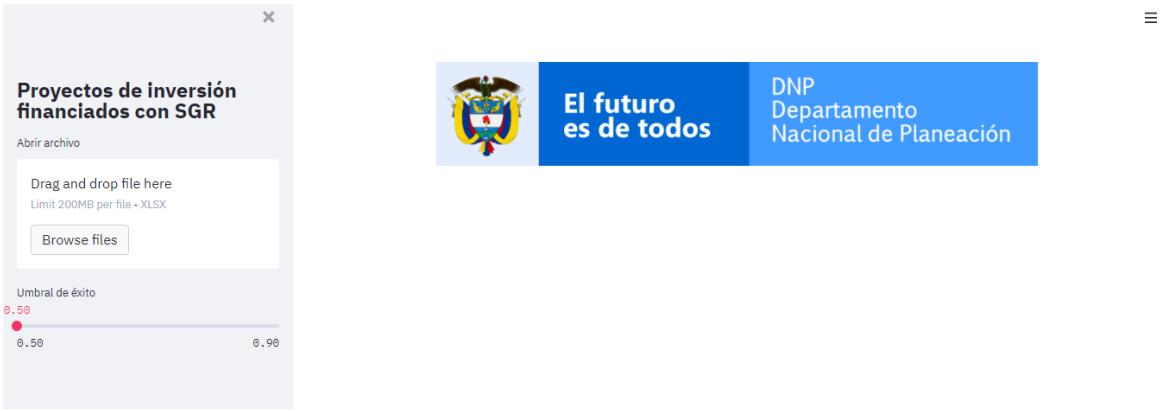
Abrir herramienta

El primer paso para utilizar la herramienta es escribir la siguiente ruta web en un navegador de un computador del DNP:

http://vdatascience:8076/

Esta ruta abrirá la aplicación, que se visualizará como en la *Ilustración 1*. Consiste en una barra lateral izquierda, donde se podrán cargar los documentos de Excel de insumo y escoger el umbral de éxito (lo cual se explicará más adelante).

Ilustración 1. Vista de la herramienta apenas abierta



Fuente: herramienta de proyectos de inversión

Archivo de Excel de insumo

El insumo de la herramienta es un archivo de Excel que contiene la información necesaria para hacer las predicciones de los proyectos de inversión. Toda la información se consigue de las bases de datos de GESPROY. El archivo debe tener una única hoja y las siguientes variables de los proyectos de inversión que quieren ser revisados (los nombres de las variables tienen que estar en mayúsculas y exactamente como se muestran a continuación):

Tabla 16. Columnas de insumo

Columna	Descripción
BPIN	Número BPIN del proyecto de inversión
NOMBRE DEL PROYECTO	Nombre del proyecto

Columna	Descripción
VALOR SGR	Valor numérico del proyecto con recursos provenientes de SGR
PGN	Valor total del Presupuesto General de la Nación
SGP	Valor total del Sistema General de Participaciones
VALOR NO SUIFP	Valor no incluido en SUIFP
TOTAL PROYECTO	Valor numérico total del proyecto
FECHA INICIO PROGRAMACIÓN INICIAL	Fecha de inicio del proyecto según la programación inicial. Tiene el formato fecha "DD/ MM/AAAA"
FECHA FINAL PROGRAMACIÓN INICIAL	Fecha de fin del proyecto según la programación inicial. Tiene el formato fecha "DD/ MM/AAAA"
FECHA INICIAL PROGRAMACIÓN ACTUAL	Fecha inicial del proyecto según la programación actual. Tiene el formato fecha "DD/ MM/AAAA"
FECHA APROBACIÓN	Fecha de aprobación del proyecto. Tiene el formato fecha "DD/ MM/AAAA"
REGIÓN	Región donde se ejecuta el proyecto
ENTIDAD EJECUTORA	Entidad que ejecuta el proyecto
SECTOR-SUIFP	Sector SUIFP al que pertenece el proyecto
ENFOQUE DIFERENCIAL	Especifica si el proyecto tuvo un enfoque sobre alguna comunidad

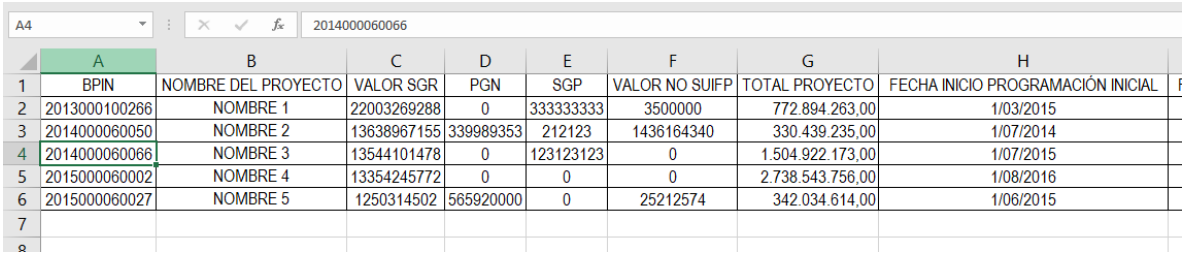
Atención

Para que la herramienta funcione correctamente, se tienen que ingresar las variables con los nombres que se muestran en la Tabla 16, teniendo en cuenta las mayúsculas, tildes y espacios. Así es como están escritas en las bases de GESPROY. Adicionalmente, los nombres de las etiquetas o valores en las variables con texto deben ser exactamente las mismas que se encuentran en las bases de GESPROY

En caso de que haya bases de datos que no contengan las columnas exactas como se muestran en la Tabla 16, será necesario calcularlas con el resto de variables. Si llega a suceder que una columna no existe en una nueva base de datos, se debe ingresar con valores vacíos si es de texto o con ceros si es numérica

La Ilustración 2 y la Ilustración 3 muestran el ejemplo de un mismo archivo de Excel con las columnas y sus respectivos valores a ser ingresados a la herramienta.

Ilustración 2. Archivo de Excel de insumo – 1



A4 | 2014000060066

	A	B	C	D	E	F	G	H
1	BPIN	NOMBRE DEL PROYECTO	VALOR SGR	PGN	SGP	VALOR NO SUIFP	TOTAL PROYECTO	FECHA INICIO PROGRAMACIÓN INICIAL
2	2013000100266	NOMBRE 1	22003269288	0	333333333	3500000	772.894.263,00	1/03/2015
3	2014000060050	NOMBRE 2	13638967155	339989353	212123	1436164340	330.439.235,00	1/07/2014
4	2014000060066	NOMBRE 3	13544101478	0	123123123	0	1.504.922.173,00	1/07/2015
5	2015000060002	NOMBRE 4	13354245772	0	0	0	2.738.543.756,00	1/08/2016
6	2015000060027	NOMBRE 5	1250314502	565920000	0	25212574	342.034.614,00	1/06/2015
7								

Fuente: elaboración propia

Ilustración 3. Archivo de Excel de insumo – 2 (continuación)

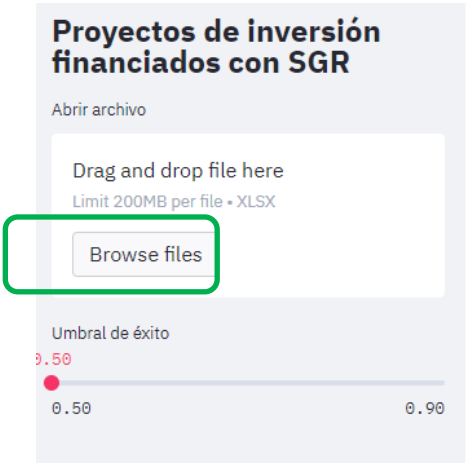
	I	J	K	L	M	N	O
1	FECHA FINAL PROGRAMACIÓN INICIAL	FECHA INICIAL PROGRAMACIÓN ACTUAL	FECHA APROBACIÓN	REGIÓN	ENTIDAD EJECUTORA	SECTOR-SUIFP	ENFOQUE DIFERENCIAL
2	31/03/2015	1/03/2014	27/02/2013	Centro Sur	Departamento De Amazonas	Transporte	Sin Enfoque Diferencial
3	30/11/2019	1/07/2014	27/02/2013	Centro Sur	Departamento De Amazonas	Relaciones Exteriores	Población Indígena
4	30/11/2016	1/10/2015	20/12/2012	Eje Cafetero	Corporación Ruta N Medellín	Salud Y Protección Social	Población Indígena
5	31/05/2017	1/03/2014	20/12/2012	Del Llano	Departamento De Arauca	Transporte	Sin Enfoque Diferencial
6	30/11/2016	1/11/2013	20/12/2012	Caribe	Luruaco	Deporte Y Recreación	Población Indígena
7							

Fuente: elaboración propia

Ingreso de archivo de Excel de insumo a la herramienta

Para ingresar el archivo de Excel, con las especificaciones de la Sección 0, se presiona sobre el botón “Browse files”, el cual se encuentra en la barra lateral izquierda, como se muestra en la *Ilustración 4*

Ilustración 4. Botón de carga de archivo de Excel



Fuente: herramienta de proyectos de inversión

Luego de presionar el botón, se podrá buscar el archivo de Excel en una carpeta del computador y abrirlo. Una vez abierto, la herramienta tendrá el siguiente aspecto:

Ilustración 5. Herramienta con Excel cargado

Fuente: herramienta de proyectos de inversión

Tabla de resultados

Una vez cargado el archivo de Excel, aparecerá debajo de la imagen con el logo DNP una tabla que mostrará si los proyectos de inversión serán problemáticos o no y también algunas columnas con información específica de cada proyecto. Las columnas de esa tabla son las siguientes:

Tabla 17. Nombre y descripción de las columnas de la herramienta

Columna	Descripción
BPIN	Número BPIN del proyecto de inversión
NOMBRE DEL PROYECTO	Nombre del proyecto de inversión
Predicción	Especifica si el proyecto es problemático o no, según el umbral de éxito escogido. Contiene los valores "P" (problemático) y "NP" (no problemático)
Probabilidad problemático	Muestra la probabilidad de que el proyecto sea problemático según el modelo de predicción.
Problema por fecha	Muestra si es posible establecer que el problema del proyecto es por ampliaciones en más del 20% de los tiempos de ejecución iniciales. En caso de ser así, aparecerá la palabra “Sí”
Problema por valor monetario	Muestra si es posible establecer que el problema del proyecto se da por incrementos en más del 20% de los valores monetarios iniciales. En caso de ser así, aparecerá la palabra “Sí”

Fuente: elaboración propia

La tabla de resultados se ve de la siguiente manera en la herramienta, con las primeras cuatro variables:

Ilustración 6. Tabla de resultados en la herramienta
(los valores son de ejemplo y no los verdaderos)

	BPIN	NOMBRE DEL PROYECTO	Predicción	Probabilidad problemát…	F
0	2013000100266	INVESTIGACIÓN INVENTAR…	P	0.9517	
2	2014000060066	CONSTRUCCIÓN DE UNA BA…	P	0.9106	
4	2015000060027	ESTUDIOS Y DISEÑOS PAR…	P	0.8953	
3	2015000060002	ESTUDIOS Y DISEÑOS PAR…	P	0.8851	
1	2014000060050	ESTUDIOS Y DISEÑOS PAR…	P	0.5787	

Fuente: herramienta proyectos de inversión

Con la barra deslizable en la parte inferior, es posible ver el resto de las variables que se encuentran a la derecha de la tabla. Estas indican si es posible establecer que el proyecto de inversión tenga problemas por aumentos del 20% en tiempos de ejecución o en valores monetarios. La Tabla 18 presenta estas dos últimas columnas.

Tabla 18. Tabla de resultados de la herramienta – últimas dos columnas.
(Los resultados son de ejemplo)

Problema por fecha	Problema por valor mon…
Sí	Sí
Sí	
Sí	Sí
Sí	Sí
	Sí

La tabla de la herramienta está ordenada de mayor a menor a partir de la variable “Probabilidad problemático”, donde la primera fila corresponde al proyecto de inversión con mayor probabilidad de ser problemático.

Seleccionar umbral de éxito

El umbral de éxito mide la probabilidad mínima que debe tener el resultado de un proyecto de inversión para ser catalogado como problemático según el modelo de aprendizaje de máquina. El umbral por defecto es 0.5, es decir, que si el proyecto es considerado problemático es porque el modelo concluye que tienen una probabilidad mayor a 0.5 de serlo.

Entre más alto sea este umbral de éxito, cada proyecto necesitará una probabilidad mayor para ser considerado problemático. Es decir, entre mayor sea el umbral de éxito más estricto es el modelo para concluir que un proyecto sea problemático. El umbral de éxito puede escogerse en la barra lateral izquierda con el botón deslizable que se encuentra debajo del título “Umbral de éxito”. Pueden escogerse los siguientes valores para el umbral de éxito: 0.5; 0.6; 0.7; 0.8; 0.9. Al cambiar el umbral de éxito se cambian las variables “Predicción”, “Problema por fecha” y “Problema por valor monetario”. La *Ilustración 7* muestra dónde se encuentra el umbral de éxito.

Ilustración 7. Umbral de éxito

Fuente: herramienta proyectos de inversión

La Ilustración 8 y la Ilustración 9 muestran cómo varía la tabla de resultados con un umbral de éxito de 0.5 y 0.9, respectivamente (los proyectos dentro de la tabla no corresponden a proyectos reales). En la Ilustración 8 se ve que todos los proyectos de inversión son considerados problemáticos según el umbral de 0.5.

Ilustración 8. Tabla con umbral de éxito de 0.5

Fuente: herramienta proyectos de inversión

La Ilustración 9 muestra que con un umbral de éxito de 0.9 únicamente dos proyectos de inversión son considerados problemáticos.

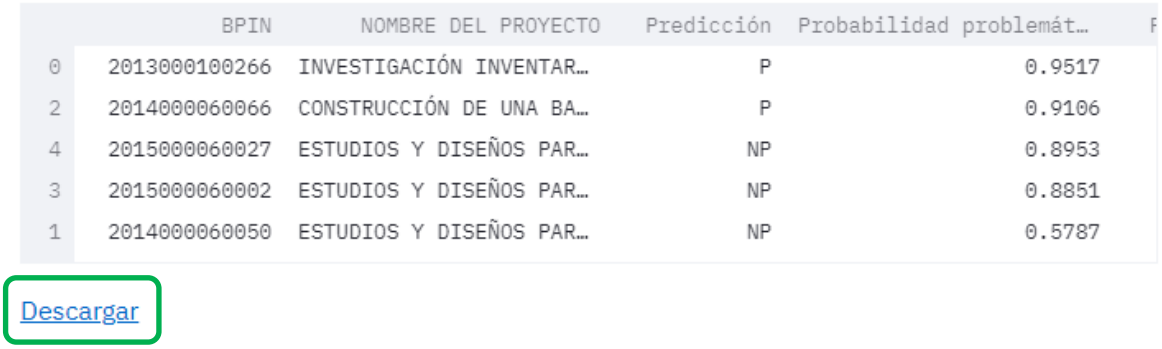
Ilustración 9.Tabla con umbral de éxito de 0.9

Fuente: herramienta proyectos de inversión

Descargar resultados

La descarga de la tabla de resultados, con el umbral de éxito escogido, se hace presionando sobre el botón "Descargar" que se encuentra justo debajo de la tabla, como lo muestra la Ilustración 10.

Ilustración 10. Botón de descarga de tabla con resultados



	BPIN	NOMBRE DEL PROYECTO	Predicción	Probabilidad problemát…	P
0	2013000100266	INVESTIGACIÓN INVENTAR…	P	0.9517	
2	2014000060066	CONSTRUCCIÓN DE UNA BA…	P	0.9106	
4	2015000060027	ESTUDIOS Y DISEÑOS PAR…	NP	0.8953	
3	2015000060002	ESTUDIOS Y DISEÑOS PAR…	NP	0.8851	
1	2014000060050	ESTUDIOS Y DISEÑOS PAR…	NP	0.5787	

Descargar

Fuente: herramienta proyectos de inversión

Al presionar sobre este botón la tabla se descargará directamente a la carpeta "Descargas" del computador en un archivo de Excel.

6. Conclusiones y recomendaciones

A partir de la metodología desarrollada y de los resultados obtenidos para cumplir los objetivos de este entregable, planteados en el plan de trabajo, se presentan a continuación las principales conclusiones obtenidas por el equipo de la UCD y las principales recomendaciones para la continuación del proyecto.

1. Los modelos para predecir el IGPR de los proyectos de inversión no dieron resultados satisfactorios, por lo cual no se recomienda aplicarlos en la práctica. Es posible utilizar la variable de IGPR si se cuenta con un

nivel histórico lo suficientemente amplio, de tal manera que se pueda observar una evolución de los proyectos de inversión

2. La variable dependiente utilizada en el modelo identifica si los proyectos de inversión tienen incrementos en más del 20% en los tiempos de ejecución iniciales o en los valores monetarios iniciales
3. La herramienta de visualización de resultados se puede acceder desde un computador del DNP conectado al internet de la entidad y con ella se puede predecir si nuevos proyectos de inversión tendrían dificultades o no

7. Socialización

Los resultados fueron compartidos con los miembros de la DVR y se subió la herramienta de visualización de resultados en los servidores de la UCD, para poder accederla en la intranet del DNP.

ANEXOS

Tabla 19. Aumentos significativos en tiempos de ejecución sin años 2019 y 2020

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo retrasos de más del 20% en tiempos	0.5	0,856	0,839	0,815	1 (proyecto con dificultad): 2543 0 (proyecto sin dificultad): 1695
	0.6	0,825	0,857	0,813	
	0.7	0,786	0,879	0,807	
	0.8	0,731	0,893	0,786	
	0.9	0,605	0,921	0,732	
Proyecto tuvo retrasos de más del 50% en tiempos	0.5	0,803	0,795	0,788	1 (proyecto con dificultad): 2222 0 (proyecto sin dificultad): 2016
	0.6	0,755	0,814	0,781	
	0.7	0,697	0,836	0,770	
	0.8	0,619	0,858	0,746	
	0.9	0,488	0,906	0,705	
Proyecto tuvo retrasos de más del 80% en tiempos	0.5	0,753	0,765	0,781	1 (proyecto con dificultad): 1944 0 (proyecto sin dificultad): 2294
	0.6	0,698	0,787	0,775	
	0.7	0,628	0,822	0,767	
	0.8	0,542	0,866	0,751	
	0.9	0,422	0,920	0,718	

Fuente: elaboración propia

Tabla 20. Aumentos significativos en valores monetarios iniciales sin los años 2019 y 2020

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo incrementos de más del 20% en su valor monetario inicial	0.5	0,414	0,829	0,913	1 (proyecto con dificultad): 548 0 (proyecto sin dificultad): 3690
	0.6	0,399	0,851	0,913	
	0.7	0,361	0,871	0,910	
	0.8	0,308	0,907	0,906	
	0.9	0,216	0,958	0,897	
Proyecto tuvo incrementos de más del 50% en su valor monetario inicial	0.5	0,546	0,847	0,955	1 (proyecto con dificultad): 347 0 (proyecto sin dificultad): 3891
	0.6	0,514	0,864	0,954	
	0.7	0,465	0,886	0,951	
	0.8	0,392	0,899	0,946	
	0.9	0,288	0,962	0,941	
Proyecto tuvo incrementos de más del 80% en su valor monetario inicial	0.5	0,536	0,790	0,956	1 (proyecto con dificultad): 305 0 (proyecto sin dificultad): 3933
	0.6	0,526	0,818	0,957	
	0.7	0,475	0,835	0,955	
	0.8	0,411	0,884	0,954	
	0.9	0,321	0,938	0,950	

Fuente: elaboración propia

Tabla 21. Cambios significativos en tiempos de ejecución y valores monetarios sin los años 2019 y 2020

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo retrasos de más del 20% en tiempos e incrementos de más del 20% en su valor monetario inicial	0.5	0,475	0,888	0,947	1 (proyecto con dificultad): 383 0 (proyecto sin dificultad): 3855
	0.6	0,454	0,915	0,947	
	0.7	0,420	0,930	0,945	
	0.8	0,354	0,941	0,940	
	0.9	0,276	0,973	0,934	
Proyecto tuvo retrasos de más del 50% en tiempos e incrementos de más del 50% en su valor monetario inicial	0.5	0,620	0,909	0,976	1 (proyecto con dificultad): 231 0 (proyecto sin dificultad): 4007
	0.6	0,594	0,936	0,976	
	0.7	0,520	0,954	0,973	
	0.8	0,472	0,971	0,971	
	0.9	0,413	0,988	0,968	
Proyecto tuvo retrasos de más del 80% en tiempos e incrementos de más del 80% en su valor monetario inicial	0.5	0,652	0,895	0,981	1 (proyecto con dificultad): 191 0 (proyecto sin dificultad): 4047
	0.6	0,630	0,918	0,981	
	0.7	0,572	0,939	0,979	
	0.8	0,547	0,936	0,978	
	0.9	0,440	0,965	0,974	

Fuente: elaboración propia

Tabla 22. Cambios significativos en tiempos de ejecución o valores monetarios sin los años 2019 y 2020

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
Proyecto tuvo retrasos de más del 20% en tiempos o incrementos de más del 20% en su valor monetario inicial	0.5	0,846	0,822	0,784	1 (proyecto con dificultad): 2708 0 (proyecto sin dificultad): 1530
	0.6	0,810	0,846	0,784	
	0.7	0,772	0,868	0,779	
	0.8	0,713	0,890	0,760	
	0.9	0,591	0,916	0,704	
Proyecto tuvo retrasos de más del 50% en tiempos o incrementos de más del 50% en su valor monetario inicial	0.5	0,799	0,784	0,768	1 (proyecto con dificultad): 2338 0 (proyecto sin dificultad): 1900
	0.6	0,753	0,807	0,764	
	0.7	0,694	0,831	0,753	
	0.8	0,613	0,858	0,731	
	0.9	0,480	0,903	0,685	
Proyecto tuvo retrasos de más del 80% en tiempos o incrementos de más del 80% en su valor monetario inicial	0.5	0,748	0,755	0,760	1 (proyecto con dificultad): 2058 0 (proyecto sin dificultad): 2180
	0.6	0,695	0,779	0,756	
	0.7	0,620	0,809	0,744	
	0.8	0,539	0,851	0,730	
	0.9	0,413	0,913	0,696	

Fuente: elaboración propia

Tabla 23. IGPR menor a umbral definido

Variable dependiente	Umbral de éxito	Recall	Precision	Accuracy	valores
	0.5	0,652	0,641	0,654	

Proyecto tuvo IGPR menor a 0.32	0.6	0,542	0,667	0,647	1 (proyecto con dificultad): 4296 0 (proyecto sin dificultad): 4572
	0.7	0,416	0,693	0,628	
	0.8	0,274	0,725	0,598	
	0.9	0,124	0,821	0,562	
Proyecto tuvo IGPR menor a 0.62	0.5	0,813	0,693	0,666	1 (proyecto con dificultad): 5407 0 (proyecto sin dificultad): 3461
	0.6	0,726	0,712	0,654	
	0.7	0,601	0,737	0,626	
	0.8	0,437	0,757	0,571	
	0.9	0,223	0,813	0,495	

Fuente: elaboración propia