Dirección de Desarrollo Digital

Unidad de Científicos de Datos

DEPARTAMENTO NACIONAL DE PLANEACIÓN

DEPARTAMENTO NACIONAL DE PLANEACIÓN

ANALÍTICA PREDICTIVA DE PROYECTOS DE INVERSIÓN FINANCIADOS CON RECURSOS SGR – ETAPA 2

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Economía Naranja y Desarrollo Digital - Unidad de Científicos de Datos • Dirección de Seguimiento, Evaluación y Control del Sistema General de Regalías
Sector	Territorial
Tecnologías utilizadas	Python
Fuentes de datos	• GESPROY • IGPR

Contenido

1. Presentación

Este proyecto es la continuación de un proyecto desarrollado en 2021, que tiene como objetivo predecir qué proyectos pueden presentar problemas de ejecución en el futuro dado un conjunto de variables. En esta etapa, se siguieron directrices proporcionadas por la Dirección de Seguimiento, Evaluación y Control del Sistema General de Regalías (DSECSGR) para mejorar la métrica del modelo, se eliminaron algunas variables y se redefinieron otras. Además, se construyeron diferentes modelos para incluir nuevas variables independientes. Después de varias discusiones con la DSECSGR, se determinó que la mejor opción sería utilizar las variables del modelo de 2021 e implementar las nuevas definiciones de otras variables para generar el modelo final de esta segunda etapa.

This marks the second stage of a project that was initiated in 2021 with the objective of developing a model that can predict potential difficulties that might arise during the execution of a project with royalty funds. In this second stage, we adhered to the guidelines provided by the Dirección de Seguimiento, Evaluación y Control del Sistema General de Regalías (DSECSGR) and made changes to some variables and redefined others to improve the model's metric. Moreover, we created various models to include new independent variables. After several discussions with the DSECSGR, it was concluded that the best course of action would be to use the variables from the 2021 model and implement the new definitions of other variables to create the final model for this second stage.

2. Objetivos del proyecto

2.1. General

Desarrollar un modelo de predicción de proyectos financiados con fondos de regalías que podrían presentar dificultades en el futuro por tiempo de ejecución o por costo programado o costo ejecutado.

2.2. Específicos

1. Realizar un análisis de la importancia de las variables y discutir los resultados con la DSECSGR para determinar su inclusión en el modelo final. Además, redefinir las variables propuestas por la DSECSGR para la construcción de variables dependientes, en caso de ser necesario.
2. Implementar modelos de balanceo de datos para evitar sesgos en el modelo final y garantizar su precisión.
3. Evaluar diversos modelos de aprendizaje automático para los problemas de clasificación de las variables dependientes, incluyendo cambios significativos en los tiempos de ejecución iniciales, cambios significativos en el valor monetario original de la inversión y cambios significativos en el valor monetario programado.
4. Desarrollar una herramienta de visualización que permita utilizar el modelo implementado para identificar proyectos que puedan presentar dificultades en el futuro de manera sencilla y eficiente.

3. Metodología

3.1. Adecuación de las bases de datos

En este punto se analizan las bases de datos dadas: GESPROY e IGPR. Se tuvieron en cuenta datos hasta el segundo trimestre del 2022. Además, se tuvo en cuenta los elementos en común por medio de la llave BPIN y se realizó una limpieza en cada una de las bases de datos que implica tener en cuenta valores nulos, observar si hay filas duplicadas, analizar si hay diferentes tipos de datos en cada variable, estudiar los valores atípicos. Una vez se realiza esta limpieza de datos se filtran los datos con respecto a la variable ESTADO DETALLE seleccionando únicamente a los proyectos que están en estado de ejecución. Puesto que se busca estudiar los proyectos que están en ejecución.

3.2. Selección de variables

Para la selección de variables se usó la metodología RFECV (Recursive Features Elimination Cross Validation) para la selección de variables. Se desarrollaron modelos con la selección de variables hecha por esta metodología y se sometió a discusión con la DSECSGR. De esta discusión se llegó a la conclusión que la selección de variables adecuada corresponde a la hecha en el modelo del 2021 con las siguientes adecuaciones:

- Eliminación de la variable Enfoque Diferencial
- Redefinición de las siguientes variables de esta manera:
- *"**longitud inicial del proyecto**": Diferencia entre las variables FECHA FINAL PROGRAMACIÓN ACTUAL y FECHA INICIO PROGRAMACIÓN ACTUAL*
- *"**retraso inicio proyecto**": Diferencia entre las variables MINIMA FECHA ACTA INICIO y FECHA INICIO PROGRAMACIÓN INICIAL*
- *"**demora en el inicio del proyecto**": Diferencia entre las variables MINIMA FECHA ACTA INICIO y MINIMA FECHA DE SUSCRIPCIÓN CONTRATOS*

3.3. Adecuación de las variables independientes y dependientes

Como se mencionó en la Subsección 3.2 la selección de variables se realizó de acuerdo con la selección de variables efectuada en el 2021 con las modificaciones mencionadas en la subsección 3.2. Es decir, que la selección final de variables independientes es mostrada en la .

Tabla 1: Variables independientes tomadas de las bases de datos del IGPR y GESPROY.

Columna	Descripción
BPIN	Número BPIN del proyecto de inversión
NOMBRE DEL PROYECTO	Nombre del proyecto
VALOR SGR	Valor numérico del proyecto con recursos provenientes de SGR
PGN	Valor total del Presupuesto General de la Nación
VALOR 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 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 FINAL PROGRAMACIÓN ACTUAL	Fecha final del proyecto según la programación actual. Tiene el formato fecha "DD/MM/AAAA"
REGION EJECUTOR	Región donde se ejecuta el proyecto
ENTIDAD EJECUTORA	Entidad que ejecuta el proyecto
SECTOR SUIFP	Sector SUIFP al que pertenece el proyecto
MINIMA FECHA ACTA DE INICIO	Fecha en la que se firmó el primer contrato del proyecto.

Fuente: Elaboración propia

Ahora, la selección de variables dependientes es la siguiente:

Tabla 2: Variables dependientes creadas a partir de variables tomadas de la bases de datos IGPR y GESPROY.

Variables dependientes	Descripción
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.
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.
Cambios significativos en el valor monetario programado	Muestra si el proyecto tuvo cambios en el valor programado.

Fuente: Elaboración propia.

3.4. Modelos de predicción

Se usaron dos modelos: Un modelo base de clasificación: RandomForest y a partir de este modelo se usó el XGBoost que mejoró el rendimiento de ejecución. Dentro de estos dos modelos se usaron distintas metodologías para ajustar el modelo. Es decir, en este caso los datos están desbalanceados (muchos proyectos sin dificultades a futuro y pocos proyectos con dificultades a futuro). Al ser datos desbalanceados los modelos tienden a ser sesgados (realizan mal las predicciones). Para balancearlos se usaron distintas metodologías. Una es conocida como SMOTE y consiste en crear datos artificiales en la clase con menos datos (en este caso los proyectos con dificultades a futuro). La otra metodología es conocida como Undersampling y consiste en borrar datos en la clase que tiene mayor información (en este caso en los proyectos sin dificultades a futuro). SMOTE y Undersampling se usaron tanto en el modelo RandomForest como en el XGBoost. Además de estas metodologías se usó calibración de la probabilidad para hallar el Umbral de éxito óptimo para mejorar las métricas del modelo.

3.5. Interpretación de resultados

La interpretación de los resultados se hizo con respecto a la métrica *f1-score.* No se incluye la métrica de accuracy debido a que no arroja buenos resultados cuando los datos son desbalanceados. Estas métricas están basadas en el concepto de matriz de confusión mostrado en la Tabla 3.

Tabla 3 : Matriz de confusión.

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

Fuente: Elaboración propia.

Las clasificaciones **Negativo** y **Positivo** en la Tabla 3 son convenciones usadas cuando el problema de clasificación tiene únicamente dos clases, dónde la clase **Positivo** representa la clase minoritaria y la clase **Negativo** representa la clase mayoritaria. En el problema de predicción de proyectos con dificultades a futuro la clase **Positivo** sería la clase de los proyectos que presentarían dificultades a futuro y la clase **Negativo** sería la clase de los proyectos que no presentan dificultad a futuro.

Ahora, los valores *Verdaderos negativos (vn)* representan las predicciones correctas de la clase **Negativo**. Los valores *Falsos positivos (fp)* representan las predicciones incorrectas asociadas a la clase **Positivo** cuando su valor real es la clase **Negativo**. Los valores *Falsos negativos (fn)* representan las predicciones incorrectas asociadas a la clase **Negativo** cuando su valor real es la clase **Positivo**. Los valores *Verdaderos positivos (vp)* representan las predicciones correctas de la clase **Positivo.**

La matriz de confusión expresada en términos del problema de clasificación de proyectos con dificultades a futuro tendría la forma expresada en la Tabla 4.

Tabla 4: Matriz de confusión para el caso de predicción de proyectos con dificultades a futuro.

		Categoría predicha	
		Sin dificultad	**Con dificultad**
Categoría Original	**Sin dificultad**	Predicción correcta: proyecto no tiene dificultad	Predicción incorrecta: proyecto tiene dificultad

	Con dificultad	Predicción incorrecta: proyecto no tiene dificultad	Predicción correcta: proyecto tiene dificultad

Fuente: Elaboración propia.

Dado que las métricas usadas en el modelo están basadas en la matriz de confusión entonces se mostrará las definiciones de la métrica usada expresada en la **¡Error! No se encuentra el origen de la referencia.**.

La métrica *f1-score* representa un balance entre *calidad* del modelo y la *cantidad* de predicciones correctas del modelo. De ahí que la métrica *f1-score* sea la métrica escogida para determinar si un proyecto presenta dificultades a futuro o no. Otra razón para escoger la métrica *f1-score* es que es la adecuada para datos desbalanceados como en este caso de predicción de proyectos que van a presentar dificultad a futuro.

Así, la métrica *f1-score* definida en términos de las clasificaciones: *Verdaderos negativos, Falsos positivos, Falsos negativos, Verdaderos positivos* es la mostrada en la formula Ecuación 1:

Ecuación 1: Ecuación del f1-score

$$f1 - score = 2\,x \left(\frac{\frac{vp}{vp + fn}\ x\ \frac{vp}{vp + fp}}{\frac{vp}{vp + fn} + \frac{vp}{vp + fp}} \right)$$

Fuente: Elaboración propia

3.6. Umbral de éxito

El umbral de éxito representa la probabilidad con la que se predice si un proyecto va a presentar dificultades en el futuro o no. De ahí que el umbral de éxito esté relacionado directamente a la métrica *f1-score* ya que de acuerdo al valor de la métrica *f1-score* y al umbral de éxito escogido se determina si un proyecto va a presentar dificultades a futuro o no.
Para el modelo del 2022 se dejó un umbral óptimo de 0.8 dado que si dicho umbral es menor entonces puede efectuar predicciones que no sean fiables.

3.7. Variable dependiente del IGPR

De la discusión con la *DSECSGR* se determinó que las variables finales serían las del modelo del 2021. Sin embargo, al contar con estas variables únicamente no se lograron buenos resultados con respecto a la variable del IGPR. De ahí, que no se muestran los resultados de esta variable en este informe.

4. Resultados

En esta sección se presentan los resultados hallados a lo largo del proyecto. Las variables escogidas son las usadas en el modelo del año 2021. Se escogieron de esta manera dado que los modelos para escoger nuevas variables no arrojaron buenos resultados para la DSECSGR. Así las variables independientes son las mostradas en la . Las variables dependientes son las mostradas en la Tabla 2.

Se incluyó la variable dependiente *Cambios significativos en el valor monetario programado* dado que podría estar relacionado a la variable *TIENE AJUSTES*. La variable TIENE AJUSTES determina si un proyecto ha sufrido ajustes en el valor monetario programado. Es decir que si un proyecto ha modificado su valor monetario programado es posible que pueda presentar dificultades en el futuro.

4.1. Resultados con respecto a cada una de las variables dependientes

Los resultados obtenidos se realizaron con respecto a la métrica *f1-score* para cada una de las variables dependientes como lo muestra la Tabla 5. Los resultados de la Tabla 5 corresponden al modelo XGBoost desarrollado con las

variables especificadas en la . Es necesario tener en cuenta que el *Umbral de éxito* está fijo y es 0.8. Por esta razón no se presenta un rango de valores como en el modelo 2021.

Tabla 5: valores f1-score por cada variable dependiente.

Variables dependiente	Valor f1-score
Cambios significativos en tiempos de ejecución iniciales	0.94
Cambios significativos en el valor monetario original de la inversión	0.84
Cambios significativos en el valor monetario programado	0.85

Fuente: Elaboración propia.

4.2. Selección del Umbral de éxito

El umbral de éxito se escogió teniendo en cuenta el mejor valor de la métrica *f1-score* para cada una de las variables dependientes. También este valor se escogió después de la discusión con la DSECSGR ya que comentaron que siempre escogían proyectos con un umbral de éxito mayor a 0.9.

5. Herramienta de visualización de resultados

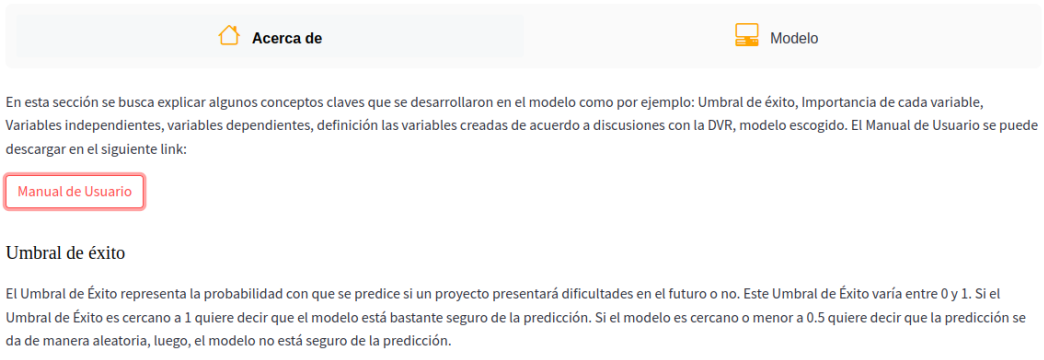
En esta sección se presenta la herramienta de visualización y que es posible usarla desde un computador conectado a la intranet del DNP.

5.1. Interfaz de la aplicación

La siguiente ruta abrirá la aplicación, que se visualizará como en la Figura 1. Esta aplicación consta de dos partes. La primera parte es *Acerca de* la cual muestra conceptos básicos usados en el desarrollo del modelo. Además, ofrece un botón para descargar el *Manual de Usuario.*

Aquí va la ruta de despliegue.

Figura 1: Interfaz de la aplicación al ingresar .



Fuente: Elaboración propia.

La segunda parte es *Modelo* que consta del modelo desarrollado junto a unas estadísticas de las predicciones desarrolladas. Aquí se encuentra un botón *Browse files* que permite cargar el archivo de Excel con los datos de los proyectos a predecir si van a presentar dificultades en el futuro.

6. Conclusiones y Recomendaciones

1. Se puede afirmar que el Índice General de Proyectos de Regalías (IGPR) es un indicador útil para detectar proyectos que pueden presentar problemas de ejecución en el futuro. Sin embargo, en este estudio se

encontró que el IGPR actual no fue útil para el modelo propuesto debido a la falta de información histórica para más años y la falta de estandarización del cálculo. Se recomienda la exploración de variables adicionales que permitan estandarizar el cálculo del IGPR y su aplicación a más años de información histórica para hacerlo más útil para el modelo propuesto y en futuros estudios relacionados con proyectos de regalías.

2. Para el cálculo de las variables dependientes de la Tabla 2, se redefinieron las variables: *longitud inicial del proyecto; retraso inicio proyecto;* y *demora en el inicio del proyecto,* siguiendo los lineamientos de la DSECSGR, como se mencionó en la Sección 3.2. Estas nuevas definiciones mejoraron la precisión de la estimación para determinar si un proyecto enfrentará dificultades de ejecución en el futuro.
3. Durante el estudio de relevancia y selección de características para mejorar la precisión del modelo de predicción, se identificaron variables que podrían mejorar su rendimiento. Sin embargo, después de discutir con expertos temáticos de la DSECSGR, se decidió no incluirlas en el modelo final. A pesar de esto, se recomienda analizar estas variables para futuras etapas del proyecto, ya que podrían ser de gran utilidad.
4. Se aplicaron técnicas de balanceo de datos (*SMOTE* y *undersampling*) en cada variable dependiente para contrarrestar el desequilibrio de clases en la base de datos. Los resultados demostraron que la utilización de estas técnicas permitió mejorar la precisión en la clasificación, la generalización del modelo y evitó el sesgo hacia los datos de entrenamiento. El método *SMOTE* permitió generar nuevas instancias sintéticas de la clase minoritaria, mientras que el *undersampling* disminuyó el número de instancias de la clase mayoritaria.
5. En este trabajo se evaluaron diversos modelos de aprendizaje automático para predecir si un proyecto con fondos de regalías y en estado de ejecución presentaría dificultades en el futuro. Tras comparar el desempeño de varios modelos, se seleccionó *XGBoost* como el modelo final debido a que obtuvo el mejor puntaje en la métrica f1-score y logró mejorar significativamente los tiempos de ejecución. Con una probabilidad superior al 84% (Tabla 5), el modelo final ofrece un alto nivel de precisión en la predicción de problemas futuros en los proyectos. En conjunto con la implementación de técnicas de balanceo de datos, se logró aumentar la generalización del modelo y evitar el sesgo en los datos de entrenamiento. Los resultados obtenidos demuestran la eficacia del enfoque propuesto y podrían ser de gran utilidad para la toma de decisiones en la gestión de proyectos financiados por regalías. Además, se desarrolló una herramienta de visualización que utiliza el modelo implementado para evaluar nuevos proyectos de regalías y estimar si tendrán problemas de ejecución en el futuro de una manera más sencilla.

7. Socialización

Los resultados serán compartidos por medio de la herramienta de visualización, informe final y manual de usuario con la DSECSGR y se subirá la herramienta de visualización de resultados en los servidores de la UCD, para poder accederla en la intranet del DNP.

Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.