



**MODELO DE PREDICCIÓN DE ALARMAS DE FALLA DE ROCAS EN OBRAS
SUBTERRÁNEAS A PARTIR DE DATOS DE INSTRUMENTACIÓN GEOTÉCNICA.**

Norbey Arcila Quintero

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor

José David Ruiz Álvarez, Doctor (PhD) en Física

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2021

Cita	(Arcila-Quintero, 2021)
Referencia	Arcila-Quintero, N (2021). <i>Modelo De Predicción de Alarmas de Falla de Rocas en Obras Subterráneas a Partir de Datos de Instrumentación Geotécnica</i> . [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.
Estilo APA 7 (2020)	



Especialización en Analítica y Ciencia de Datos, Cohorte II.



Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Jesús Francisco Vargas Bonilla.

Jefe departamento: Diego José Luis Botia Valderrama.

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Dedicatoria

Dedicado a Luisa, mi esposa.

Quien siempre me anima y me apoya

Infinito Amor y Agradecimiento hacia ella.

Tabla de contenido

Tabla de contenido	4
Lista de tablas.....	5
Lista de figuras	6
Resumen	8
Abstract	10
Introducción	12
1 Planteamiento del problema.....	2
1.1 Problema de negocio.	2
1.2 Aproximación desde la analítica de datos.	2
1.3 Origen de los datos.	3
1.4 Métricas de desempeño.	3
2 Datos	4
2.1 Datos originales.....	4
2.2 Dataset.	4
3 Proceso de Analítica.	6
4 Metodología.	9
4.1 Baseline.	9
4.2 Iteraciones y Evolución.	16
5 Resultados	18
5.1 Métricas.	18
5.2 Evaluación Cualitativa.	20
6 Conclusiones	21
7 Referencias.....	22

Lista de tablas

Tabla 1. Valores de los parámetros de la Ecuación 1 para la generación de los datos sintéticos para cada nivel de alerta.	4
--	---

Lista de figuras

Ilustración 1.	A la izquierda, serie temporal de desplazamiento de un geomaterial. El rectángulo amarillo representa la ventana de tiempo ampliada hacia la derecha, en la que se traza la velocidad inversa y la línea punteada roja representa la tendencia lineal. Modificado de Emanuele Intrieri, 2019.....	8
Illustration 2.	On the left, a geomaterial displacement time series. The yellow rectangle represents the time window enlarged to the right, in which the inverse velocity is plotted, and the red dotted line represents the linear trend. Modified from Emanuele Intrieri, 2019.....	10
Ilustración 3.	Representación de un registro de deformación de un extensómetro previo al colapso dentro de una excavación subterránea.	5
Ilustración 4.	Series de tiempo sintéticas dentro de sus respectivos Niveles de Alerta y sus umbrales.	3
Ilustración 5.	Inverso de la velocidad y su pronóstico de falla con relación a los niveles alerta.	3
Ilustración 6.	A la Izquierda se presenta los datos sintéticos generados para una serie de tiempo de 48 horas, se resalta el periodo entre 12 y 24 horas que es cuando se presenta la mayor tasa de cambio en la deformación, detallándose a la derecha.	5
Ilustración 7.	Visualización del Dataset de datos sintéticos.....	5
Ilustración 8.	Diagrama deflujo del proceso de analítica	8
Ilustración 9.	Curvas de deformación con variaciones en la selección del periodo de tiempo a emplear, detalles de la generación de estas curvas en el repositorio Evaluación_tiempo. ...	10
Ilustración 10.	Métricas de resultados y matriz de confusión Kernel Lineal.....	11
Ilustración 11.	Métricas de resultados y matriz de confusión Kernel Polinómico de grado 2..	12
Ilustración 12.	Métricas de resultados y matriz de confusión Kernel Polinómico de grado 3..	12
Ilustración 13.	Métricas de resultados y matriz de confusión Kernel Polinómico de grado 4..	13
Ilustración 14.	Métricas de resultados y matriz de confusión Kernel Polinómico de grado 5..	14
Ilustración 15.	Métricas de resultados y matriz de confusión Kernel Polinómico de grado 6..	14
Ilustración 16.	Métricas de resultados y matriz de confusión Kernel Sigmoide.....	15
Ilustración 17.	Métricas de resultados y matriz de confusión Kernel RBF	15

Ilustración 18. Comportamiento generalizado de la deformación en el instrumento Exten_E3_P2. Se resalta en amarillo los episodios de mayor aceleración en la deformación. 18

Ilustración 19. Métricas de resultados y matriz de confusión data de validación completa. 19

Ilustración 20. Métricas de resultados y matriz de confusión data de validación balanceada. . 19

Resumen

Se plantea un modelo para predicción de niveles de riesgos por fallas de rocas en obras subterráneas para la generación temprana de alertas, el desarrollo del proyecto tiene como base la metodología semi-empírica propuesta por Fukuzono (1985), quien identificó una relación entre el tiempo de aceleración de una falla de un material geológico (geomaterial) era inversamente proporcional a la tasa de cambio de la deformación, de lo cual se deriva que el tiempo de falla del geomaterial se puede pronosticar extrapolando la tendencia lineal hacia cero de la gráfica del inverso de la velocidad de deformación contra el tiempo (Fukuzono, 1985), ver Ilustración 1.

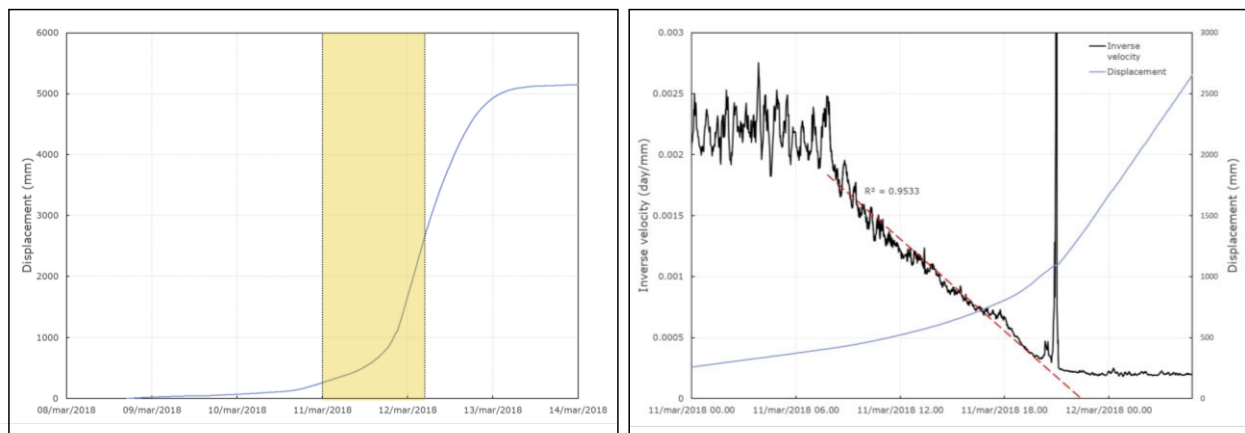


Ilustración 1. A la izquierda, serie temporal de desplazamiento de un geomaterial. El rectángulo amarillo representa la ventana de tiempo ampliada hacia la derecha, en la que se traza la velocidad inversa y la línea punteada roja representa la tendencia lineal. Modificado de Emanuele Intrieri, 2019.

El modelo utiliza el algoritmo de aprendizaje supervisado de Máquinas de Soporte Vectorial para clasificar una serie temporal de datos de deformación de un geomaterial registrados por instrumentos instalados para dicho fin. Dado lo anterior, el volumen de datos para entrenamiento a partir de estos registros es limitado por el número de instrumentos, la cantidad de tiempo registrando lecturas y la cifra y variedad de alarmas reportadas. Por tanto, debió recurrirse a la generación de datos sintéticos que cubriesen el volumen de información requerida para el entrenamiento del modelo, mediante el uso de una ecuación de la curva logística, en su tramo con mayor tasa de cambio.

A partir de los datos sintéticos se generan las primeras iteraciones para la puesta en operación del modelo, con cada iteración se definieron distintos parámetros de la implementación de scikit-learn [sklearn.svm](https://scikit-learn.org/stable/modules/svm.html), buscando los que entregasen un mejor desempeño del modelo. Finalmente, se obtuvieron mejores resultados implementando un modelo de clasificación con un Kernel polinómico de grado 2.

<https://github.com/Narcilaq/Monografia>

Palabras clave: Instrumentación Geotécnica, Niveles de Alerta, Máquinas de Vectores de Soporte, Datos Sintéticos.

Abstract

A model is proposed for predicting risk levels due to rock failures in underground works for the early generation of alerts, the development of the project is based on the semi-empirical methodology proposed by Fukuzono (1985), who identified a relationship between time of acceleration of a failure of a geological material (geomaterial) was inversely proportional to the rate of change of the deformation, from which it follows that the failure time of the geomaterial can be predicted by extrapolating the linear trend towards zero from the graph of the inverse of the strain rate against time (Fukuzono, 1985), see Ilustración 1.

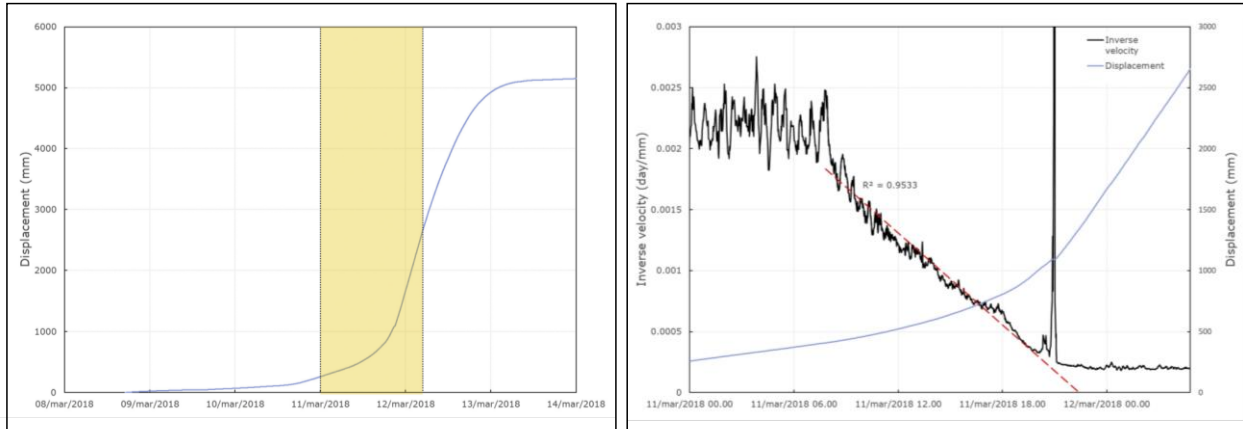


Illustration 2. On the left, a geomaterial displacement time series. The yellow rectangle represents the time window enlarged to the right, in which the inverse velocity is plotted, and the red dotted line represents the linear trend. Modified from Emanuele Intrieri, 2019.

The model uses the supervised learning algorithm of Vector Support Machines to classify a time series of deformation data of a geomaterial recorded by instruments installed for this purpose. Given the above, the volume of data for training from these records is limited by the number of instruments, the amount of time recording readings, and the number and variety of reported alarms. Therefore, it was necessary to resort to the generation of synthetic data that covered the volume of information required for the training of the model, using an equation of the logistic curve, in its section with the highest rate of change.

From the synthetic data, the first iterations are generated to put the model into operation. With each iteration, different parameters of the scikit-learn [sklearn.svm](https://scikit-learn.org/stable/), implementation were defined,

looking for those that delivered a better performance of the model. Finally, better results were obtained by implementing a classification model with a polynomial kernel of degree 2.

<https://github.com/Narcilaq/Monografia>

Keywords: Geotechnical Instrumentation, Alert Levels, Support Vector Machine, Synthetic Data

Introducción

La construcción de obras civiles son espacios donde confluyen distintos tipos de riesgos bajo diversos niveles de alertas, especialmente si son de tipo subterráneo. Este documento plantea un modelo para predicción de niveles de riesgos por fallas de rocas en obras subterráneas para la generación temprana de alertas utilizando el modelo teórico del inverso de la velocidad y herramientas analíticas de aprendizaje supervisado de máquinas de soporte vectorial (SVM).

El modelo estará en capacidad de clasificar entre 4 niveles de alertas distintos periodos de una serie de tiempo de deformaciones registradas por instrumentos instalados para dicho fin, afrontando la dificultad del limitado y poco variado número de datos para entrenamiento del modelo mediante el uso de datos sintéticos, la validación final del modelo se realiza mediante las observaciones tomadas por el instrumento.

1 Planteamiento del problema

1.1 Problema de negocio.

En el desarrollo de proyectos de ingeniería específicamente en las obras subterráneas se emplean distintos instrumentos que permiten controlar el estado de las deformaciones en las excavaciones sub - superficiales, los registros de datos se compilan como observaciones ordenadas bajo una secuencia temporal, lo que permite determinar la evolución de las deformaciones en el tiempo de la obra específica.

El seguimiento a las deformaciones permite tener control sobre la evolución de la obra respecto a las condiciones esperadas por diseño, el comportamiento de estos registros es usado para definir alertas durante el desarrollo del proyecto, una adecuada y sobre todo oportuna una identificación de cada una de alerta permite salvaguardar el correcto desarrollo del proyecto y de las personas que allí se encuentran, por lo cual es crítico emplear herramientas analíticas que optimicen el proceso de seguimiento de esta instrumentación.

1.2 Aproximación desde la analítica de datos.

El modelo implementado utilizará el método de Máquinas de Soporte Vectorial (SVM) para clasificación de alertas en series de tiempo. Las alertas son definidas mediante el pronóstico de falla del geomaterial al extrapolar con tendencia lineal a cero en la salida gráfica del inverso de la velocidad de la deformación.

Teniendo definidas las alertas se generarán los datos de entrenamiento que permitan evaluar las series de tiempo mediante una clasificación multiclase. El modelo permitirá definir con mayor eficiencia si la serie de datos registrada por el instrumento presenta un tipo de alerta donde se deban tomar medidas correctivas o de advertencia dentro de una obra subterránea.

1.3 Origen de los datos.

Los datos corresponden a series de tiempo de deformaciones en geomateriales registradas por instrumentos tipo extensómetros, los cuales se utilizan para medir la convergencia o cierre del contorno de excavación de una obra subterránea. Los datos para entrenamiento y las primeras iteraciones son de origen sintético, contruidos mediante el uso de una ecuación de la curva logística, los datos para la evaluación final del modelo son tomados a partir de las observaciones de deformación de los extensómetros que registran lecturas con una frecuencia horaria desde su instalación, alcanzado en algunos casos acumulado de datos de hasta dos años.

1.4 Métricas de desempeño.

El modelo implementado utiliza un algoritmo de clasificación, por lo que se utilizarán métricas relacionados para este tipo de tareas, se utilizará el método estimador del score, el método de la matriz de confusión y el método de reporte de clasificación, que contiene las métricas precision, recall y 'F-score', todas estas métricas se encuentran incluidas en la librería scikit-learn.

Las métrica de negocio se miden a partir de las observaciones reales de los instrumentos teniendo como referencia alertas críticas de falla donde efectivamente ocurrieron colapsos de geomateriales dentro de obras subterráneas, se evaluaría el acierto del modelo relacionando el pronóstico de falla, respecto a la materialización de la misma, se esperaría que el modelo indique la alerta correcta con un porcentaje de acierto no menor del 95% y el pronóstico de falla presente un desfase menor a 3 unidades de tiempo de toma las lecturas (minutos, horas, días), entre el momento de ocurrencia de la falla y el momento pronosticado.

2 Datos

El modelo consiste en una aplicación de la analítica de datos en la metodología semi-empírica del inverso de la velocidad para el pronóstico de falla por ruptura en geomateriales. Este método fue propuesta inicialmente por Fukuzono (1985), y aplicado por otros autores como Rose, N.D. (2007), Osasan, K.S. (2014), Carlà (2016), Intrieri (2019) (Fukuzono, 1985) (Crosta & Agliardi, 2003) (Rose & Hungr, 2007) (Osasan & Stacey, 2014) (Carlà, y otros, 2016) (Intrieri, Carlà, & Gigli, 2019) entre otros. En este sistema se utilizan datos de instrumentos de medición de deformaciones compilados como una serie de tiempo para pronosticar un momento de falla.

Para entrenar un modelo analítico que identifique diversos niveles de alerta, se necesitan gran cantidad de datos que recreen la variedad de comportamiento posibles que pueden darse durante el proceso de deformación bajo los distintos grados de alerta. Esta variabilidad de datos no siempre es detectada por lo instrumentos instalados para este fin, ya sea porque no se presente la suficiente cantidad de instrumentos, tiempo de recolección de datos o heterogeneidad de comportamiento de estos. Lo anterior, implica que se requiera de datos sintéticos que simulen distintos comportamientos de deformación dentro de los umbrales de alerta identificados. La construcción de estos datos y conformación del Dataset se presentará a continuación.

2.1 Datos originales.

Los datos para la aplicación del modelo se tomarán de las observaciones de deformación registradas por extensómetros ubicados en excavaciones subterráneas de una obra de ingeniería en curso, de donde ya se han identificados los periodos donde se generaron las mayores deformaciones y se tienen etiquetadas los niveles de alarma.

Para la construcción de los datos sintéticos se tuvo como base para su generación el comportamiento típico de una deformación en el tiempo, en conjunto con la relación del inverso de la tasa de cambio de la deformación y el momento de falla de un geomaterial identificada por Fukuzono (Fukuzono, 1985). Concerniente al comportamiento típico de las deformaciones, es

habitual observar que este se modele como curvas de tendencia parabólica, exponencial, logarítmicas, o logísticas (Crosta & Agliardi, 2003) (Rose & Hungr, 2007) (Intrieri, Carlà, & Gigli, 2019), sin importar el tipo de curva o ecuación seleccionada, todos los autores coinciden en señalar que la principal de alerta de un potencial colapso dentro de una obra subterránea en un momento t_f podría ocurrir cuando una deformaciones o convergencias de magnitud ΔD se generan en periodos de tiempo t_i cada vez más pequeños respecto al anterior, tal como se representa en la Ilustración 3.

El presente trabajo empleará la Ecuación 1, de tipo logística, para la generación de los datos sintéticos y recrear así un comportamiento típico de la deformación de geomateriales. Así, $-L$ representa la asíntota inferior de la curva; $-k$ es la tasa de crecimiento de la deformación; t corresponde a la abscisa o variable dependiente, para este caso, el tiempo en horas; t_0 indica el punto de medio entre la asíntota superior e inferior y el punto de inflexión de la curva; y el parámetro *var* es un factor de variabilidad que imprime la aleatoriedad a los datos sintéticos para la conformación de distintas series de tiempo.

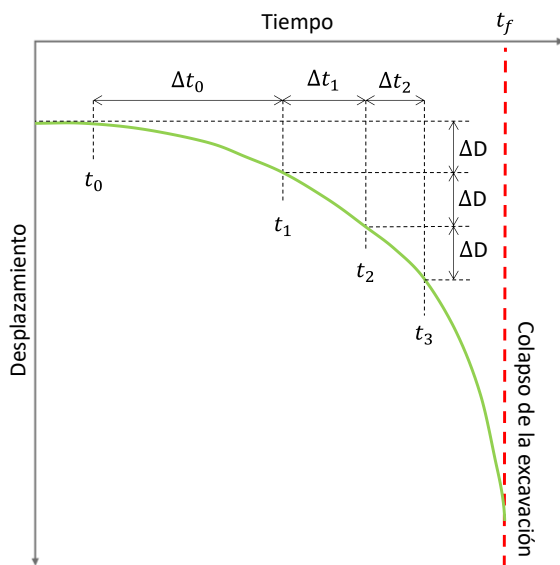


Ilustración 3. Representación de un registro de deformación de un extensómetro previo al colapso dentro de una excavación subterránea.

Ecuación 1. Ecuación logística para la generación de datos sintéticos.

$$\text{Curva de Deformación} = \frac{-L}{1 + e^{(-k \cdot (t-t_0)) \cdot var}}$$

En este modelo se clasificarán 4 clases diferenciados como niveles de alarma. Los datos sintéticos dentro de cada clase estarán limitados comportamientos umbrales que separan un tipo de nivel de alerta de otro. Utilizando la Ecuación 1 y la relación de Fukuzono, se trazaron tres series de tiempo en la que pronostica un lapso de falla de 21, 15 y 8 que corresponderán a los umbrales 1, 2 y 3, respectivamente, así, en la Ilustración 4 se observa un ejemplo con 5 series de tiempo para cada nivel de alerta con su respectivo umbral.

De esta manera, los datos sintéticos dentro del nivel de alerta 1 tendrán pronósticos de falla mayores a 21 días a partir del dato de la hora 12, para el nivel de alerta 2 se tendrán series de tiempo donde se esperaría una falla entre 21 y 15 días luego del dato de la hora 12, así hasta el nivel de alerta 4 donde se tendrían series de tiempo con pronósticos de falla menores a 8 días. En Ilustración 5 se presenta un detalle con la serie de datos de los umbrales 2 y 3, expresado en el inverso de la velocidad (Inv_U3, Inv_U2), junto con la proyección del momento de falla (líneas punteadas Tend Inv_U3 y Tend Inv_U2) y los niveles de alerta generados entre umbrales.

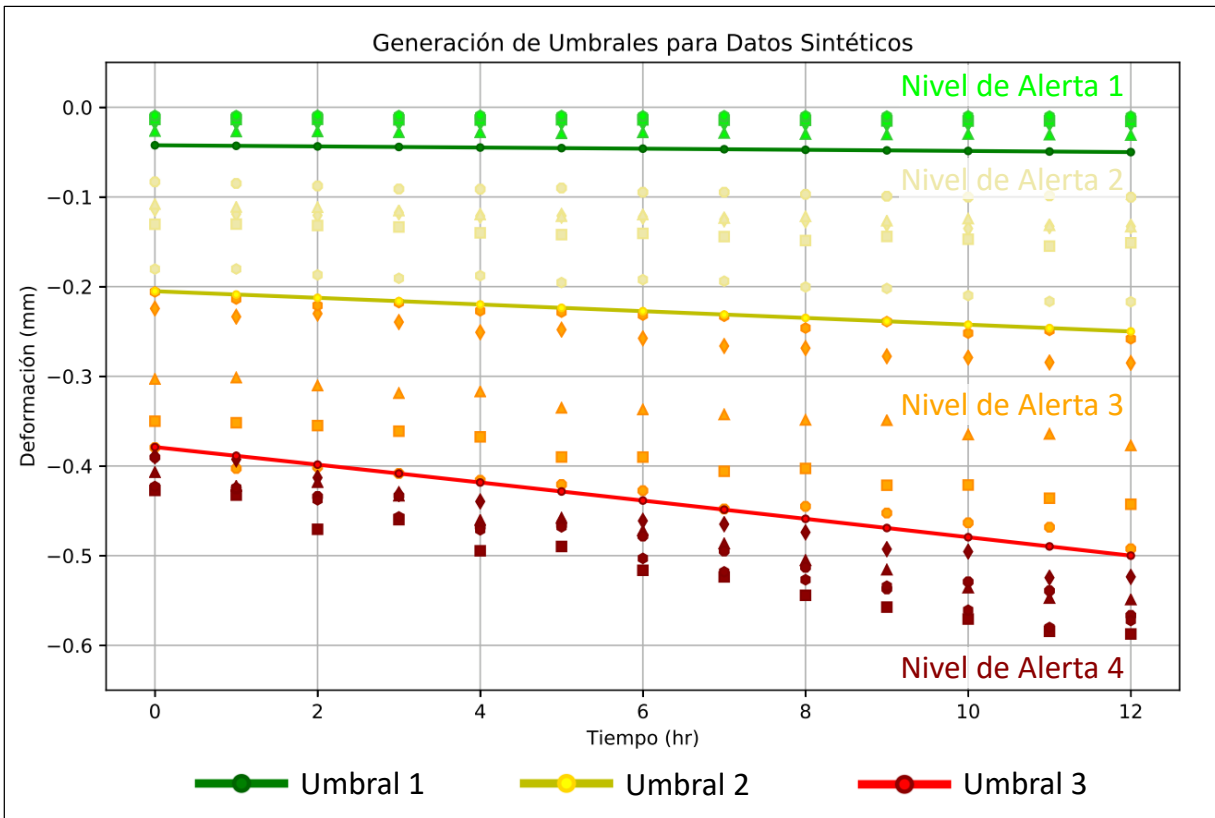


Ilustración 4. Series de tiempo sintéticas dentro de sus respectivos Niveles de Alerta y sus umbrales.

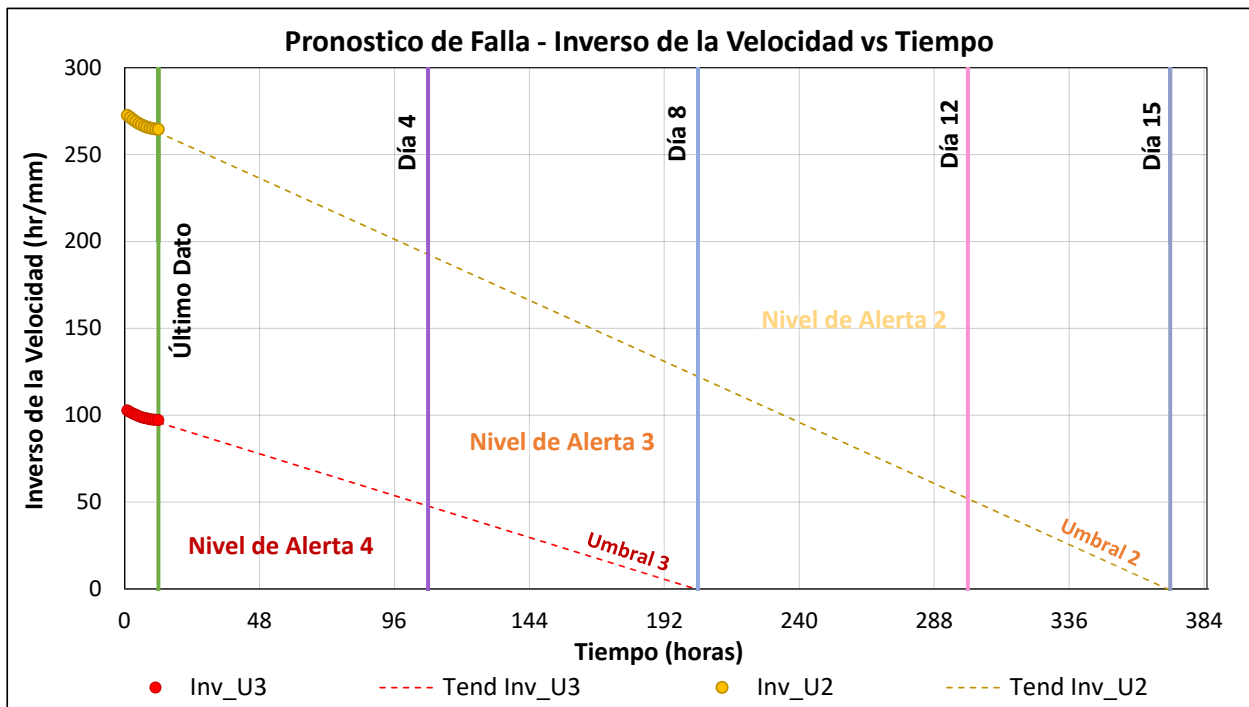


Ilustración 5. Inverso de la velocidad y su pronóstico de falla con relación a los niveles alerta.

2.2 Dataset.

Definida la ecuación de los datos sintéticos y los umbrales para los niveles de alarma, se generaron mil (1000) datos para cada clase agrupados en 4 listas, que finalmente se unirían en solo un Dataframe con 4000 series de tiempo etiquetadas. Los valores de los parámetros para la generación de los datos en cada nivel de alerta se presentan en la Tabla 1. Los datos generados se podrían visualizar como se presenta en la Ilustración 6, donde muestra los datos de la ecuación sin aplicar la variabilidad, los datos con la variabilidad, el rango y dominio de la ecuación y el periodo de 12 horas de datos a utilizar. Una vez agrupados en un Dataframe las 4 listas con los datos generados y su etiqueta del nivel de alerta se barajan aleatoriamente las filas del Dataset y se reinician los índices de manera que se tenga la mayor variabilidad en los datos. Por último, el Dataset inicial tendrá una dimensión de 4000×13 , similar a como se presenta en la Ilustración 7 y como se detalla en el [Notebook 01 Generación de Datos](#).

Tabla 1. Valores de los parámetros de la Ecuación 1 para la generación de los datos sintéticos para cada nivel de alerta.

Parámetro	Nivel de Alerta 1		Nivel de Alerta 2		Nivel de Alerta 3		Nivel de Alerta 4	
	Valor mínimo	Valor máximo	Valor mínimo	Valor máximo	Valor mínimo	Valor máximo	Valor mínimo	Valor máximo
L (mm)	0,01000	0,10000	0,10001	0,50000	0,50001	1,00000	1,05000	1,20000
K	0,01000	0,02560	0,02560	0,03024	0,03024	0,04117	0,04118	0,05265
var	0,95000	1,05000	0,95000	1,05000	0,95000	1,05000	0,95000	1,05000
t(h)	12	24	12	24	12	24	12	24
t0	24		24		24		24	

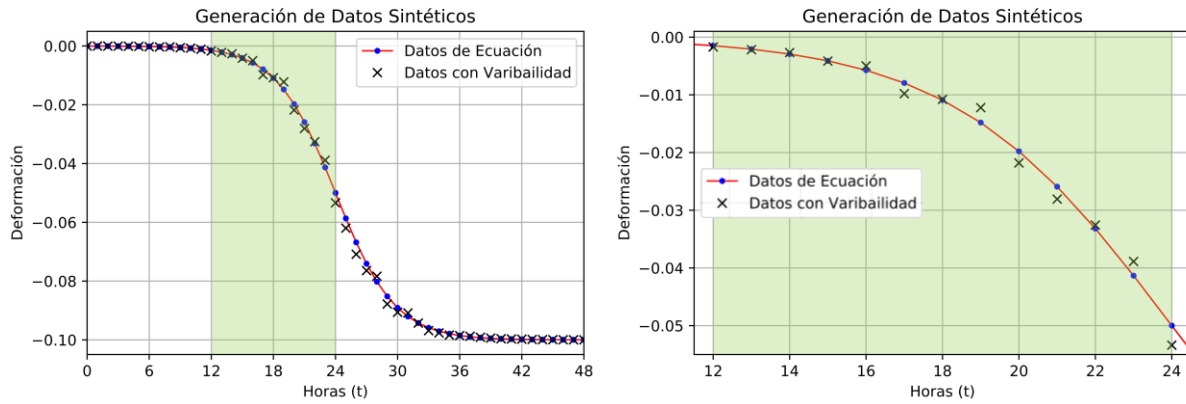


Ilustración 6. A la Izquierda se presenta los datos sintéticos generados para una serie de tiempo de 48 horas, se resalta el periodo entre 12 y 24 horas que es cuando se presenta la mayor tasa de cambio en la deformación, detallándose a la derecha.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Target
0	-0.022558	-0.023032	-0.024023	-0.025349	-0.026347	-0.028589	-0.028350	-0.029052	-0.031695	-0.032396	-0.033315	-0.033515	-0.035920	Alerta 3
1	-0.005890	-0.007091	-0.007399	-0.008642	-0.009114	-0.010009	-0.011171	-0.012715	-0.014414	-0.015241	-0.017518	-0.018832	-0.020903	Alerta 4
2	-0.031982	-0.030778	-0.031850	-0.032960	-0.032796	-0.034148	-0.035528	-0.036158	-0.036193	-0.037175	-0.038341	-0.040353	-0.039709	Alerta 2

Ilustración 7. Visualización del Dataset de datos sintéticos.

3 Proceso de Analítica.

El proceso analítico de este trabajo consta de 6 etapas separadas inicialmente en dos ramas de trabajo que convergen en la etapa 5 del proyecto. En la Ilustración 8, se muestran las 6 etapas numeradas y diferenciadas por distintas tonalidades en los recuadros del flujograma.

El proceso inicia con la concepción del modelo (tonalidades rosadas), donde se identifica la necesidad de negocio y la base teórica para el desarrollo del modelo predictivo (inverso de la velocidad) y se definen los niveles de alerta que deberán predecirse al final del modelo (Alerta Verde, Alerta Amarilla, Alerta Naranja y Alerta Roja). Posterior a la primera etapa se genera la ramificación del proceso, esta obedece principalmente a la afinidad del origen de los datos, por un lado, se encuentran las etapas asociadas a datos de origen sintético en contraste con la otra rama asociada a etapas desarrolladas con los datos observados por los instrumentos.

Sobre la rama asociada a datos de origen sintéticos se encuentra la etapa de generación de datos de entrenamiento identificada por las tonalidades grises en el flujograma de la Ilustración 8, los por menores de esta etapa ya fueron descritos en el Ítem 2 de este informe. Prosiguiendo por esta rama se continúa con la etapa de entrenamiento y selección del modelo la cual se presenta en tonalidad amarilla clara, durante esta etapa utilizando únicamente los datos sintéticos se evaluaron las distintas transformaciones de las características de los datos mediante las funciones Kernel ofrecidas en la librería [sklearn.svm](#) (lineal, polinómica, RBF y sigmoide), en paralelo con las variación de la función de la forma de la clasificación (uno contra uno u OvO y uno contra el resto u OvR). Ejecutadas estas iteraciones, a partir de los resultados de las métricas (Ítem 1.4, para ver métricas utilizadas), se define los parámetros a utilizar en el modelo.

Adentrándose a la rama asociada con los datos observados, se inicia con la etapa de registro y procesamiento de datos, identificada por los procesos marcados en coloración verde, esta etapa consta de la obtención de los datos mediante instrumentación y su preprocesamiento para la entrada de estos registros al modelo en producción. No obstante, para que el modelo entre en producción se debe pasar previamente por la etapa de validación del modelo, identificada por el color morado

en la Ilustración 8, esta validación se realiza ejecutando datos observados etiquetados y validando el desempeño del modelo, si el resultado es aceptable según las métricas del negocio (ver Ítem 1.4), se procede con la etapa del Modelo de producción, de lo contrario se debe volver a ejecutar las etapas 2 y 3. Finalmente la sexta última etapa, “Modelo en Producción” consistirá en la aplicación del algoritmo desarrollado para la ejecución del monitoreo periódico del proyecto de excavación subterránea en estudio. El detalle programático de todo este proceso se puede encontrar en el repositorio [Monografía](#).

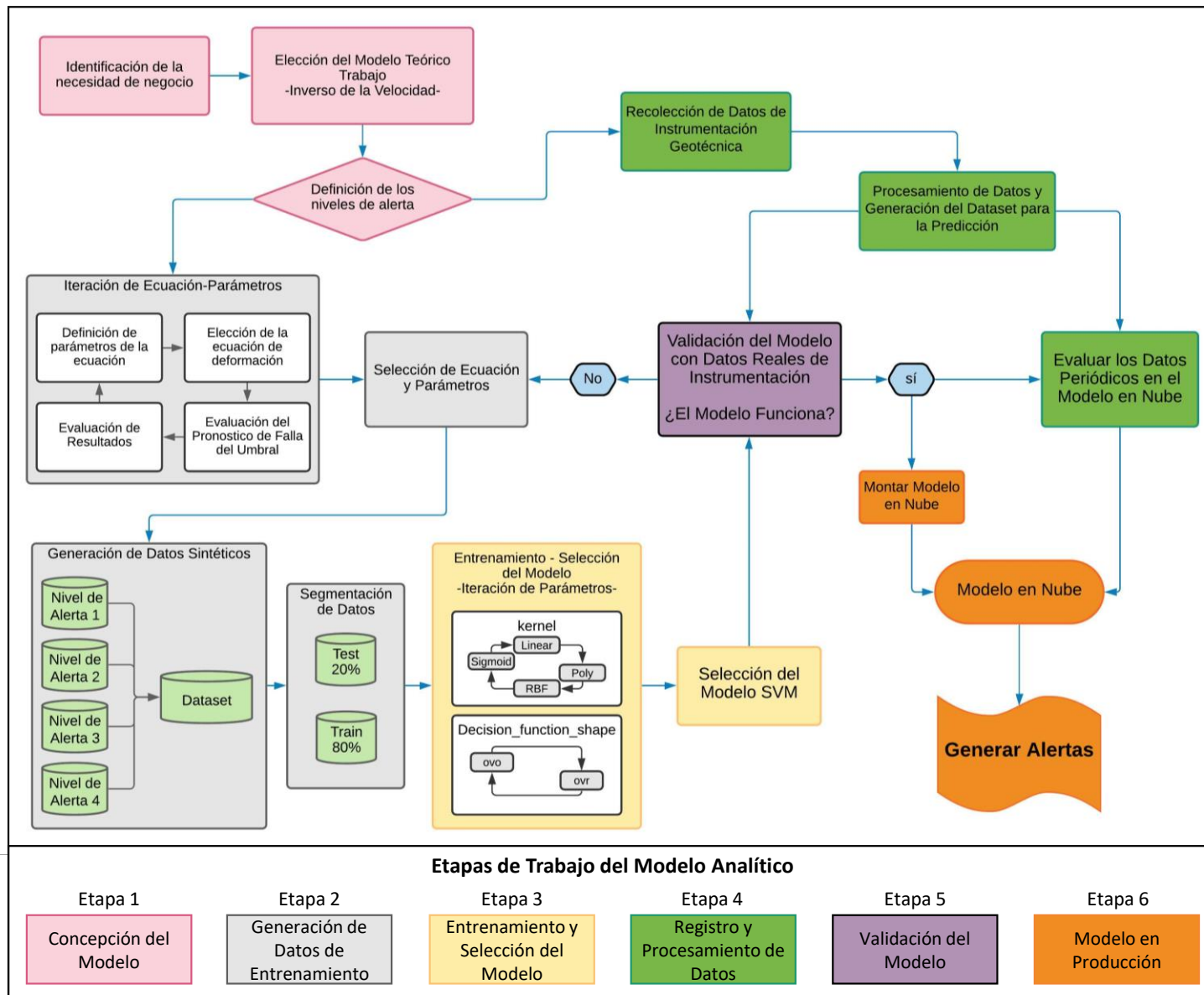


Ilustración 8. Diagrama deflujo del proceso de analítica

4 Metodología.

4.1 Baseline.

La primera iteración tuvo como principal fin la generación de los datos sintéticos y la definición de los parámetros del modelo que ofreciesen un mejor desempeño a la luz de los resultados de las pruebas de validación.

La generación de los datos sintéticos se realizó aprovechando la ecuación de una curva logística, de allí los primeros retos consistieron en la selección de los parámetros a ingresar en la ecuación y el tramo de la curva a utilizar. Los datos de la tasa de crecimiento de la deformación k y el valor L de la asíntota estaban definidos por los umbrales y el nivel de alerta por lo cual se fijaron desde el inicio de la propuesta de trabajo (ver Ilustración 8) y sus complejidades ya se discutieron en el Ítem 2. Respecto al tramo de la curva, inicialmente se planteó el uso de una curva asociada un periodo de datos de 24 horas con un punto de inflexión en la curva en $t_0 = 12$ y se usarían los datos de la hora 0 a la hora 12 (ver gráficas superiores en Ilustración 9), No obstante, se encontró que se obtendría un resultado similar realizando una curva con un periodo de 48 horas con un punto de inflexión en $t_0 = 24$, y se usarían los datos de la hora 12 a la hora 24 (ver gráficas inferiores en Ilustración 9). Finalmente se eligió la segunda opción pues ofrecía la posibilidad de tener datos de tiempo previos, lo cual resultó más conveniente para otras iteraciones que se detallarán más adelante.

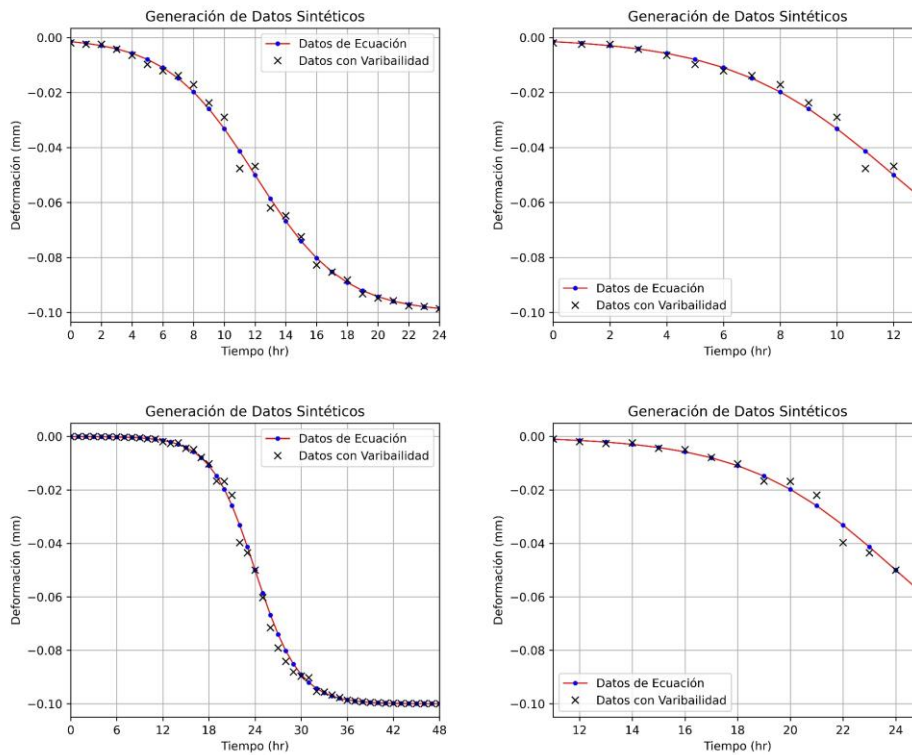


Ilustración 9. Curvas de deformación con variaciones en la selección del periodo de tiempo a emplear, detalles de la generación de estas curvas en el repositorio [Evaluación tiempo](#).

Definido el rango de tiempo (eje x) a utilizar para la generación de los datos sintéticos, durante la primera iteración se prosiguió con la definición de los parámetros a utilizar en el modelo analítico. Para esto se realizaron entrenamientos, pruebas y validaciones de resultados utilizando cada una de las transformaciones de las características de los datos mediante las funciones Kernel ofrecidas en la librería [sklearn.svm](#) (lineal, polinómica, RBF y sigmoide), además a la par que se realizaba este proceso de variación de los Kernel, se analizaban las diferencias de resultados al aplicar el cambio de la función de la forma de la clasificación entre uno contra uno (OvO) y uno contra el resto u (OvR).

Respecto al cambio de función de la forma se observó que para este modelo cualquiera de las 2 modalidades arroja el mismo resultado en las métricas de resultados, de esta manera se decide implementar el modelo mediante la metodología de uno contra uno (OvO), pues es la recomendada por defecto en la librería de [sklearn.svm](#) para clasificaciones multiclase.

Respecto al cambio de características mediante la aplicación de las funciones Kernel las métricas de los implementados se observan de la Ilustración 10 a la Ilustración 17. Allí se muestra que los mejores resultados se obtuvieron mediante los Kernel de tipo lineal y polinómico de grado dos. Si bien la utilizando un Kernel lineal se obtienen mejores resultados a nivel general, la diferencia entre resultados en mínimo, además, el método polinómico de grado 2 ofrece un mejor desempeño para la clasificación de alertas de tipo 3 y 4, las cuales son consideradas las más críticas. Por lo anterior, se define como modelo a implementar y un sistema de clasificación SVM con Kernel polinómico de grado 2 con una forma de clasificación tipo ‘OvO’. En el [Notebook 03 Primera Iteracion](#), se encuentra el detalle programático de este proceso.

SVM Kernel Lineal				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,94	1	0,97	210
Alerta_2	0,97	0,93	0,95	206
Alerta_3	1	0,94	0,97	206
Alerta_4	0,97	1	0,98	178
<hr/>				
accuracy			0,97	800
macro avg	0,97	0,97	0,97	800
weighted avg	0,97	0,97	0,97	800

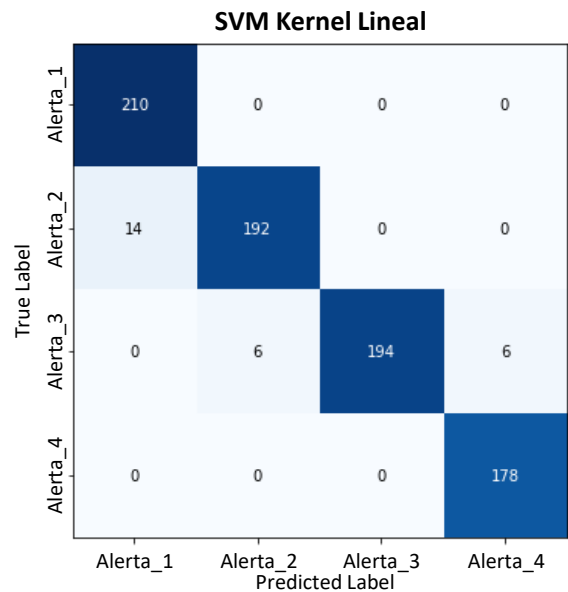


Ilustración 10. Métricas de resultados y matriz de confusión Kernel Lineal

SVM Kernel Polinómico de grado 2				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,9	1	0,95	210
Alerta_2	0,99	0,87	0,93	206
Alerta_3	0,99	1	0,99	206
Alerta_4	1	1	1	178
<hr/>				
accuracy	-	-	0,96	800
macro avg	0,97	0,97	0,97	800
weighted avg	0,97	0,96	0,96	800

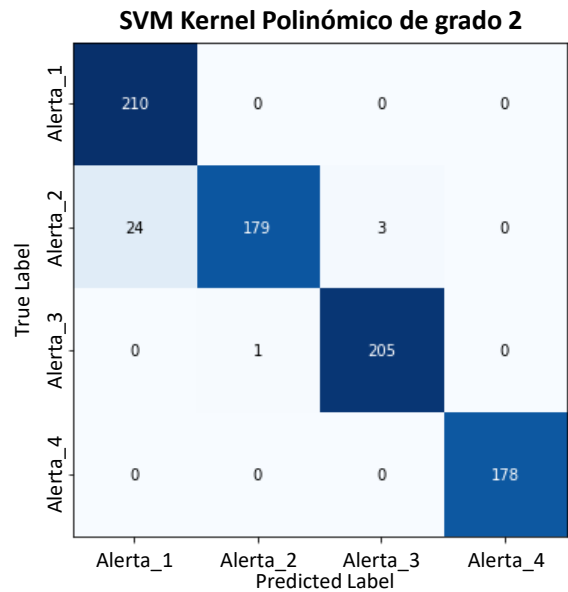


Ilustración 11. Métricas de resultados y matriz de confusión Kernel Polinómico de grado 2

SVM Kernel Polinómico de grado 3				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,85	1,00	0,92	210
Alerta_2	0,99	0,81	0,89	206
Alerta_3	0,99	1,00	0,99	206
Alerta_4	1,00	1,00	1,00	178
<hr/>				
accuracy			0,95	800
macro avg	0,96	0,95	0,95	800
weighted avg	0,96	0,95	0,95	800

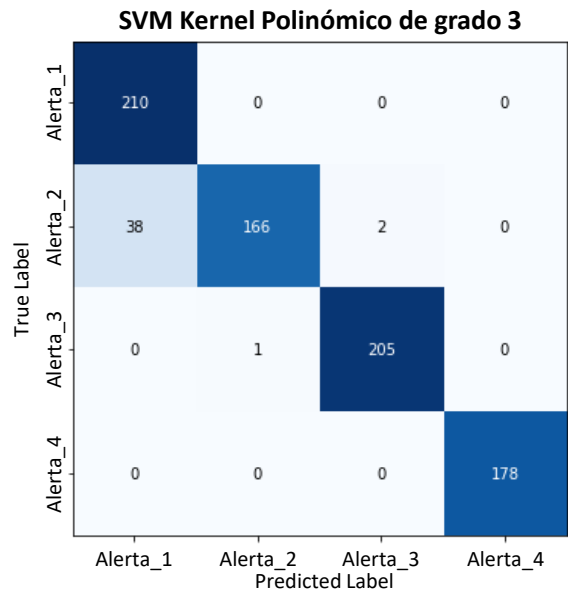


Ilustración 12. Métricas de resultados y matriz de confusión Kernel Polinómico de grado 3

SVM Kernel Polinómico de grado 4				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,80	1,00	0,89	210
Alerta_2	1,00	0,73	0,85	206
Alerta_3	0,99	1,00	1,00	206
Alerta_4	1,00	1,00	1,00	178
accuracy			0,93	800
macro avg	0,95	0,93	0,93	800
weighted avg	0,94	0,93	0,93	800

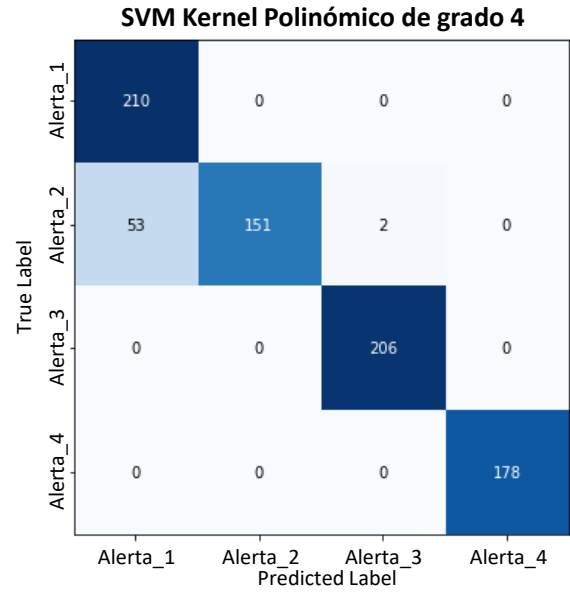


Ilustración 13. Métricas de resultados y matriz de confusión Kernel Polinómico de grado 4

SVM Kernel Polinómico de grado 5				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,76	1,00	0,86	210
Alerta_2	1,00	0,67	0,80	206
Alerta_3	0,99	1,00	1,00	206
Alerta_4	1,00	1,00	1,00	178
Summary Metrics				
accuracy			0,92	800
macro avg	0,94	0,92	0,92	800
weighted avg	0,93	0,92	0,91	800

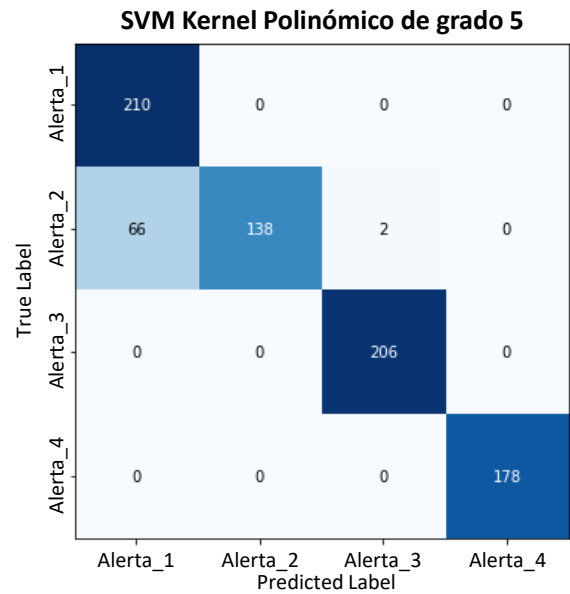


Ilustración 14. Métricas de resultados y matriz de confusión Kernel Polinómico de grado 5

SVM Kernel Polinómico de grado 6				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,73	1,00	0,84	210
Alerta_2	1,00	0,61	0,76	206
Alerta_3	0,99	1,00	1,00	206
Alerta_4	1,00	1,00	1,00	178
Summary Metrics				
accuracy			0,90	800
macro avg	0,93	0,90	0,90	800
weighted avg	0,93	0,90	0,90	800

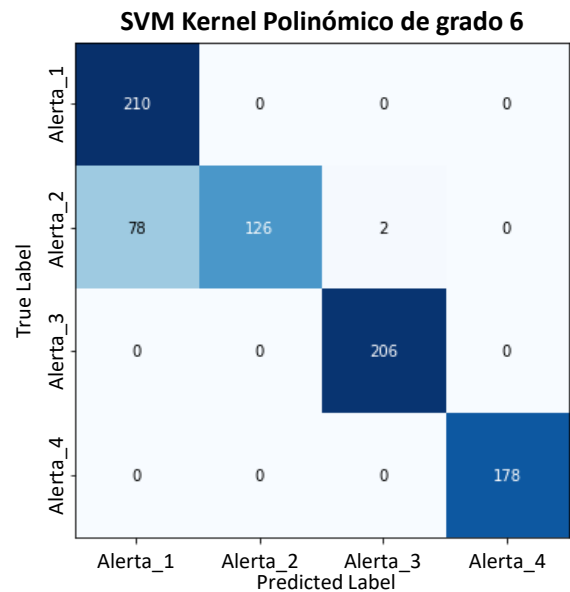


Ilustración 15. Métricas de resultados y matriz de confusión Kernel Polinómico de grado 6

SVM Kernel Sigmoide				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	1,00	0,73	0,84	210
Alerta_2	0,00	0,00	0,00	206
Alerta_3	0,00	0,00	0,00	206
Alerta_4	0,02	0,04	0,02	178
Summary				
accuracy			0,20	800
macro avg	0,25	0,19	0,22	800
weighted avg	0,27	0,20	0,23	800

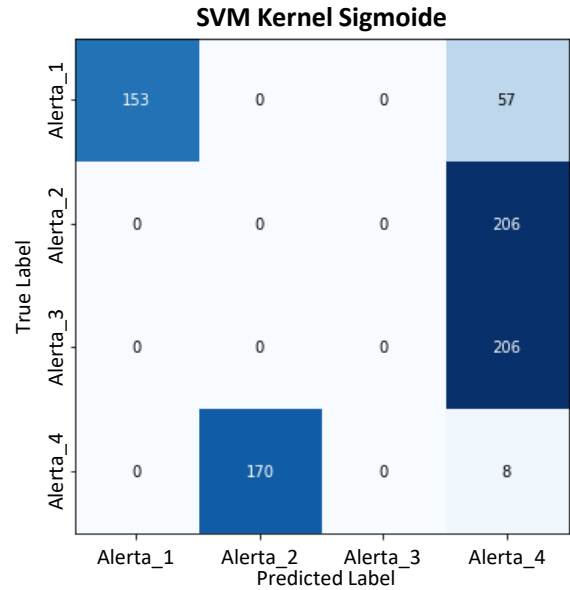


Ilustración 16. Métricas de resultados y matriz de confusión Kernel Sigmoide

SVM Kernel RBF				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,94	1,00	0,97	210
Alerta_2	0,97	0,93	0,95	206
Alerta_3	0,99	0,96	0,98	206
Alerta_4	0,98	1,00	0,99	178
Summary				
accuracy			0,97	800
macro avg	0,97	0,97	0,97	800
weighted avg	0,97	0,97	0,97	800

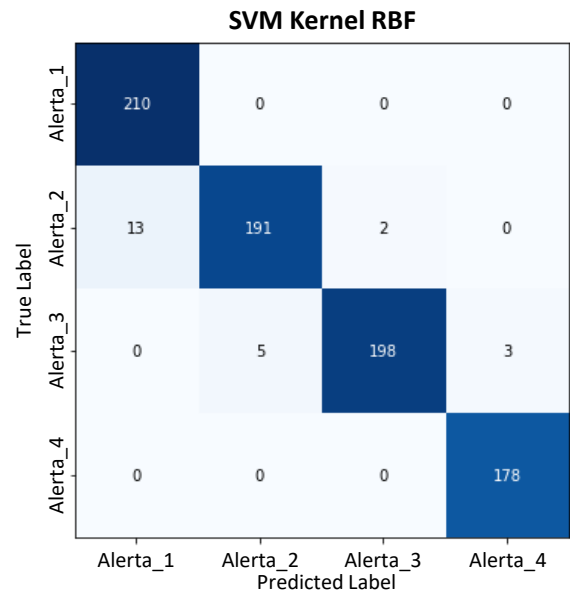


Ilustración 17. Métricas de resultados y matriz de confusión Kernel RBF

4.2 Iteraciones y Evolución.

Una vez definidos los parámetros del modelo a utilizar luego de la primera iteración se procede a hacer una validación mediante el uso de datos observados, para esto se utiliza el archivo [04 Exten E3 P2](#), que contiene datos de deformación registrados por un instrumento tipo extensómetro. Para la ejecución del modelo estos datos son procesados mediante una función que realiza un Look-back, que consiste en una estructuración de datos que reporta el dato para fecha i junto los datos de las fechas $i-12$, una vez adecuado el Dataset en esta estructura se genera la evaluación del modelo.

Al realizar la evaluación del modelo se hace evidente que el entrenamiento no tuvo en cuenta la razón de cambio entre registros si no la magnitud de los valores ingresados, generando niveles de alerta críticos (Alerta_4) cuando se sobrepasaban ciertas magnitudes sin importar que la deformación en el periodo de tiempo evaluado se mantuviese estable, lo cual va en contravía el concepto fundamental de la teoría del inverso de la velocidad. Siguiendo la ruta planteada en la Ilustración 8, se generaron nuevos datos para realizar un nuevo entrenamiento del modelo.

El proceso de generación de nuevos datos se detalla en el [Notebook 06 Generación de nuevos datos](#), como tal el proceso de creación de los datos sintéticos para entrenamiento no varía mucho, no obstante teniendo en cuenta el principio fundamental de la teoría del inverso de la velocidad se agrega una etapa adicional en la generación de los datos que consiste en calcular la tasa de cambio entre registros, a fin de mantener las predicciones para reportes de 12 horas se utiliza el dato de la hora 11 de manera que se tengan la tasa de cambio de cambio del espectro de 12 a 24 horas, de aquí la conveniencia de no haber tomado para entrenamiento un rango de tiempo de 0 a 12, como ya se había mencionado.

Los Dataset resultantes de la nueva generación de datos se encuentran en el repositorio [Monografía](#) con los nombres [2do_Dataset](#) y [2do_Dataset_Tasas](#).

Ya con el nuevo Dataset se procede a generar una nueva iteración la cual se detalla programáticamente en el [Notebook 10 Iteración Final](#), en esta nueva iteración se utilizan el Dataset [2do_Dataset_Tasas](#) generado en el [Notebook 06](#) y adicionalmente se carga la data del

extensómetro E3_P2 con el procesamiento de Look-back a 13 horas y los datos etiquetados evidenciados en el archivo [09 Exten E3 P2 + Label](#) del repositorio.

En esta iteración se entrena el modelo con la totalidad de los nuevos datos sintéticos de tasa a la vez que también se evalúan todos los datos registrados por el extensómetro. Al ejecutar el modelo de esta manera se observa un amplio desbalance entre los datos registrados por el instrumento como niveles de Alerta_1 y Alerta_2 contra los niveles de 3 y 4 más críticos y menos comunes en registro. Por lo anterior, se procede a un nuevo Dataset muestreando de manera aleatoria una cantidad de datos idéntica a la cantidad de alertas registrada en el nivel 4 (69). Los resultados de estas iteraciones se discutirán en el siguiente capítulo.

5 Resultados

5.1 Métricas.

Realizadas las nuevas iteraciones con los datos de validación ajustando el entrenamiento del modelo a las tasas de cambio, se obtiene una mejoría respecto al modelo anterior que se basa en la deformación neta, no obstante, es indudable el desbalance en el número de registros de niveles de alerta a evaluar. Al observar la Ilustración 18, es evidente ante inspección visual que son más reducidos los episodios de alta aceleración que podían corresponder con un Nivel de Alerta_4, en relación con periodos de tiempo en que la pendiente de la deformación es menos inclinada.

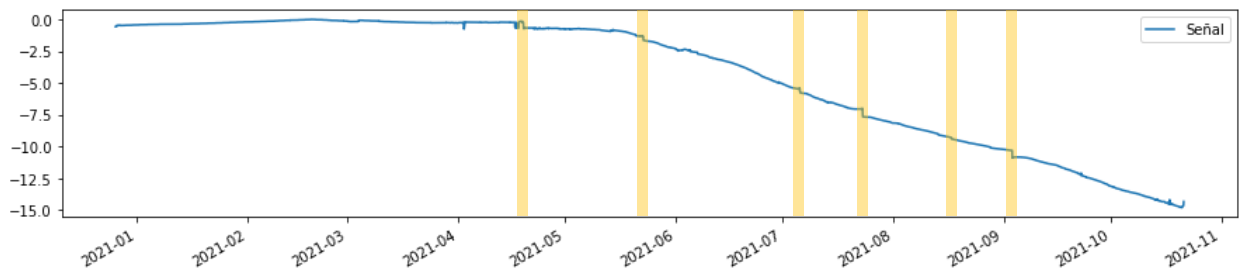


Ilustración 18. Comportamiento generalizado de la deformación en el instrumento Exten_E3_P2. Se resalta en amarillo los episodios de mayor aceleración en la deformación.

Las métricas de validación del modelo con todos los datos observados por el instrumento se presentan en la Ilustración 19, allí se constata el desbalance ya mencionado en la cantidad de alertas evaluadas, siendo de 8635 para el nivel de Alerta_1 y 69 para el Nivel de Alerta_4. A nivel general el modelo presenta un valor sobresaliente en la exactitud ($accuracy = 0,87$). No obstante, para este modelo es de mayor prioridad las alertas críticas (Alerta_3 y Alerta_4) por lo cual, los bajos valores de precisión y el valor F para estas clasificaciones, requieren de la búsqueda de alternativas de evaluación de las métricas del modelo.

En busca de una validación y evaluación del modelo más balanceado, se hace un muestreo aleatorio de niveles de alerta etiquetados en los datos de validación reportados por el instrumento, igualando la cantidad de registros para cada nivel de alerta a 69 y así observar el desempeño del modelo de clasificación, especialmente para el Nivel de Alerta_3 y Nivel de Alerta_4.

SVM Kernel Polinómico de grado 2 - Data Completa				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,99	0,93	0,96	8635
Alerta_2	0,82	0,73	0,77	3299
Alerta_3	0,13	0,75	0,23	189
Alerta_4	0,47	0,87	0,61	69
<hr/>				
accuracy			0,87	12192
macro avg	0,60	0,82	0,64	12192
weighted avg	0,93	0,87	0,89	12192

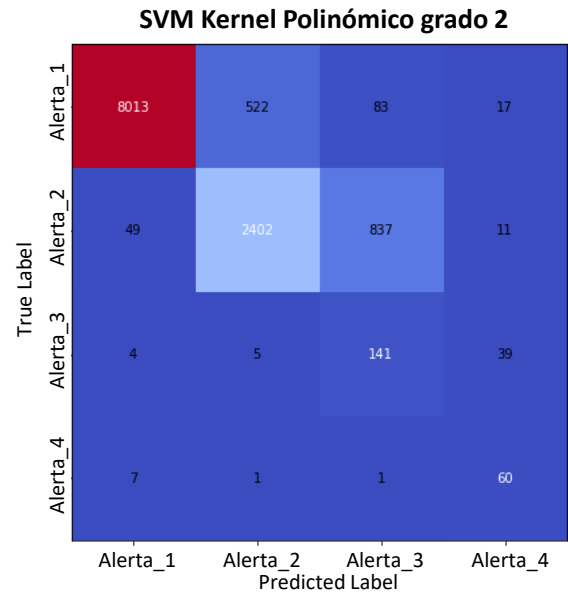


Ilustración 19. Métricas de resultados y matriz de confusión data de validación completa.

En contraste con el ejercicio anterior, donde se evaluaron todos los datos, en la validación con datos balanceados se observa una disminución en la exactitud general del modelo pasando de un accuracy de 0,87 a uno de 0,82. Sin embargo, se nota una especial mejoría en los valores de precisión y el valor F en los niveles de alerta 3 y 4, lo que constituye un resultado más satisfactorio.

SVM Kernel Polinómico de grado 2 - Data Balanceada				
Nivel de Alerta	Precision	Recall	F1-score	Support
Alerta_1	0,86	0,93	0,90	69
Alerta_2	0,86	0,72	0,79	69
Alerta_3	0,73	0,74	0,73	69
Alerta_4	0,81	0,87	0,84	69
<hr/>				
accuracy			0,82	276
macro avg	0,82	0,82	0,81	276
weighted avg	0,82	0,82	0,81	276

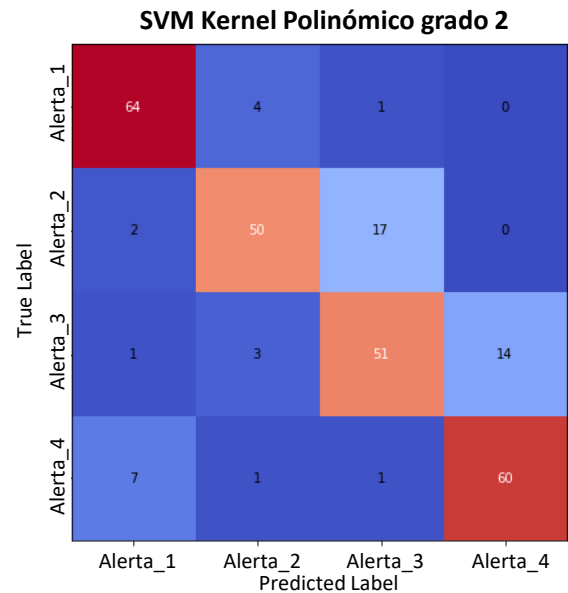


Ilustración 20. Métricas de resultados y matriz de confusión data de validación balanceada.

5.2 Evaluación Cualitativa.

En las primeras ejecuciones del modelo utilizando únicamente los datos sintéticos se presentaron resultados bastante prometedores basándonos en valores de las métricas de desempeño, especialmente para los modelos lineales y polinómicos de hasta grado 3 donde la exactitud alcanzaba valores mayores de 0,95, no obstante, al tratarse de datos sintéticos tanto para entrenamiento y test, era más fácil incurrir en un sobreajuste, sin embargo, esos ejercicios permitieron la mejor definición del modelo a desarrollar.

Al realizar las validaciones con datos reales se obtuvieron dos conclusiones principales, la inconveniencia de entrenar y evaluar el modelo mediante datos de deformación bruta y la afectación en los valores de las métricas al evaluar cuando se presenta una desproporción en el número de registros a clasificar. En ambos casos se pudo solucionar mediante procedimientos de feature engineering.

Desde las métricas de negocio se esperaba que al validar el modelo mediante datos observados por los instrumentos se obtuviese una precisión superior al 0,95, no obstante, desde el panorama del machine learning los valores finales obtenidos son considerablemente sobresalientes.

Desde el punto de vista del negocio la matriz de confusión de la Ilustración 20 muestra un comportamiento conservador en la generación de errores, pues si bien se esperaba mayor exactitud y confiabilidad en los resultados, sería más desfavorable encontrar que las imprecisiones subestiman el nivel de alerta real, lo que se observa que ocurre lo contrario encontrándose inexactitudes que sobreestiman el nivel de alerta por lo que el modelo se mantiene dentro del espectro de la seguridad que es el fin mismo de este.

6 Conclusiones

Dando seguimiento a literatura donde confluyan instrumentación geotécnica y machine learning, este proyecto es de los primeros evidenciados donde se emplee herramientas analíticas para la predicción de alarmas en obras subterráneas, teniendo como base teórica el sistema del inverso de la velocidad.

El modelo presenta un acercamiento inicial de como emprender los procedimientos de entrenamiento del modelo, enfrentándose a la limitante del número y la variabilidad de los de datos y la validación mediante datos observados.

A nivel general y desde el punto de vista del machine learning el modelo presenta unos datos sobresalientes, no obstante, se requieren nuevos ejercicios de manera que se alcance un resultado aceptable dentro de las métricas de negocio, aun así, es importante destacar el carácter conservador de las predicciones a luz de la criticidad de subestimar niveles de alerta.

7 Referencias

- Carlà, T., Intrieri, E., Di Traglia, F., Nolesini, T., Gigli, G., & Casagli, N. (2016). Guidelines on the use of inverse velocity method as a tool for setting alarm thresholds and forecasting landslides and structure collapses. *Landslides*, *14*(2), 517–534.
- Crosta, G., & Agliardi, F. (2003). Failure forecast for large rock slides by surface displacement measurements. *Canadian Geotechnical Journal*, *40*(1), 176–191.
- Fukuzono, T. (1985). A new method for predicting the failure time of a slope. *Fourth International Conference and Field Workshop on Landslides* (págs. 145-150). Tokyo, Japón: Japan Landslide Society.
- Intrieri, E., Carlà, T., & Gigli, G. (Junio de 2019). Forecasting the time of failure of landslides at slope-scale: A literature review. *Earth-Science Reviews*, *193*, 333-349. doi:<https://doi.org/10.1016/j.earscirev.2019.03.019>
- Osasan, K. S., & Stacey, T. R. (2014). Automatic prediction of time to failure of open pit mine slopes based on radar monitoring and inverse velocity method. *International Journal of Mining Science and Technology*, *24*(2), 275–280.
- Rose, N. D., & Hungr, O. (2007). Forecasting potential rock slope failure in open pit mines using the inverse velocity method. *International Journal of Rock Mechanics and Mining Sciences*, *44*(2), 308-320.