Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Área de Ingeniería e Innovación

Informe final de proyecto

Pronóstico de terremotos mediante modelos de machine learning

Integrantes: Diego Acevedo
Sebastián Jana
Renzo Zanca
Profesor: Carlos Flores
Ayudantes: Francisco Santibáñez
Fecha de realización: 20 de diciembre de 2021
Fecha de entrega: 22 de diciembre de 2021
Santiago de Chile

Resumen

Chile es uno de los países con mayor actividad sísmica en el mundo, abarcando varios puestos entre los terremotos más fuertes registrados. Estos desastres naturales pueden generar grandes efectos negativos, devastando grandes zonas del país. Estos han producido tsunamis y derrumbes, destruyendo grandes secciones de nuestro país, y quitándole la vida a muchos chilenos.

Es por esto que existe un organismo del Estado encargado de alertar a la población en caso de un desastre, conocido como el Centro de Alerta Temprana (CAT). Pero los sismos son muy impredecibles, imposibles de predecir, y muy difíciles de pronosticar. Es por esto que el CAT necesita algún tipo de herramienta que le permita facilitar esta tarea.

Se llegó a la solución de crear un modelo de machine learning que usará la información registrada por el Centro Sismológico Nacional a partir del año 2010 para poder pronosticar la magnitud máxima de un sismo un mes y determinados. Los datos que el CSN recopila que se usarán son la fecha, la latitud y longitud, la magnitud y la profundidad.

A partir de estos datos, se define una matriz de características para todos los meses desde el 2010 para cada región, y se intentará predecir la variable de máxima intensidad de sismo el próximo mes, siendo esta una variable continua.

Una vez creada la matriz de características, se crearon 3 modelos de machine learning, los cuales son: Decision Tree Regressor, Random Forest Regressor y Linear Regressor. Realizando una validación cruzada, se obtuvo que el modelo Linear regressor fue el más preciso en terminos de error absoluto medio (MAE)

Índice de Contenidos

Índice de Figuras

Índice de Tablas

1. Definición del problema

1.1. Contexto y problemática

Chile es un país de alta actividad sísmica, donde pueden haber repercusiones desastrosas, dependiendo del lugar e intensidad, pues si ocurre en una zona costera, los movimientos en sí podrían no ser el mayor problema, ya que se pueden producir tsunamis que representan una amenaza a los residentes de esta zona.

Es por esto que si se le avisara a las comunidades previamente acerca de riesgos de índole sísmica, se podrían reducir casualidades al tomar medidas preventivas tales como evacuaciones masivas.

1.2. Usuario

Puesto que el objetivo es informar a la mayor cantidad de personas posibles, el usuario se define como el organismo público Onemi, que corresponde al Centro de Alerta Temprana (CAT). Pues se dedica principalmente a informar preventivamente a la población chilena sobre desastres naturales tales como, en este caso, actividad sísmica.

De esta manera, el potencial del modelo es aprovechado al máximo, pues la información puede extenderse a lo largo de todo el país.

1.3. Problema a resolver

El problema principal es que los sismos, por naturaleza, son completamente aleatorios, es decir, es imposible predecir futura actividad sísmica.

Entonces, ¿Es posible resolver este problema?

En esencia, no, sin embargo, si bien no se puede predecir un sismo, sí se puede pronosticar. La diferencia está en que pronosticar un evento admite un porcentaje de certeza, por lo que el problema real a resolver es poder lograr una pronosticación de intensidad sísmica para distintos lugares geográficos del país en el mes posterior, reduciendo al máximo la incertidumbre.

Para lograr esto, se crea modelo que se basa en datos reales para retornar la máxima magnitud que podría tener un sismo en el próximo mes.

A modo de ejemplificar el problema, el desprevenido terremoto de 2010 dejó a más de 500 fallecidos. Si se posibilita el pronóstico de sismos intensos, las casualidades podrían reducirse en una gran cantidad.

2. Análisis de datos

En nuestro modelo se utilizarán las siguientes variables:

2.1. Fecha y hora

Esta variable permitir crear un registro histórico de los sismos registrados. A pesar de tener acceso tanto a la fecha como a la hora, esta última no resulta de gran relevancia en la construcción del modelo.

En el dataset, se encuentran registros de sismos desde el 1 de enero de 2010 hasta el 1 de noviembre de 2021. En general, se tiene un total de 618 semanas registradas.

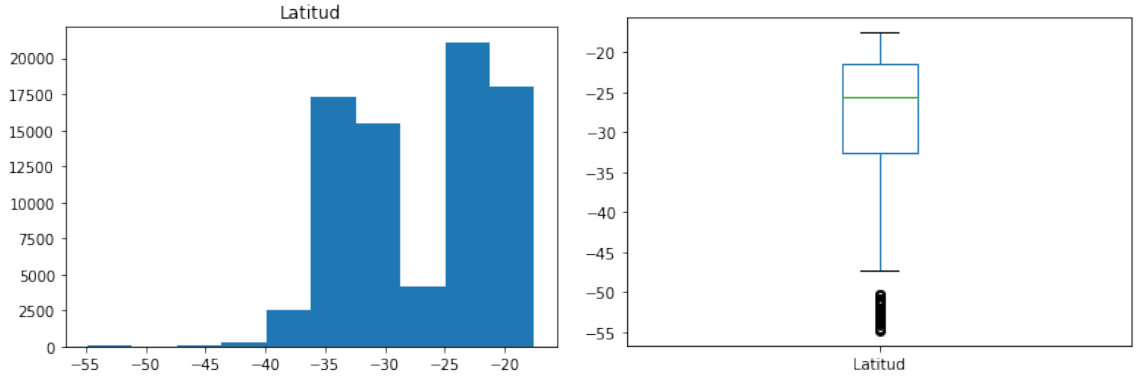
2.2. Latitud y longitud

Al verificar los valores de latitud y longitud, existen sismos muy distantes entre sí. Por ejemplo, existe un registro a 3250 km desde la costa chilena. Es por esto que se filtran los datos desde las latitudes -17 hasta -55 y desde las longitudes -64 hasta -74.

Con respecto a estos datos, se pudo obtener la siguiente información sobre la latitud y longitud de los registros sísmicos:

Tabla 1: Tabla de datos estadísticos de la latitud y longitud.

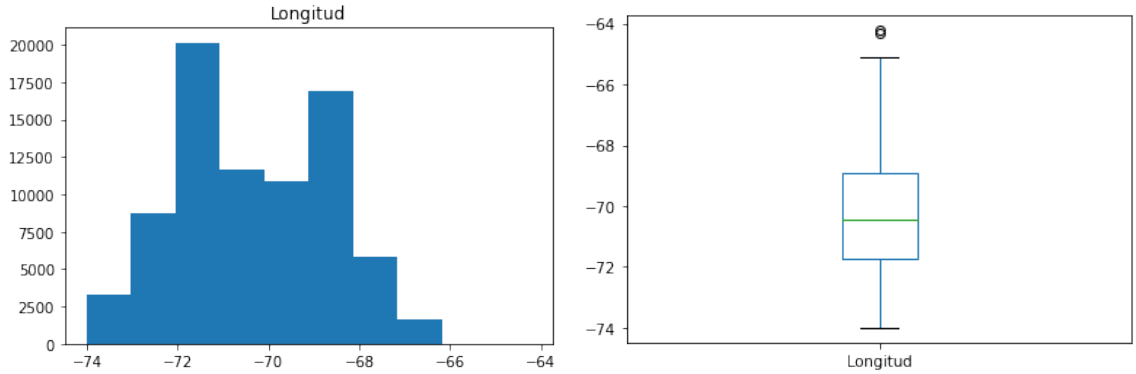
	Latitud	Longitud
Promedio	-27,0	-70,3
Mínimo	-54,8	-74,0
Máximo	-17,5	-64,2



a Histograma de la latitud

b Diagrama de caja de la latitud

Figura 1: Gráficos de las latitudes registradas para sismos en Chile desde el 1 de enero de 2010 hasta el 1 de noviembre de 2021.



a Histograma de la longitud b Diagrama de caja de la longitud

Figura 2: Gráficos de las longitudes registradas para sismos en Chile desde el 1 de enero de 2010 hasta el 1 de noviembre de 2021.

Al ver los valores máximos y mínimos, se observa que los datos varían entre -51,2 y -18,0 para latitud y entre -74,8 y -66,2 para longitud. Se verifica que los datos corresponden efectivamente al territorio chileno, al observar el siguiente mapa:

Figura 3: Mapa de Chile con las líneas de latitud y longitud.

Por otro lado, como es de interés trabajar los datos por región, estas se dividieron de forma aproximada por sus latitudes. Para ello se utilizaron los datos de la tabla 2, que corresponden a latitudes aproximadas de las fronteras de cada región

Se puede observar que la mayor cantidad de sismos se registraron en la regiones del norte de Chile, como Coquimbo, Tarapacá y Antofagasta. Les siguen principalmente regiones de la zona centro del país, como la región Metropolitana, Maule, O'Higgins y Bío Bío. Las regiones con menor actividad sísmica son las regiones del sur del país como Magallanes, Los Ríos y Los Lagos.

A partir de los datos, se evidencia una relación entre la cantidad de sismos y la latitud. En particular, se observa que al norte de Chile la actividad sísmica es mucho mayor que en el sur del país.

Tabla 2: Valores aproximados de las latitudes superior e inferior de cada región.

Región	**Latitud superior**	**Latitud inferior**
Aríca y Parinacota	-17.5	-19.1
Tarapacá	-19.1	-21.3
Antofagasta	-21.3	-25.8
Atacama	-25.8	-29.2
Coquimbo	-29.2	-32
Valparaíso	-32	-32.9
Metropolitana	-32.9	-34.1
O'Higgins	-34.1	-34.8
Maule	-34.8	-36.2
Ñuble	-36.2	-37.1
Bío Bío	-37.1	-37.8
Araucanía	-37.8	-39.4
Los Ríos	-39.4	-40.5
Los Lagos	-40.5	-43.7
Aysén	-43.7	-48.8
Magallanes	-48.8	-55.8

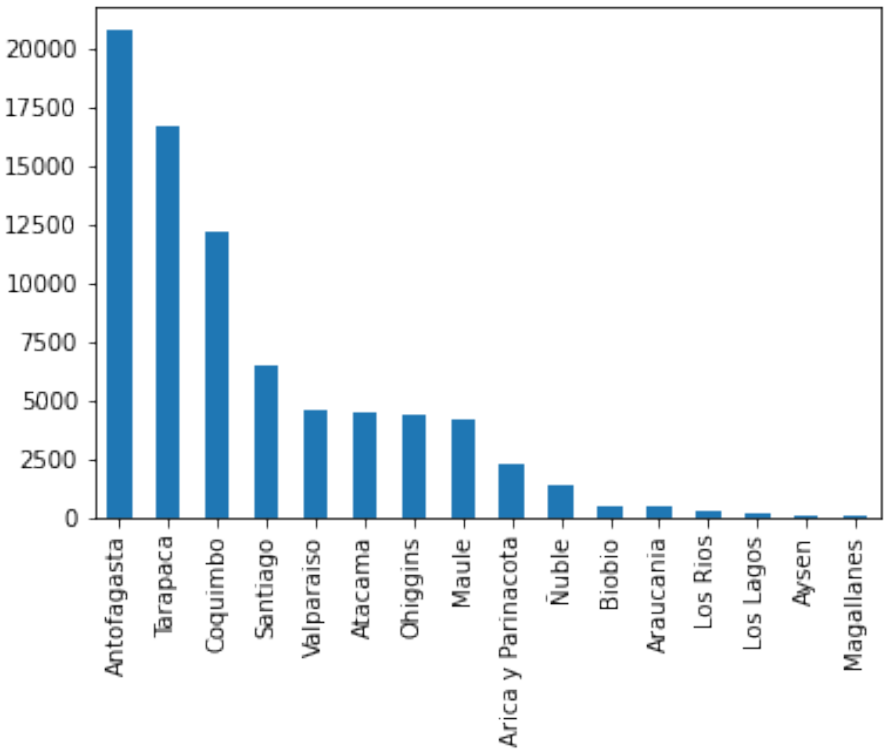


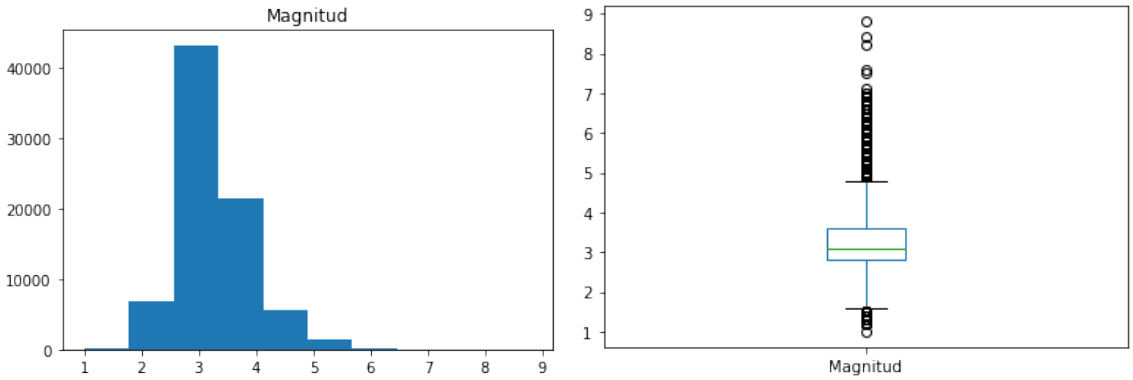
Figura 4: Cantidad de sismos registrados en las distintas regiones de Chile.

2.3. Intensidad

Para esta clasificación, se crearon tres categorías para separar a los sismos por su intensidad:

- **Leve:** Magnitud hasta 4
- **Medio:** Magnitud entre 4 y 6
- **Fuerte:** Magnitud desde 6

De los datos de los sismos registrados se obtuvo la información presentada a continuación. El sismos más fuerte registrado fue de magnitud 8,8 en la región del Ñuble. Se puede observar que la mayoría de los sismos son de intensidad baja, entre los 3 y 4. En total se registraron 8 sismos de intensidad superior a 7.



a Histograma de la magnitud

b Diagrama de caja de la magnitud

Figura 5: Gráficos de las magnitudes registradas para sismos en Chile desde el 1 de enero de 2010 hasta el 1 de noviembre de 2021.

2.4. Profundidad

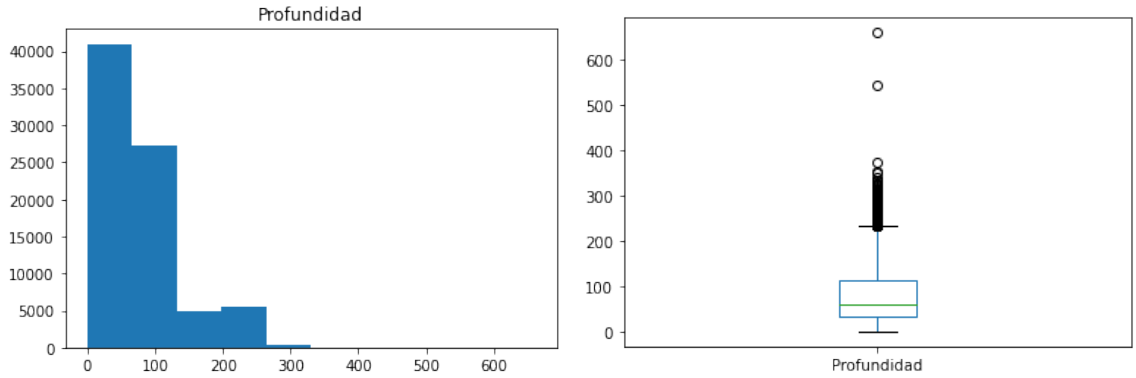
Se observa que las profundidades de los sismos se mueven entre los 2 kilómetros hasta los 297 kilómetros. Para poder estudiarlos, resulta conveniente dividir los sismos en tres categorías:

- **Superficiales:** Menor a 60 kilómetros.
- **Intermedios:** Entre 60 kilómetros y 250 kilómetros.
- **Profundos:** Más de 250 kilómetros.

Con hallazgo más importante, se pudo observar que ningún sismo fuerte fue profundo o intermedio, sino que fueron todos sismos superficiales.

Tabla 3: Distribución de profundidades por intensidad de los sismos registrados desde el 1 de enero de 2010 hasta el 1 de noviembre de 2021.

	Intensidad		
Profundidad	Leve	Medio	Terremoto
Superficial	33652	5622	89
Intermedio	33914	4856	33
Profundo	502	345	7



a Histograma de la profundidad

b Diagrama de caja de la profundidad

Figura 6: Gráficos de las profundidades registradas para sismos en Chile desde el 1 de enero de 2010 hasta el 1 de noviembre de 2021.

3. Matriz de características

Una vez analizado los datos, pasamos a definir que variables participarán en la matriz de características. Para ello, debemos considerar el problema ya definido, el cual es predecir un sismo. Para resolverlo, separaremos los datos por región y luego por mes dentro de cada región. Con esto, se usará el dataframe anterior para completar la información de cada variable. Para cada fila, se crearan las variables que serán mencionadas a continuación, las cuales representaran la matriz de características.

3.1. Variable a predecir

En primera instancia, se tenía pensado predecir la probabilidad de que ocurra un terremoto al mes siguiente. Debido a la baja cantidad de terremotos producidos en los últimos 12 años por región (máximo 4 terremotos por región, y más de la mitad de regiones sin registro de terremotos), se concluyó que la efectividad del modelo para predecir esa variable iba a ser muy baja.

Es por esto que se decidió replantear el objetivo con la intención de mejorar la efectividad del modelo buscando un nuevo enfoque. Se decidió que, para este modelo, la variable a predecir será la magnitud máxima de los sismos producidos el próximo mes en determinada región.

3.2. Variables históricas del modelo

Dentro de las variables previamente definidas en el informe de avance, se mantuvieron las variables que miden la cantidad de sismos producidos en el último mes para cada tipo de sismo (leve, medio, profundo, intermedio y superficial). Además, se mantuvo la variable que mide la magnitud máxima y mínima registrada en el último mes.

Con respecto a las variables temporales, se decidió que, en lugar de medir la cantidad meses desde el ultimo sismo fuerte, se ocupará la cantidad de días. Por otro lado, se descartó usar el tiempo desde el último sismo medio, ya que estos son tan frecuentes que no resulta relevante conocer el tiempo transcurrido desde el último.

Por otro lado, con respecto a las variables de variación, se decide medir la variación de sismos

leves y medios solamente, descartando la variación de sismos general, ya que los sismos leves y medios abarcan casi la mayoría de los registros, por lo que no es necesario calcular la variación general. Además, se decidió medir la variación de la magnitud máxima y mínima registrada. Para calcular esta medida, se decide usar el cociente entre los sismos de dos meses consecutivos, ya que produce error cuando no se tiene registro de sismos en el primer mes, debido a la división por cero. Para solucionar este problema, se definió la siguiente fórmula:

$$Var = \frac{a-b}{a+b}$$

donde a son los sismos del último mes y b los sismos del mes anterior a ese. Este resultado varía entre -1 y 1, donde -1 significa que la actividad sísmica disminuye a una completa inactividad, 1 significa que la actividad sísmica aumenta desde la completa inactividad, y 0 significa que no existe variación. En el caso de la división por cero, es decir, cuando no hay sismos en ninguno de los dos meses, significa que no hubo un aumento en la actividad sísmica, por lo que se puede definir arbitrariamente la variación como 0.

Además, se decidió agregar nuevas variables a la matriz de características. Se definieron variables de tipo espaciales, debido a la diferencia de la actividad sísmica entre regiones. En un comienzo se había pensado generar modelos para cada región, pero luego se decidió crear nuevas variables que distingan entre regiones, las cuales son:

- es_Arica
- es_Coquimbo
- es_Maule
- es_LosRios
- es_Tarapaca
- es_Valparaiso
- es_Ñuble
- es_LosLagos
- es_Antofagasta
- es_Santiago
- es_BioBio
- es_Aysen
- es_Atacama
- es_Ohiggins
- es_Araucania
- es_Magallanes

El valor de esta variable es 1 si corresponde a esa región, y 0 si no corresponde.

3.3. Resultado de la matriz de características

Una vez definidas todas la variables, se obtiene una matriz de características con dos llaves, 28 variables históricas y una variable a predecir. A continuación se presenta un ejemplo del resultado final de la matriz:

Tabla 4: Ejemplo de las primeras filas de la matriz de características del modelo

	Fecha	Region	dias_UltimoTerremoto	CantLeves	CantMedios
1	2010-03-01	Arica	59	2.0	0.0
2	2010-03-01	Tarapaca	59	26.0	7.0
3	2010-03-01	Antofagasta	22	17.0	25.0
4	2010-03-01	Atacama	59	0.0	2.0
5	2010-03-01	Coquimbo	59	15.0	6.0

	CantSuper	CantInter	CantProf	Max_UltimoMes	Min_UltimoMes
1	1.0	1.0	0.0	3.8	3.2
2	3.0	30.0	0.0	5.1	2.5
3	11.0	26.0	6.0	6.0	2.1
4	1.0	1.0	0.0	4.2	4.1
5	10.0	11.0	0.0	5.1	2.6

	Var_SismosLeves	Var_SismosMedios	Var_Max	Var_Min
1	0.333333	-1.000000	-0.126437	-0.085714
2	0.106383	0.076923	-0.019231	-0.038462
3	0.062500	0.190476	-0.032258	-0.045455
4	-1.000000	-0.200000	-0.066667	0.138889
5	-0.090909	-0.076923	-0.055556	0.000000

	es_Arica	...	es_Magallanes	Max_MesSiguiente
1	1.0	...	0.0	3.7
2	0.0	...	0.0	5.3
3	0.0	...	0.0	5.9
4	0.0	...	0.0	5.9
5	0.0	...	0.0	5.1

4. Modelos

Una vez lista la matriz de características, se definen las características y la etiqueta para crear los modelos. Las características (variable x en el programa) corresponde a todas las variables definidas en la matriz de características, a excepción de la variable región, la cual no se utiliza en el modelo al ser un dato de tipo string. La etiqueta (variable y en el programa) corresponde a la variable a predecir, o sea la variable max mes siguiente.

Posteriormente, se dividen los datos de x e y en datos de entrenamiento (x train, y train) y los datos de prueba (x test, y test), con una proporción de 75 para entrenar 25 para validar. Los datos de entrenamiento sirven para que el modelo busque patrones en las características que permiten predecir la variable deseada. Los datos de test sirven para luego cuantificar que tan eficaz fue el modelo en predecir la variable.

Con los datos ya separados, se obtienen una predicción (y pred en el programa) distinto para cada modelo. Luego se graficaron los datos y se calcularon los siguientes parámetros para ver la eficacia del modelo: error cuadrático medio (MSE), raíz de error cuadrático (RMSE) y error absoluto medio (MAE). El error absoluto medio porcentual fue omitido, debido a errores al dividir por cero, que se producen en meses sin sismos. En total, se probaron 3 modelos de machine learning para el problema, los cuales son: Decision Tree Regressor, Random Forest Regressor y Linear Regressor. A continuación, presentaremos los resultados para cada uno de ellos. Debemos tener en consideración que los resultados varían según la vez en que fue ejecutado el código.

4.1. Decision Tree Regressor

Este modelo busca crear un árbol de decisión con las características de la matriz, para luego realizar una predicción. Para crear el modelo se utilizó el comando DecisionTreeRegressor().fit(x train,y train).predict(x test) de la libreria sklearn.tree. Se obtuvieron los siguientes resultados:

- MSE: 2.2935
- RMSE: 1.51443
- MAE: 0.979642

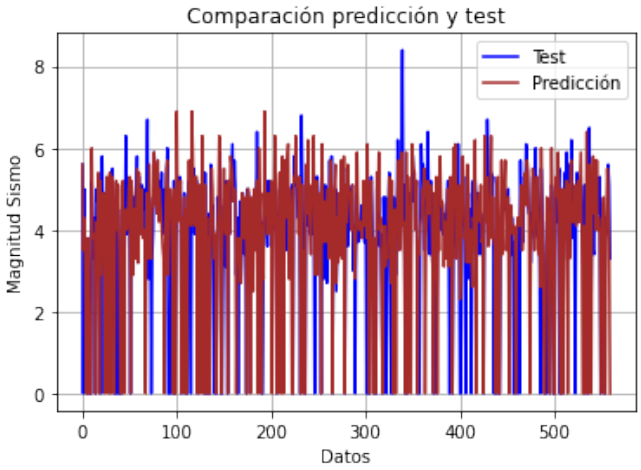


Figura 7: Comparación predicción y test en modelo Decision Tree Regressor

4.2. Random Forest Regressor

Este modelo funciona similar al anterior, con la diferencia que realiza multiples arboles de decisiones. Para crear el modelo se utilizó el comando RandomForestRegressor().fit(x train,y train).predict(x test) de la libreria sklearn.ensemble. Se obtuvieron los siguientes resultados:

- MSE: 1.4346266
- RMSE: 1.197759
- MAE: 0.8126875

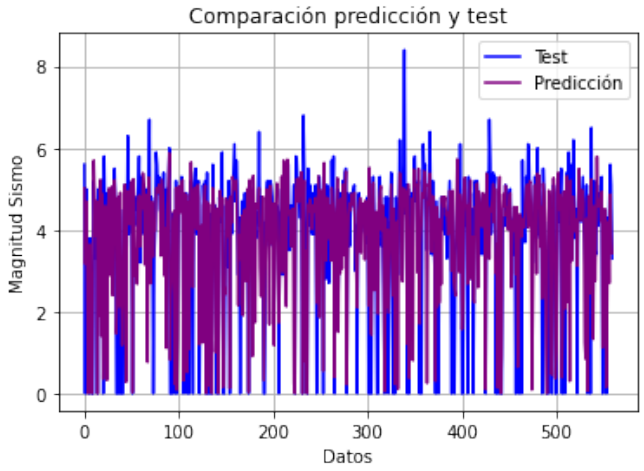


Figura 8: Comparación predicción y test en modelo Random Forest Regressor

4.3. Linear Regressor

Este modelo busca predecir una relación entre las variables x e y, realizando una aproximación lineal entre ellas. Luego, el modelo realiza una predicción con esta aproximación. Para crear el modelo se utilizó el comando LinearRegression().fit(x train,y train).predict(x test) de la librería sklearn.linearmodel. Se obtuvieron los siguientes resultados:

- MSE: 1.20255
- RMSE: 1.0966
- MAE: 0.79728

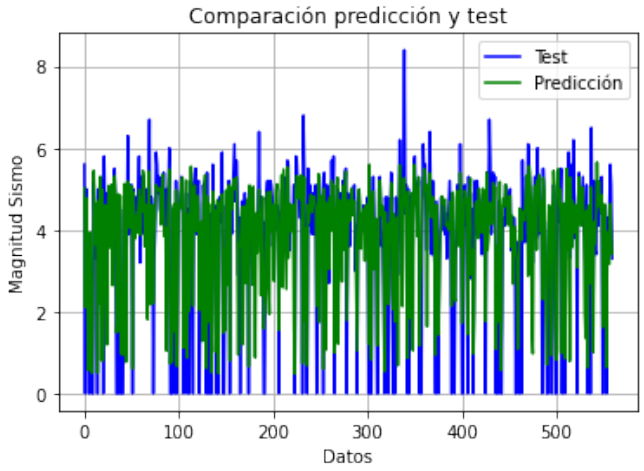


Figura 9: Comparación predicción y test en modelo Linear Regressor

4.4. Validación cruzada

La validación cruzada es un método para decidir cual es el modelo con mayor precisión. Consiste en dividir los datos de prueba en 10, para luego calcular una predicción para cada conjunto de datos. De esta manera, se evita que un modelo tenga mejor predicción que otro debido a que utilizó mejores datos para entrenar. Se debe decidir que parámetro se calculará en la predicción, el cual se utilizó el error absoluto medio (MAE) en nuestro caso. Finalmente, se calcula un promedio del MAE de los 10 conjuntos de datos y se entrega el resultado. Para ejecutarlo, se utiliza el comando cross val score de la librería sklearn.modelselection. Los resultados obtenidos fueron los siguientes:

- Decision Tree Regressor: 1.04669642
- Random Forest Regressor: 0.85501696
- Linear Regressor: 0.8178376

Como el modelo Linear Regressor obtuvo un mejor resultado (MAE más bajo), se concluye que es el más preciso. Sin embargo, Random Forest Regressor posee un resultado muy similar, por lo que también podría ser utilizado, pero con una precisión un poco menor.

5. Conclusiones

Se observa que los resultados del modelo para los tres métodos utilizados no son tan precisos como se esperaría para poder ser usados en un ámbito cotidiano. Para mejorar esto, se podría, por un lado, agregar más variables al modelo que no fueron consideradas en el proyecto, dándole más herramientas al modelo para poder predecir con mayor exactitud. Por otro lado, se podría extender el periodo de registros sísmicos de 12 años a, por ejemplo, 20 años o 30 años. Durante la experiencia se pudo observar que el modelo resulta ser más exacto en regiones con gran concentración de registros sísmicos. Otra opción que se puede realizar para mejorar el modelo es separar en varios modelos según regiones que tengan actividad sísmica similar. Cuando se realizó un modelo exclusivo para la región de Coquimbo, se pudo observar un error en la predicción menor que para el modelo con todas las regiones.

Se puede concluir que, de momento, no resulta conveniente usar el modelo generado como una herramienta para prevenir accidentes producidos por sismos, debido a la falta de exactitud en los resultados. Sin embargo, resulta prometedor esperar unos años para construir un dataset de registros sísmicos más grande que el utilizado, con el objetivo de reducir la incertidumbre de la predicción. De esta forma, se espera mejorar el modelo al tener un registro histórico más amplio.