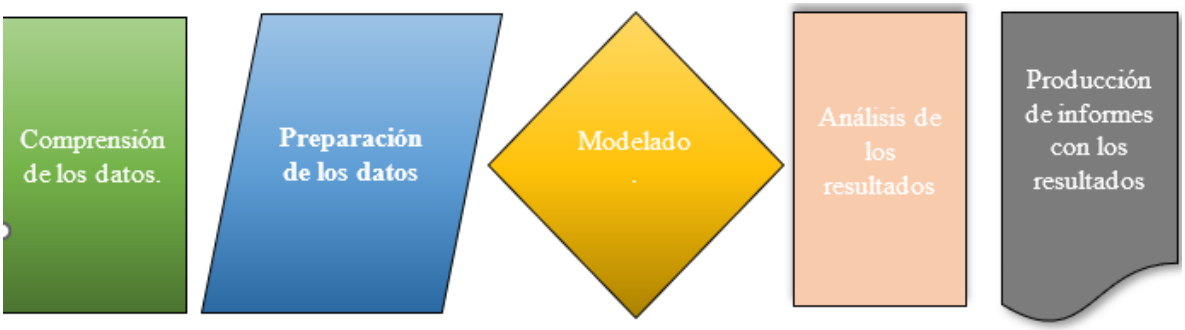
José David Pérez Rivas CI: 18.796.169

Modelar el Sistema:

El modelado predictivo de nuestro problema planteado basado en proyectar el precio de las criptomonedas, que en este caso serán las del Bitcoin y el Ethereum, será una forma de minería de datos que analizarán datos históricos de estas criptomonedas con el objetivo de identificar tendencias y luego usar esos conocimientos para predecir resultados.

Nuestro modelo para la minería de datos constara de una serie de paso:



- Etapa 1: Comprensión de los datos. Ejecución de consultas para tener muestras de los datos.
- Etapa 2: Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la minería de datos sobre ellos.
- Etapa 3: Modelado. Elección de las técnicas de modelado y ejecución de las mismas sobre los datos.
- Etapa 4: Evaluación. Análisis de los resultados obtenidos en la etapa anterior.
- Etapa 5: Distribución. Producción de informes con los resultados obtenidos en función de los objetivos y criterios. Presentación de los resultados finales.

La herramienta que vamos a utilizar para el desarrollo de este proyecto es R. Es un entorno y lenguaje de programación con un enfoque al análisis estadístico, una de las herramientas más potentes y eficientes en el mercado.

En cuanto a las técnicas a utilizar tenemos Arboles de regresión, este método para regresión y clasificación basado en árboles de decisión estratifica o segmenta el espacio del predictor en un número simple de regiones, y para obtener las predicciones se suele usar la media o moda de las observaciones de entrenamiento en la región en la que cada observación a predecir pertenece. Los árboles de decisión son simples y fáciles de interpretar, pero pueden no resultar lo suficientemente competitivos frente a otros métodos de aprendizaje supervisado en cuanto a la precisión de predicción.

Para variable dependiente cualitativa se usa un árbol de clasificación y para variable dependiente cuantitativa se usa un árbol de regresión.

Como método alternativo para pronósticos futuros utilizaremos un Modelo Random Forest.

Pasos:

1- Comprensión de los datos:

En esta fase se realiza la recolección inicial de los datos para poder establecer un primer contacto con el problema, familiarizarse con los datos y averiguar su calidad, así como identificar las relaciones

Los datos que utilizaremos en este estudio son datos reales proporcionados por la página de https://finance.yahoo.com/quote/ETH-USD/history?p=ETH-USD. Esta plataforma ya que reúne los datos reportados por las plataformas de intercambio más importantes en todo el mundo.

2- Preparación de los datos:

Para este estudio trabajaremos con las 6 variables y con un total de 2 criptomonedas: Bitcoin, Ethereum. El conjunto de datos para la criptomoneda Bitcoin fue tomado desde el 01 de Enero del año 2022 al 010 de agosto del año 2022 de la página de https://finance.yahoo.com/quote/ETH-USD/history?p=ETH-USD, la cual nos muestra un conjunto de datos diarios, para una prueba corta con la herramienta R.

La base de datos con la que se cuenta para el proyecto contiene toda la información necesaria para poder cumplir los objetivos de la minería de datos. Las 2 monedas seleccionadas tienen una gran cantidad de datos históricos.

3- Modelado:

Escoger la técnica de modelado

La técnica escogida para obtener los objetivos de minería de datos son los árboles de regresión implementada con el software R. Como método alternativo de pronóstico usaremos el método Random Forest también por R.

Creamos conjuntos de entrenamiento y prueba:

Obtenemos un subconjunto de nuestros datos que consiste en 70% del total de ellos, es decir en nuestro caso desde las fechas estipulas, para el entrenamiento El conjunto de prueba será del 30% restante a nuestros datos.

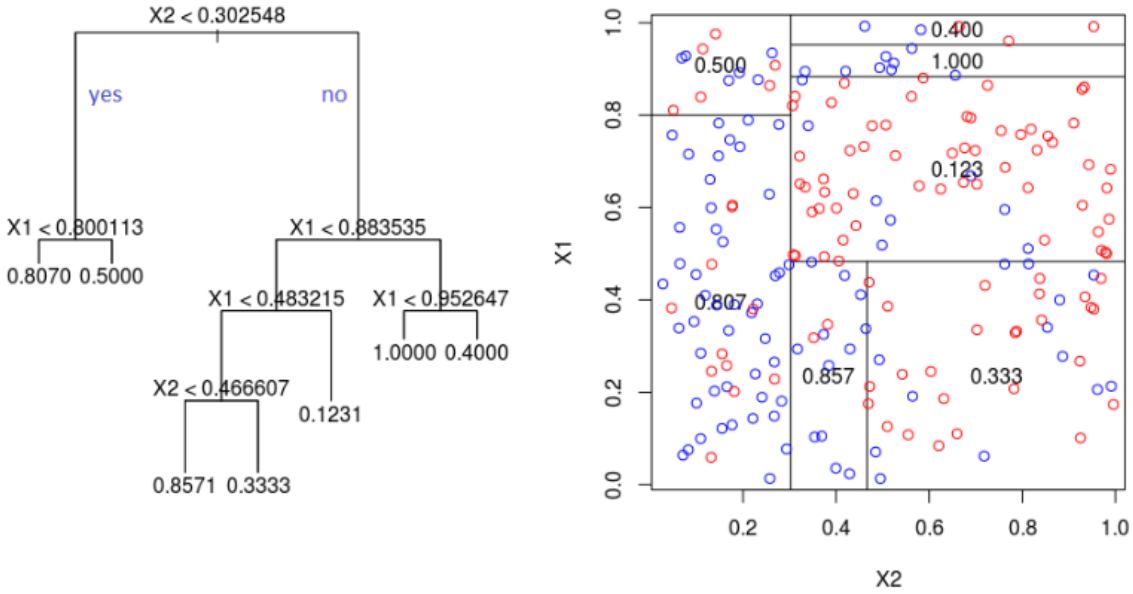
Definición del método:

“Los árboles de regresión/clasificación fueron propuestos por Leo Breiman en el libro (Breiman et al. 1984) y son árboles de decisión que tienen como objetivo predecir la variable respuesta YY en función de covariables.” https://fhernanb.github.io/libro_mod_pred/arb-de-regre.html#%C3%A1rbol-deregresi%C3%B3n

Árbol de regresión:

"Un árbol de regresión consiste en hacer preguntas de tipo ¿xk≤c? para cada una de las covariables, de esta forma el espacio de las covariables es divido en hiperrectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado Y." https://fhernanb.github.io/libro_mod_pred/arb-deregre.html#%C3%A1rbol-de-regresi%C3%B3n

"En la siguiente figura se ilustra el árbol en el lado izquierdo y la partición del espacio en el lado derecho. La partición del espacio se hace de manera repetitiva para encontrar las variables y los valores de corte c de tal manera que se minimice la función de costos $\sum (Yi - Y)$ $i=n$ $i=1$ " https://fhernanb.github.io/libro_mod_pred/arb-deregre.html#%C3%A1rbol-de-regresi%C3%B3n



"Los pasos para realizar la partición del espacio son:

1. Dado un conjunto de covariables (características), encontrar la covariable que permita predecir mejor la variable respuesta.

2. Encontrar el punto de corte c sobre esa covariable que permita predecir mejor la variable respuesta.

3. Repetir los pasos anteriores hasta que se alcance el criterio de parada.

Algunas de las ventajas de los árboles de regresión son:

- Fácil de entender e interpretar
- Requiere poca preparación de los datos.
- Las covariables pueden ser cualitativas o cuantitativas.
- No exige supuestos distribucionales.

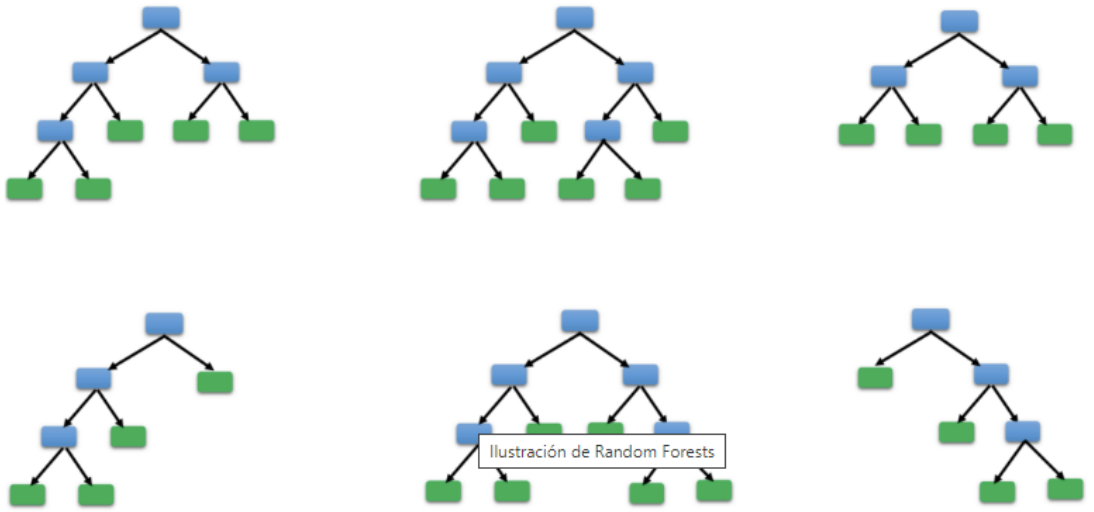
Para explicaciones más detalladas sobre las técnicas basadas en árboles recomendamos consultar el capítulo 8 de (James et al. 2013). Se recomienda también ver este video con una explicación sencilla sobre árboles" https://fhernanb.github.io/libro_mod_pred/arb-de-regre.html#%C3%A1rbol-deregresi%C3%B3n

Método alternativo:

Random Forest

Es un algoritmo de aprendizaje supervisado que utiliza un método de aprendizaje conjunto para la regresión. El aprendizaje conjunto va ser la técnica que combina las predicciones de varios algoritmos en este caso arboles de decisión de aprendizaje automático, para hacer una predicción más precisa que un solo modelo.

"Random Forest o Bosques Aleatorios fue propuesto por (Ho 1995) y consiste en crear muchos árboles para luego usarlos en la predicción de la variable de interés. A continuación se muestra una ilustración de la técnica." https://fhernanb.github.io/libro_mod_pred/rand-forests.html



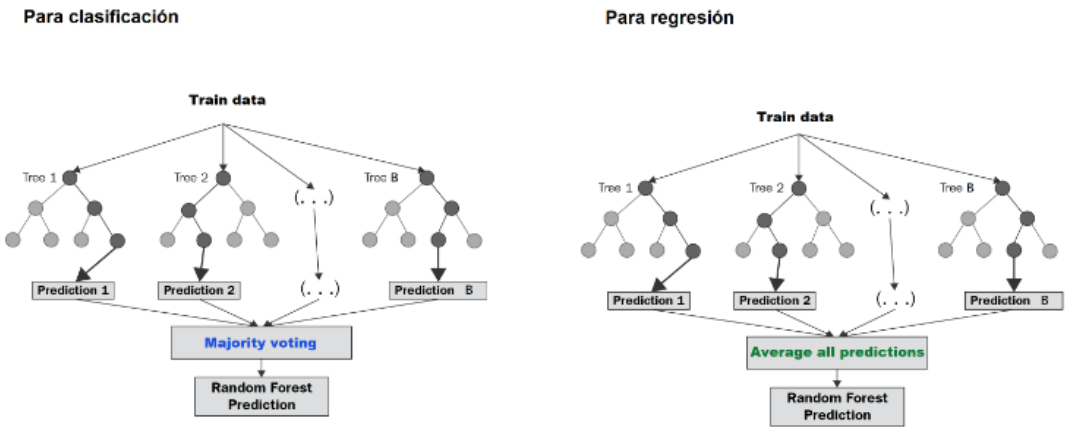
"Se inicia con un conjunto de entrenamiento que tiene n observaciones, la variable interés Y y las variables predictoras X1,X2,…,XpX1,X2,…,Xp. Luego se aplican los siguientes pasos.

1. Se construye un nuevo conjunto de entrenamiento del mismo tamaño del original usando la técnica Bootstrap. Esto se hace generando un muestreo con reemplazo y de esta forma es posible que algunas observaciones aparezcan varias veces y mientras que otras observaciones no aparezcan.

2. Se construye un árbol (de regresión o clasificación) usando en cada partición un subconjunto con k variables predictoras de las X1,X2,…,XpX1,X2,…,Xp disponibles.

3. Se repiten los pasos anteriores B veces, por lo general B=500B=500 o B=1000B=1000. De esta forma tendrán muchos árboles que luego se pueden usar para hacer predicciones de Y.

Si queremos predecir la variable Y para un caso en el cual se tienen la información (x1,x2,…,xp)T(x1,x2,…,xp)T, se toman cada uno de los B árboles creados y se predice la variable YY, de esta manera se tendrán las predicciones ^Y1,^Y2,…,^YBY^1,Y^2,…,Y^B. Luego usando estas B predicciones, se puede obtener una predicción unificada dependiendo de si el problema es de clasificación o de regresión. A continuación, una figura ilustrativa de cómo se unifican las B predicciones.” https://fhernanb.github.io/libro_mod_pred/rand-forests.html



4- Análisis de los resultados:

En esta de nuestro modelo es necesario dar seguimiento para asegurarnos que no haya errores, ir analizando las salidas generadas en este caso para los árboles de regresión y ver que tan ajustado esta nuestro modelo predictivo para la criptomoneda en la cual estemos trabajando, ya sea para Bitcoin o Ethereum.

5- Producción de informes con los resultados:

Generar un informe en términos generales sobre qué tan acertado fue nuestro modelo comparándolo y que tan ajustado está a la realidad.