INFORME PROYECTO FINAL

Título: Análisis y predicción de tsunamis por año

Objetivo: Crear un modelo que permita predecir el número de tsunamis por año

Objetivos específico:

Limpiar y analizar datos

Visualizar y comparar cómo los métodos de localización computarizados ,año 1960, han mejorado la validez de lo datos

Eliminar datos previos a 1960

Entrenar con modelo AUTO ARIMA

Evaluar el modelo

Metodología:

Exploración y preprocesamiento de datos: Se realiza un análisis exploratorio de los datos para comprender su estructura y posibles patrones. Esto puede incluir la visualización, , identificación de tendencias, estacionalidad y autocorrelaciones. Además, se pueden aplicar técnicas de preprocesamiento como la eliminación de valores atípicos o la interpolación de datos faltantes.

Selección de características: Se identifican las características relevantes que pueden influir en la serie temporal y en las predicciones futuras. Esto puede implicar la ingeniería de características, como la transformación de variables para mejorar la estacionariedad.

División de datos: Se divide el conjunto de datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba se reserva para evaluar su rendimiento.

Selección del modelo: Utilizando Auto ARIMA, se realiza una búsqueda automática de los parámetros del modelo ARIMA que mejor se ajustan a los datos de entrenamiento. Este proceso implica la exploración de diferentes combinaciones de órdenes (p, d, q) y la selección del modelo que minimice un criterio de información, como el criterio de información de Akaike (AIC) o el criterio de información bayesiano (BIC).

Entrenamiento del modelo: Se ajusta el modelo ARIMA seleccionado utilizando los datos de entrenamiento. Esto implica estimar los parámetros del modelo (coeficientes AR, coeficientes MA, etc.)

Diagnóstico del modelo: Evaluar la calidad del ajuste del modelo mediante la inspección de residuos. Verificar que los residuos se distribuyen de manera aleatoria y estén centrados en cero e identificar patrones en los residuos que puedan indicar violaciones de las suposiciones del modelo.

Validación del modelo: Se evalúa el rendimiento del modelo utilizando el conjunto de prueba. Esto implica comparar las predicciones del modelo con los valores reales y calcular métricas de rendimiento, como el error cuadrático medio (MSE) , el error absoluto medio (MAE) o coeficiente de determinación (R^2)

Conjunto de datos:

Los datos están recopilados de Kaggle

https://www.kaggle.com/datasets/noaa/seismic-waves

Hay dos dataset: sources y waves.

sources: Se refiere a los eventos que desencadenan la generación de tsunamis, como terremotos, deslizamientos de tierra submarinos, erupciones volcánicas, impactos de meteoritos u otras perturbaciones bajo el agua que causan un movimiento significativo del lecho marino.

waves: Son las ondas de agua masivas que se propagan a través del océano como resultado de un evento de tsunami. Estas olas pueden viajar grandes distancias desde su fuente original y pueden tener un impacto devastador en las áreas costeras cuando llegan a tierra firme.

Dado que "sources" se refiere a los eventos que generan tsunamis, y "waves" hace referencia a las olas de tsunami resultantes he decidido que trabajaré con el df de “sources ”ya que para mi proyecto solo necesitaré ese conjunto de datos.

Característiques generals:

EL autor de este dataset recopiló sus datos de NOAA/WDS (National Oceanic and Atmospheric Administration) Administración Nacional Oceánica y Atmosférica.

Tal y como se describe en la página, la base de datos de tsunamis de la NOAA/WDS es un listado de eventos históricos de origen de tsunamis y ubicaciones de llegada en todo el mundo desde el 2000 a.C. hasta la actualidad.

Los Centros Nacionales de Información Ambiental (NCEI) de la NOAA y el Servicio Mundial de Datos para Geofísica recopilaron y publicaron esta base de

datos de tsunamis para los centros de alerta de tsunamis, ingenieros, oceanógrafos, sismólogos y el público en general

Definició de les variables:

El dataset inicial contiene un total de 45 columnas y 2582 filas. Sin embargo después de revisar los datos y teniendo en cuenta el objetivo de mi proyecto utilizaré las siguientes columnas**:**

- SOURCE_ID: identificador único para cada registro
- YEAR: año en el que ocurrió el tsunami
- MONTH: mes del evento
- DAY: día del tsunami
- CAUSE: Código de Causa del Tsunami. Valores válidos: 0 a 11.La fuente del tsunami:
    - 0: Desconocido
    - 1:Terremoto
    - 2: Terremoto cuestionable
    - 3: Terremoto y deslizamiento de tierra
    - 4: Volcán y terremoto
    - 5: Volcán, terremoto y deslizamiento de tierra
    - 6: Volcán
    - 7: Volcán y deslizamiento de tierra
    - 8: Deslizamiento de tierra
    - 9:Meteorológico
    - 10: Explosión
    - 11: Marea astronómica
- VALIDITY: Valores válidos: -1 a 4. La validez del tsunami real se indica mediante una calificación numérica de los informes de ese evento:
    - -1: Entrada errónea
    - 0: Evento que solo causó un seiche o perturbación en un río o lago interior
    - 1: Tsunami muy dudoso
    - 2: Tsunami cuestionable
    - 3: Tsunami probable
    - 4:Tsunami definitivo
- REGION_CODE: Regional boundaries defined as follows:
    - 87 - Alaska (including Aleutian Islands)
    - 40 - Black Sea and Caspian Sea
    - 74 - Caribbean Sea
    - 78 - Central Africa
    - 84 - China, North and South Korea, Philippines, Taiwan
    - 81 - E Coast Australia, New Zealand, South Pacific Is.
    - 75 - East Coast USA and Canada, St Pierre and Miquelon
    - 76 - Gulf of Mexico

    - 80 - Hawaii, Johnston Atoll, Midway I
    - 60 - Indian Ocean (including west coast of Australia)
    - 83 - Indonesia (Pacific Ocean) and Malaysia
    - 85 - Japan
    - 86 - Kamchatka and Kuril Islands
    - 50 - Mediterranean Sea
    - 82 - New Caledonia, New Guinea, Solomon Is., Vanuatu
    - 73 - Northeast Atlantic Ocean
    - 72 - Northwest Atlantic Ocean
    - 30 - Red Sea and Persian Gulf
    - 70 - Southeast Atlantic Ocean
    - 71 - Southwest Atlantic Ocean
    - 77 - West Coast of Africa
    - 88 - West Coast of North and Central America
    - 89 - West Coast of South America
- COUNTRY: El país donde ocurrió la fuente del tsunami
- LOCATION: El país, estado, provincia o isla donde ocurrió la fuente del tsunami (por ejemplo, ingresa: Japón o Honshu).Esta es solo una ubicación geográfica aproximada. Los eventos anteriores a 1900 no fueron ubicados instrumentalmente, por lo tanto, la ubicación proporcionada se basa en la latitud y longitud de la ciudad donde ocurrieron los efectos máximos. Si hay diferentes formas de escribir el nombre de una ciudad, los nombres adicionales están entre paréntesis.

Conclusiones

El modelo no es útil para realizar la predicción de tsunamis. No se ajusta en absoluto a los datos y puede ser peor que simplemente utilizar la media como predicción

La cantidad de entradas con las que realizo el entrenamiento del modelo es muy baja. Parto de 2582 filas.Después de la limpieza y agrupación finalizo con 46 filas para entrenar el modelo

La serie tiene que ser temporal y mostrar un patrón para que el modelo pueda funcionar adecuadamente.

Para hacer este tipo de predicción también se debería tener en cuenta otros tipos de variables como por ejemplo: cantidad de desplazamiento de las placas tectónicas por años o meses.

referencia

- Kaggle**:**

https://www.kaggle.com/datasets/noaa/seismic-waves

- NOAA:

https://www.ngdc.noaa.gov/hazard/tsu_db.shtml

https://www.ngdc.noaa.gov/hazard/tsu.shtml

- Información sobre ARIMA y AUTO ARIMA:

https://www.alldatascience.com/time-series/forecasting-time-series-with-auto-arima/

https://cienciadedatos.net/documentos/py51-modelos-arima-sarimax-python