Logotipo

Descripción generada automáticamente

**Proyecto de Investigación: Proyección de Temperaturas de 5 Años por Región**

Fabián Henry Vilaxa

Prof. Patricio Soto

Inteligencia de Negocios

Iquique, 18 de octubre de 2024

Tabla de contenido

[Introducción 3](#_Toc179891685)

[Objetivos 3](#_Toc179891686)

[Descripción del Dataset 3](#_Toc179891687)

[Metodología 3](#_Toc179891688)

[Herramientas Utilizadas 3](#_Toc179891689)

[Preprocesamiento de Datos 3](#_Toc179891690)

[Visualización de Datos 3](#_Toc179891691)

[Modelos de Predicción 3](#_Toc179891692)

[Evaluación de los Modelos 3](#_Toc179891693)

[Flujo del Programa 4](#_Toc179891694)

[Resultados 4](#_Toc179891695)

[Conclusión 4](#_Toc179891696)

# Introducción

# Objetivos

Objetivos

# Descripción del Dataset

El dataset utilizado en este proyecto fue proporcionado por el profesor y contiene inicialmente 2000 registros históricos de temperatura para cinco ciudades: Tokio, París, Londres, Nueva York y Buenos Aires. Cada registro tiene las siguientes columnas:

* **Fecha**: Fecha en que se registró la temperatura, en formato 'YYYY-MM-DD'.
* **Hora**: Hora específica del registro, en formato 'HH:MM:SS'.
* **Temperatura**: Temperatura registrada en grados Celsius.
* **Ciudad**: Ciudad correspondiente al registro (Tokio, París, Londres, Nueva York, Buenos Aires).

Después de la extensión con datos ficticios, el dataset final contiene 2312 registros. Estos datos adicionales fueron generados para permitir que los modelos realizaran predicciones para periodos de hasta cinco años, cumpliendo con los requisitos del proyecto.

# Metodología

## Herramientas Utilizadas

* **Pandas**: Utilizado para la manipulación y el preprocesamiento de los datos.
* **Numpy**: Utilizado para operaciones matemáticas y generación de arrays.
* **Statsmodels**: Utilizado para la implementación de modelos ARIMA y SARIMA.
* **Scikit-Learn**: Utilizado para entrenar los modelos de Random Forest y Gradient Boosting.
* **Matplotlib y Seaborn**: Utilizados para la visualización de los resultados a través de gráficos interactivos.
* **Tkinter**: Utilizado para desarrollar la interfaz gráfica de usuario (GUI) que permite la interacción con el sistema.

## Preprocesamiento de Datos

Para obtener una visión clara de las tendencias de temperatura, se realizó un preprocesamiento de los datos de la siguiente manera:

* **Conversión de fechas**: Las columnas de fecha y hora fueron convertidas a un formato datetime, facilitando el análisis temporal.
* **Agrupación**: Los datos fueron agrupados por año, mes y ciudad, calculando la temperatura media mensual, lo que permite observar las tendencias de temperatura a lo largo del tiempo.

## Visualización de Datos

Los resultados de las proyecciones se presentan en gráficos de línea que muestran tanto las tendencias históricas como las proyecciones de temperatura para los próximos años. Se utilizaron las librerías matplotlib y seaborn para la creación de estos gráficos, y pandas para la manipulación de los datos. Además, se implementaron anotaciones interactivas en los gráficos para resaltar las temperaturas proyectadas y los datos históricos cuando el usuario pasa el cursor sobre los puntos del gráfico. A continuación se incluyen ejemplos de los gráficos generados:

* **Gráfico 1**: Tendencia histórica y proyección de temperatura para Tokio.
* **Gráfico 2**: Tendencia histórica y proyección de temperatura para París.

## Modelos de Predicción

Para la proyección de la temperatura se emplearon dos enfoques principales:

1. **Modelos de Series Temporales**:
   * **ARIMA (Autoregressive Integrated Moving Average)**: Este modelo se utilizó para capturar tanto las tendencias como los patrones estacionales presentes en los datos históricos. ARIMA es especialmente útil para series temporales con componentes lineales.
   * **SARIMA (Seasonal ARIMA)**: Además de los componentes de ARIMA, SARIMA incorpora componentes estacionales, lo cual lo hace adecuado para datos con estacionalidad recurrente. Este modelo se utilizó para reflejar patrones estacionales anuales que se observan en los datos de temperatura.
2. **Modelos de Machine Learning**:
   * **Random Forest**: Se utilizó este modelo de bosque aleatorio debido a su capacidad para manejar no linealidades complejas en los datos. El modelo fue entrenado utilizando múltiples retrasos (lags) de las temperaturas históricas para predecir los valores futuros.
   * **Gradient Boosting**: Este enfoque se utilizó para mejorar la precisión de las predicciones al combinar múltiples modelos débiles para formar un modelo más robusto. El modelo de Gradient Boosting fue entrenado con los mismos datos históricos que el Random Forest, enfocándose en la reducción del error mediante ajustes iterativos.

## Evaluación de los Modelos

Para determinar qué modelo era el más adecuado para realizar la proyección, se utilizaron dos métricas de error fundamentales:

* **Error Absoluto Medio (MAE)**: Esta métrica calcula el error promedio absoluto entre los valores reales y las predicciones, proporcionando una medida sencilla de la precisión del modelo.
* **Error Cuadrático Medio (RMSE)**: Esta métrica penaliza los errores más grandes de manera más severa, lo cual permite identificar modelos que no solo minimizan el error promedio, sino que también evitan grandes desviaciones.

Uno de los desafíos fue evaluar los modelos cuando la cantidad de datos históricos era menor que el periodo de proyección. Originalmente, el dataset contaba solo con dos años de datos históricos, mientras que se nos requería realizar proyecciones de hasta cinco años. Esto significaba que, al evaluar el **MAE** y el **RMSE** para predicciones de más de dos años, no había suficientes datos históricos para realizar una comparación directa. Para solventar este problema, se extendió el dataset utilizando datos ficticios, permitiendo a los modelos generar proyecciones y así calcular las métricas de error. La falta de suficientes datos históricos afecta la capacidad de evaluar la precisión de las proyecciones a largo plazo, ya que no existe una referencia real contra la cual comparar las predicciones para esos años adicionales.

Los resultados de las evaluaciones indicaron lo siguiente:

* **ARIMA** y **SARIMA** presentaron un buen desempeño en ciudades con patrones de temperatura estacionales bien definidos, como Londres y París.
* **Random Forest** y **Gradient Boosting** mostraron un mejor rendimiento en escenarios donde existían no linealidades más complejas, como en Tokio y Nueva York.

Los modelos con el mejor desempeño fueron seleccionados para realizar las proyecciones a cinco años. El **MAE** y el **RMSE** para cada ciudad y modelo se presentan en la interfaz, lo cual permite comparar fácilmente el rendimiento de cada uno.

# Implementación y Flujo del Programa

El programa se desarrolló con una interfaz gráfica de usuario (GUI) en Python, utilizando tkinter. El sistema permite al usuario seleccionar la ciudad, el año hasta el cual desea proyectar (de 1 a 5 años), y el modelo predictivo a utilizar. La GUI muestra un gráfico de las temperaturas históricas junto con la proyección seleccionada, lo cual permite una visualización clara e interactiva de los datos.

El proceso de implementación incluyó las siguientes funcionalidades:

* **Selección de datos**: El usuario puede cargar un archivo Excel con los datos históricos a través de la GUI.
* **Selección del modelo de predicción**: Se permite al usuario elegir entre los modelos ARIMA, SARIMA, Random Forest y Gradient Boosting.
* **Visualización de resultados**: La interfaz muestra los gráficos de temperatura con anotaciones interactivas, y presenta también métricas de error como MAE y RMSE para evaluar el rendimiento del modelo seleccionado.

# Resultados

**Gráficos Estadísticos**

* Los gráficos muestran la tendencia esperada de la temperatura en las cinco ciudades seleccionadas para los próximos cinco años.
* Se presentan tanto los datos históricos como las predicciones, facilitando la comparación visual entre ambos.
* Cada gráfico incluye una anotación interactiva que permite al usuario visualizar los valores específicos de la temperatura proyectada y los datos históricos al pasar el cursor sobre ellos.

# Conclusión

Este proyecto proporciona una herramienta para la proyección de la temperatura en las principales ciudades del mundo durante los próximos cinco años, utilizando técnicas avanzadas de predicción. Los resultados obtenidos destacan tendencias importantes que podrían ser útiles en la toma de decisiones relacionadas con el cambio climático y la gestión de recursos. La integración de diferentes enfoques de modelado permite abordar tanto patrones lineales como no lineales, ofreciendo una visión más completa de las posibles variaciones de temperatura.