

**PROYECTO FINAL**

**INTEGRANTES:**

**Carlos Alberto Serna Tuta**

**Luis Alejandro Londoño Builes**

**PROFESOR:**

**Nicolas Prieto**

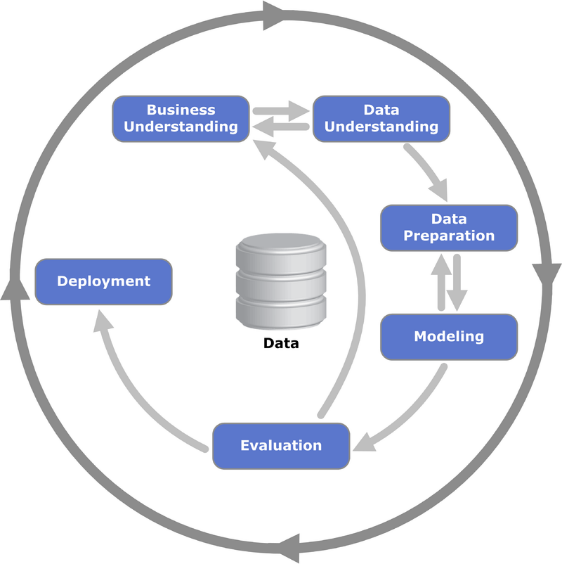
**MATERIA:**

**Análisis de datos**

**UNIVERSIDAD DE MEDELLIN**

**MEDELLIN**

**2023**

**ETAPAS DEL CRISP-DM**

1. **Entendimiento del negocio.**
2. **Entendimiento de los datos**
3. **Preparación de los datos**
4. **Modelado**
5. **Evaluación**
6. **Despliegue**
7. **Entendimiento del negocio.**

Luego de cargar los datos, podemos deducir que estos podrían ser útiles para analizar la movilidad de turistas en Medellín y para comprender mejor las tendencias de viaje.

El objetivo principal de este dataset es entender y modelar el flujo de entrada y salida de pasajeros en los países. Para esto, se cuenta con dos conjuntos de datos, uno que representa la llegada de pasajeros a el país y otro que representa la salida de pasajeros del país.

**PREGUNTAS DE NEGOCIO**

* ¿Cómo se espera que evolucione el número de llegadas de pasajeros en el futuro cercano?
* ¿Cuáles son los meses con mayor demanda de llegadas de pasajeros y cómo podemos aprovechar esta información para mejorar la planificación y promoción turística?
* ¿Cómo podemos optimizar la asignación de recursos y servicios para manejar eficientemente las llegadas de pasajeros en períodos de alta demanda?

**RIESGOS**

* Existe el riesgo de que los modelos no capturen todas las variables relevantes que afectan las llegadas de pasajeros, lo que podría llevar a predicciones inexactas.
* Los cambios significativos en el entorno o las circunstancias externas, como eventos inesperados o regulaciones de viaje, podrían alterar la precisión de las predicciones y afectar el rendimiento del modelo.
* Los errores en los datos históricos o la falta de datos actualizados podrían afectar la calidad de las predicciones y limitar la capacidad del modelo para capturar cambios recientes en las tendencias.

**CRITERIOS DE ÉXITO**

* Los criterios de éxito podrían basarse en la precisión de las predicciones del modelo en comparación con los valores reales de llegadas de pasajeros.
* Otro criterio de éxito podría ser la capacidad del modelo para identificar patrones estacionales y tendencias a largo plazo de manera efectiva.
* El modelo también podría considerarse exitoso si se utiliza como una herramienta útil para la planificación y asignación de recursos, permitiendo una gestión más eficiente de las llegadas de pasajeros.

1. **Entendimiento de los datos**

Se podría realizar un análisis a largo plazo de las tendencias y patrones de la llegada y salida de pasajeros en los diferentes países.

También se podría explorar la posibilidad de utilizar el análisis de series de tiempo para predecir la llegada y salida de pasajeros en diferentes países en el futuro, teniendo en cuenta las tendencias históricas.

1. **Preparación de los datos**

**Carga de los datos:** En el caso de nuestro proyecto se utilizaron 2 dataset, por lo tanto, se importaron ambos dataset, uno contiene los datos de llegadas de los pasajeros y el otro los datos de salida de los pasajeros.

**Análisis exploratorio de los datos**: Se llevó a cabo un análisis exploratorio para comprender la estructura y características de los datos. Esto incluyó examinar las dimensiones de los dataset, se visualizaron las primeras filas, se revisaron los tipos de datos de cada columna, se identificaron valores nulos y se observó las estadísticas descriptivas básicas.

**Manipulación de datos:** Se realizaron diversas manipulaciones en los datos para prepararlos adecuadamente para su análisis y modelado. Algunas de las manipulaciones comunes incluyeron:

1. **Modelado**

Se realizó un análisis exploratorio de los datos para comprender las características de la serie temporal y detectar posibles patrones estacionales, tendencias y estacionariedad.

Se dividió el conjunto de datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utilizó para ajustar los parámetros del modelo y el conjunto de prueba se utilizó para evaluar el rendimiento del modelo.

Se exploraron diferentes configuraciones de parámetros para el modelo ARIMA y SARIMA, incluyendo el orden de autoregresión (p), el orden de diferenciación (d) y el orden de la media móvil (q), así como los parámetros estacionales (P, D, Q, S).

Se evaluó el rendimiento del modelo utilizando métricas como el error cuadrático medio (MSE) y el error absoluto medio (MAE) en el conjunto de prueba. Estas métricas permitieron comparar las predicciones del modelo con los valores reales y determinar la precisión del modelo.

1. **Evaluación**

Se realizó un análisis de los residuos del modelo elegido para verificar si se cumplían las suposiciones de ruido blanco. Esto es importante para validar la calidad del modelo.

Se generaron predicciones utilizando el modelo ajustado en el conjunto de prueba. Para las evaluaciones "one step ahead", se generaron predicciones de un paso a futuro, es decir, se pronosticó el siguiente valor de la serie temporal en cada paso. Para las evaluaciones "multi-step ahead", se generaron predicciones múltiples pasos a futuro.

Se calcularon las métricas de evaluación para comparar las predicciones del modelo con los valores reales en el conjunto de prueba. Las métricas utilizadas comúnmente incluyen el Error Absoluto Medio (MAE) y el Error Cuadrático Medio (MSE). Estas métricas permiten medir la discrepancia entre las predicciones y los valores reales, proporcionando una medida de la precisión del modelo.

1. **Despliegue**

Se entrenó un modelo ARIMA utilizando los datos de la columna valor en ambos dataset.

Se obtuvieron las predicciones para un horizonte futuro específico utilizando el método get\_forecast().

Se extrajeron las predicciones medias y los intervalos de confianza correspondientes.

Se realizó una gráfica que mostraba los datos reales, las predicciones en la muestra, las predicciones a futuro y los intervalos de confianza.

Este proceso permite visualizar las predicciones a futuro junto con los datos reales y los intervalos de confianza correspondientes. El gráfico resultante proporciona una representación visual del rendimiento del modelo y ayuda a evaluar la precisión de las predicciones.