ANÁLISIS Y MODELADO LA'S BIKE SHARE

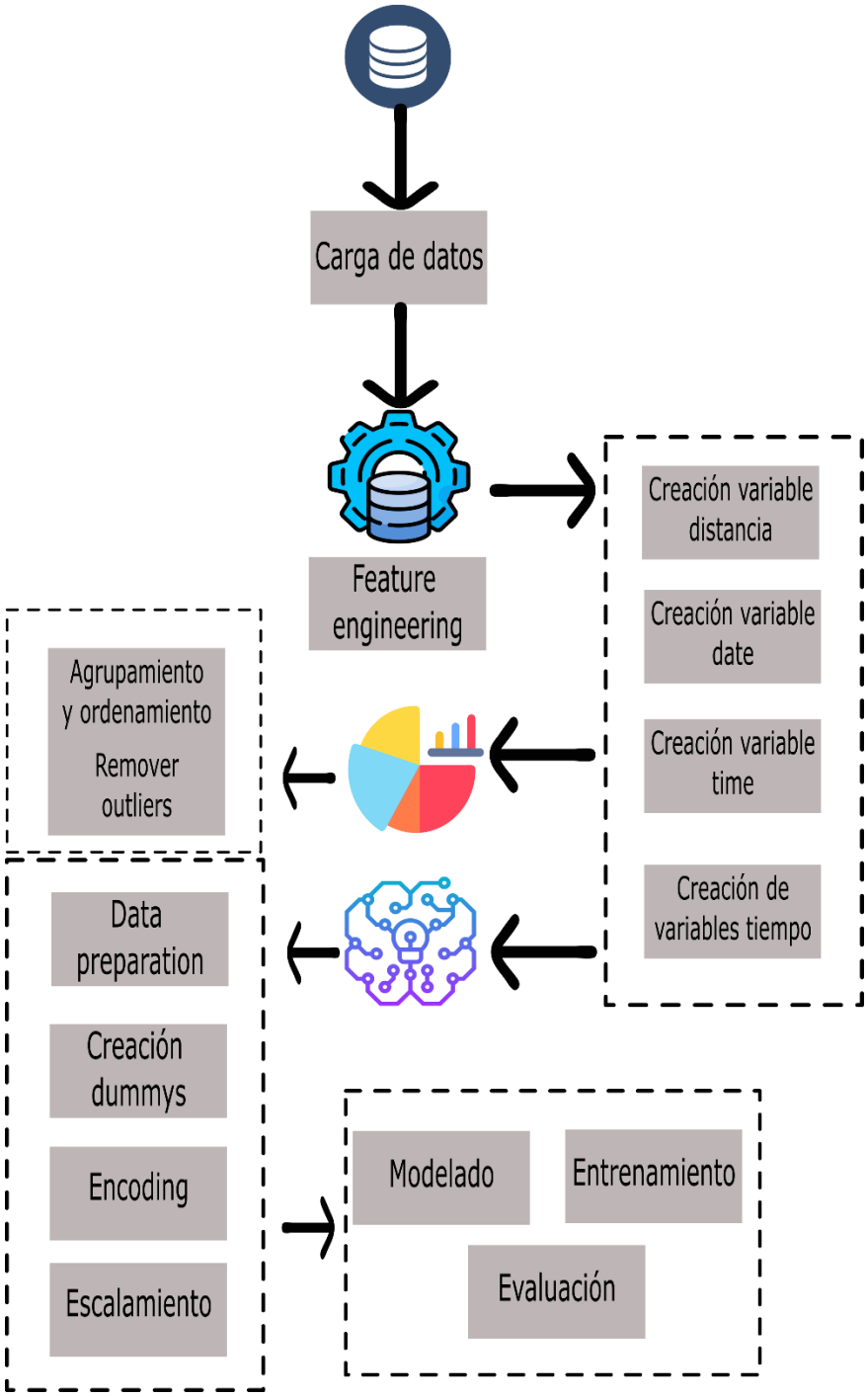
Reporte y resultados del análisis para Arkon data.

DESCRIPCIÓN BREVE

Análisis correspondiente al sistema compartido de transporte mediante bicicleta utilizado en Los Angeles. En el presente documento se presenta el resultado del ejercicio realizado para la posición de Machine learning engineer.

JOSE LUIS CERVANTES VARELA
Data scientist

Proceso realizado para el análisis

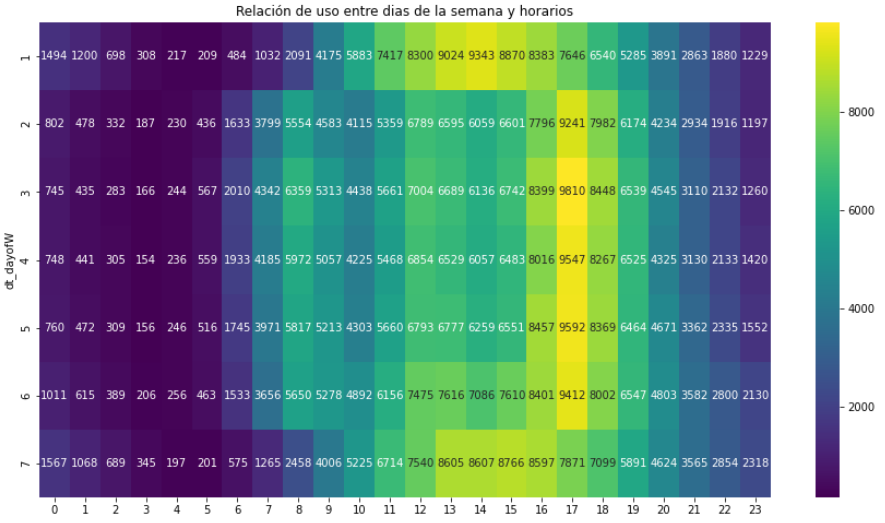


RESULTADOS DE LOS ANÁLISIS

- **¿Cuál es el comportamiento de los usuarios considerando en cada día de la semana?**

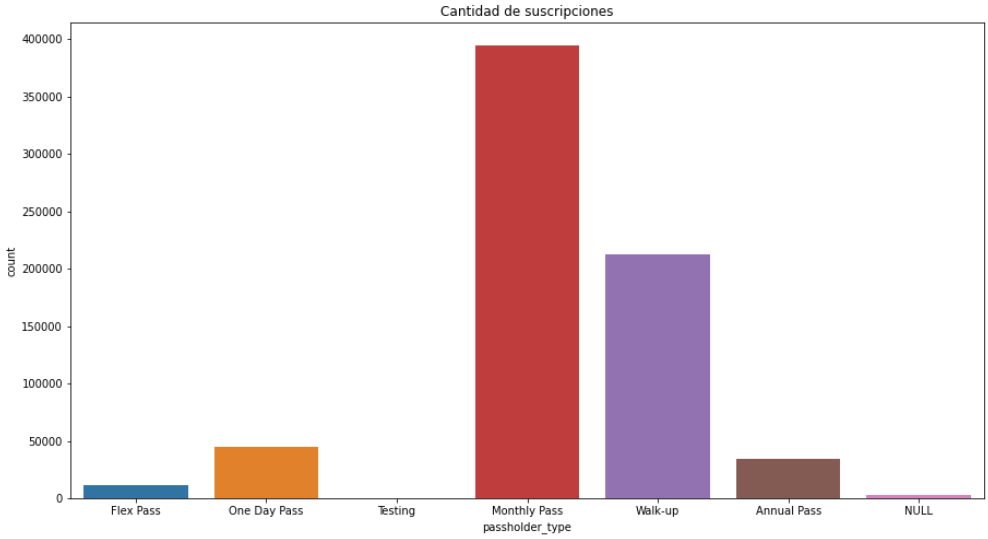
Conclusiones: El servicio suele utilizarse más entre los meses de Junio y Noviembre, siendo que el día de la semana que más se utiliza es el Lunes en el Mes de Octubre.

¿Cómo se comportan los usuarios en los diferentes horarios del día?



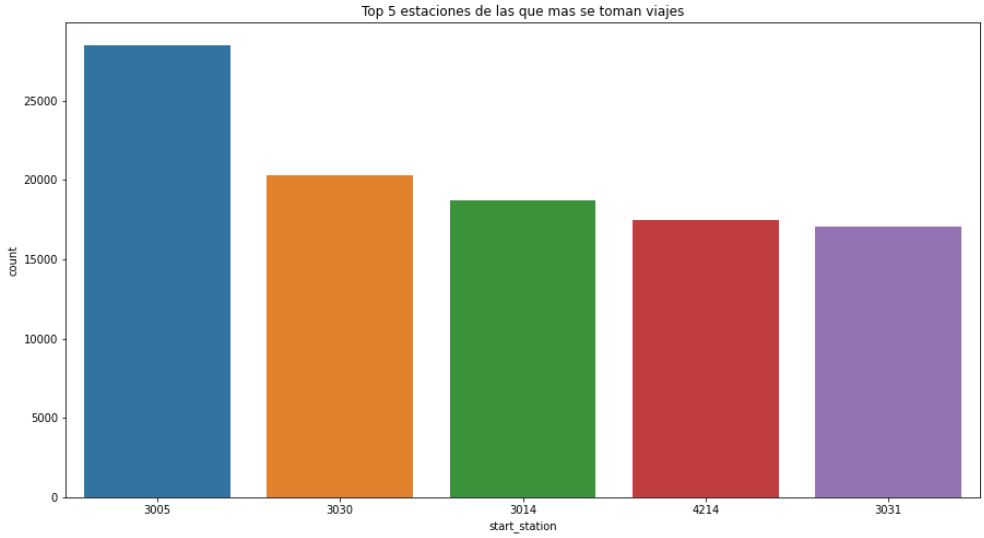
Conclusiones: El servicio suele utilizarse entre los horarios de las 6am y las 19pm. Pero la hora de mas usuarios es a las 17pm.

- **¿Cuál es la suscripción más solicitada?**

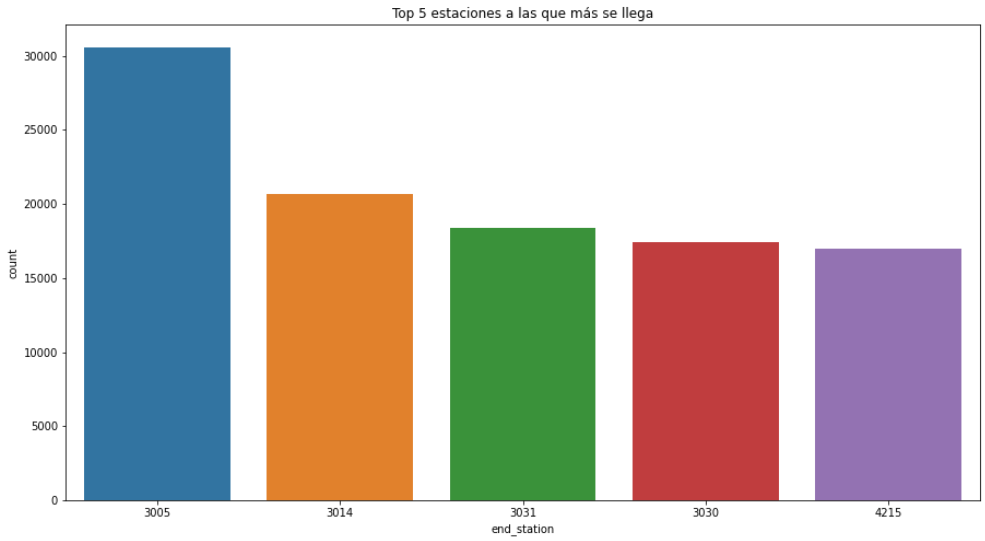


Conclusiones: El pase al que más usuarios están suscritos es el Monthly pass, seguido por el de walk up. Aún cuando el Annual pass es más económico no muchos usuarios se suscriben a el.

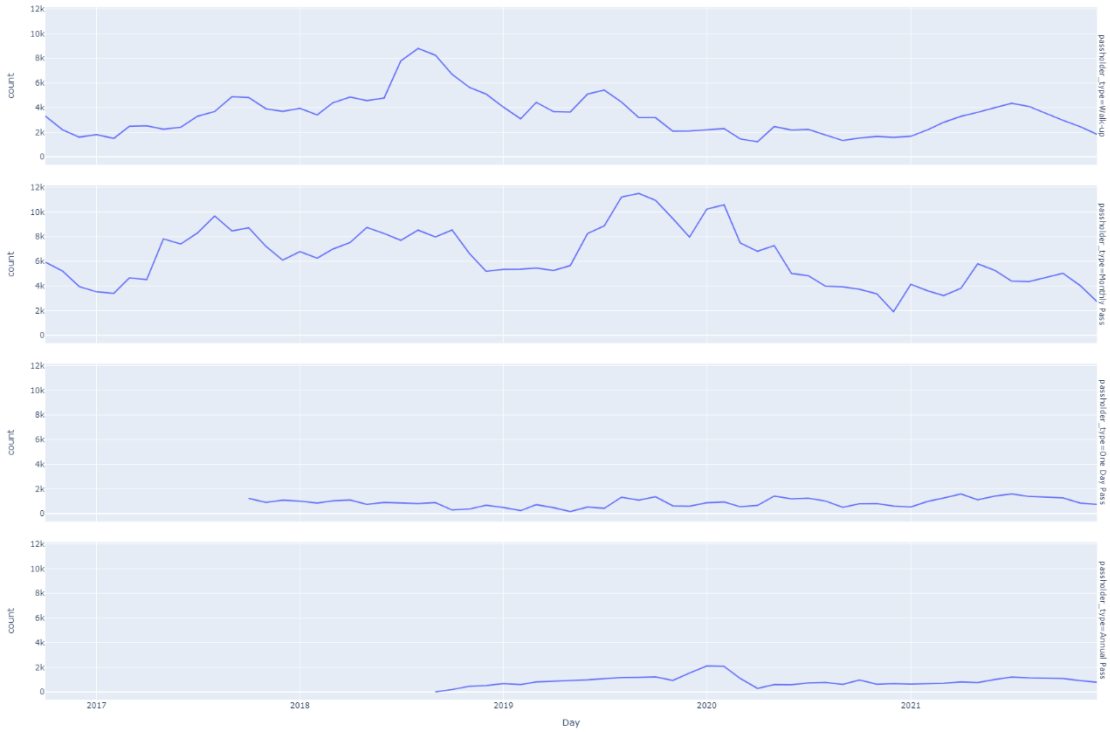
- **¿Cuáles son las estaciones de las que más viajes se ha realizado?**



- **¿Cuáles son las estaciones a las que más viajes se hacen?**

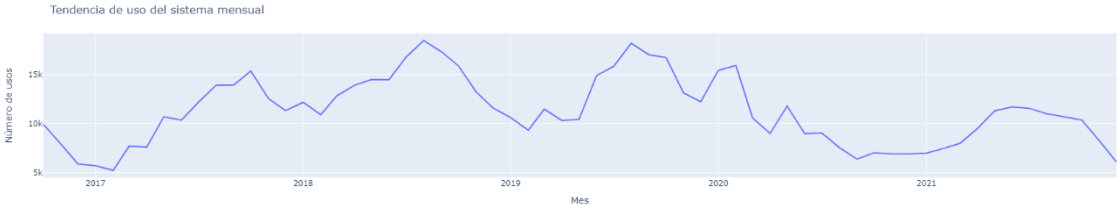


- **¿Qué tendencia siguen las suscripciones?**



Conclusiones: Como se vio en un gráfico anterior los servicios mas utilizados son el monthly pass y el walk up. Lamentablemente se han ido suscribiendo menos usuarios a estos servicios. El covid puede haber afectado a este servicio.

- **¿Cuál es la tendencia de uso del servicio?**



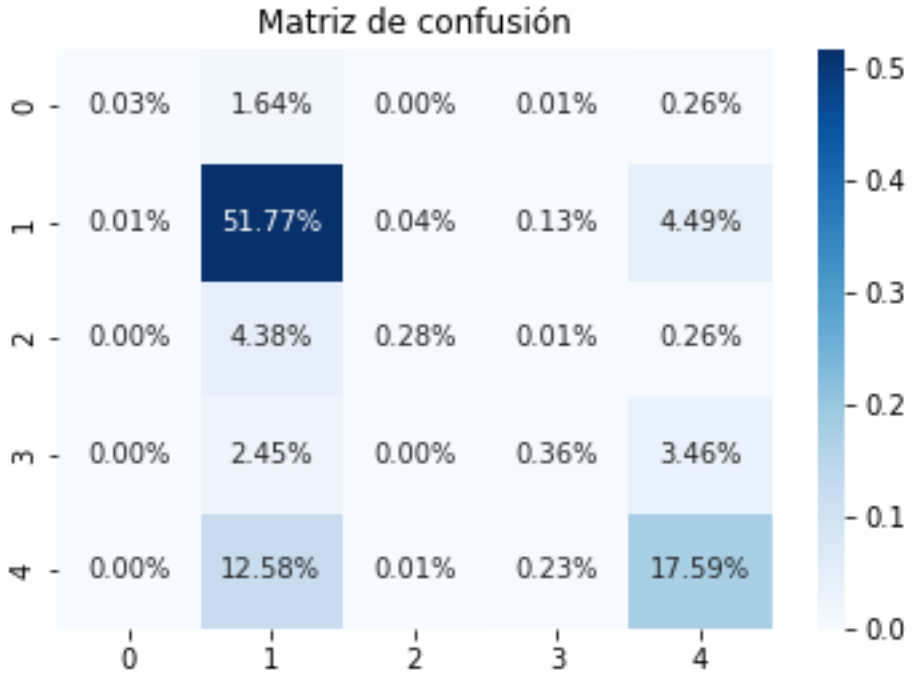
Conclusiones: El grafico de uso del servicio general confirma también el descenso en el uso del servicio. Esta tendencia decreciente comenzó en Febrero del 2020, justo en los comienzos de la pandemia.

RESULTADOS DEL MODELO

Para el entrenamiento del modelo para clasificar el tipo de pase, se selecciono un modelo random forest, el cual contiene los siguientes parámetros:

- Estimadores = 1000
- Depth = 12
- Features = 9

Matriz de correlación



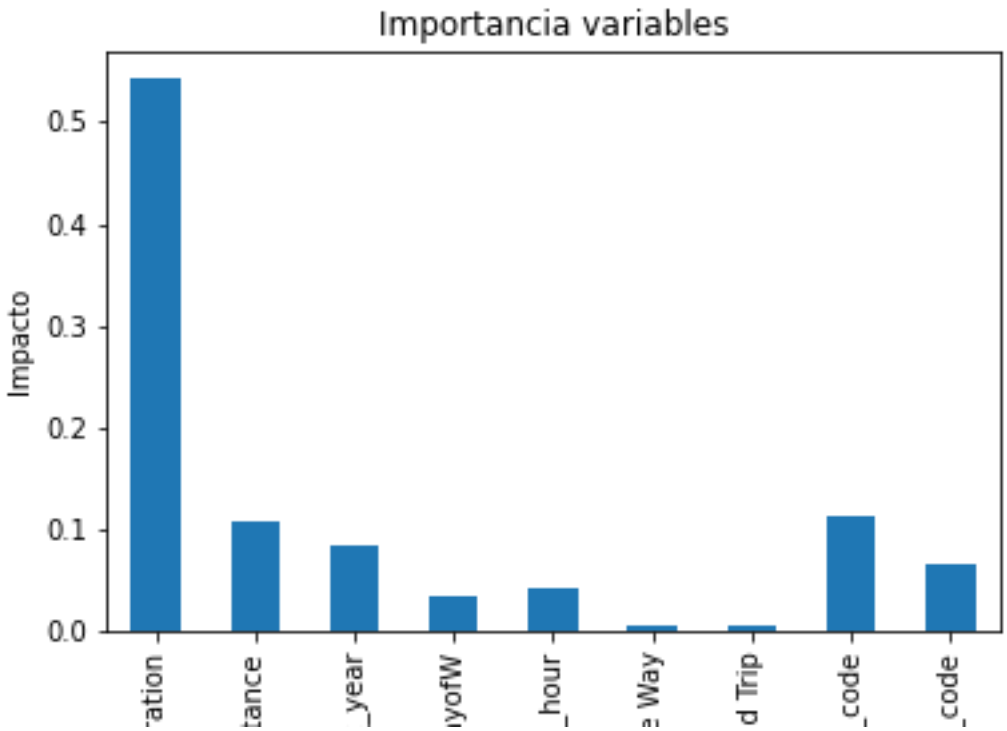
Conclusiones: En la matriz de correlación podemos observar que el modelo predice de mejor manera la clase 1, correspondiente a Monthly pass, lo cual se entiende por la cantidad de ejemplos que se tiene, no sucede así con los demás pases, observando que en la clase 4 suele confundirse al predecir con la clase 3 o 1.

Métricas

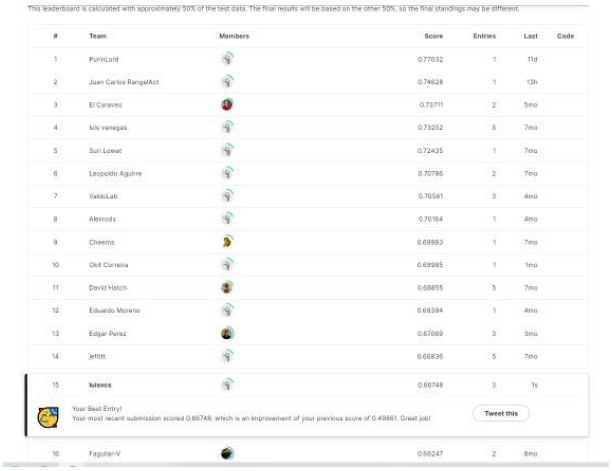
El accuracy es: 0.7003485714285714

Conclusiones: El accuracy es únicamente del 0.7, es decir que solamente predice correctamente aproximadamente el 70% de los ejemplos de test, no es lo óptimo y debería buscar mejorarse.

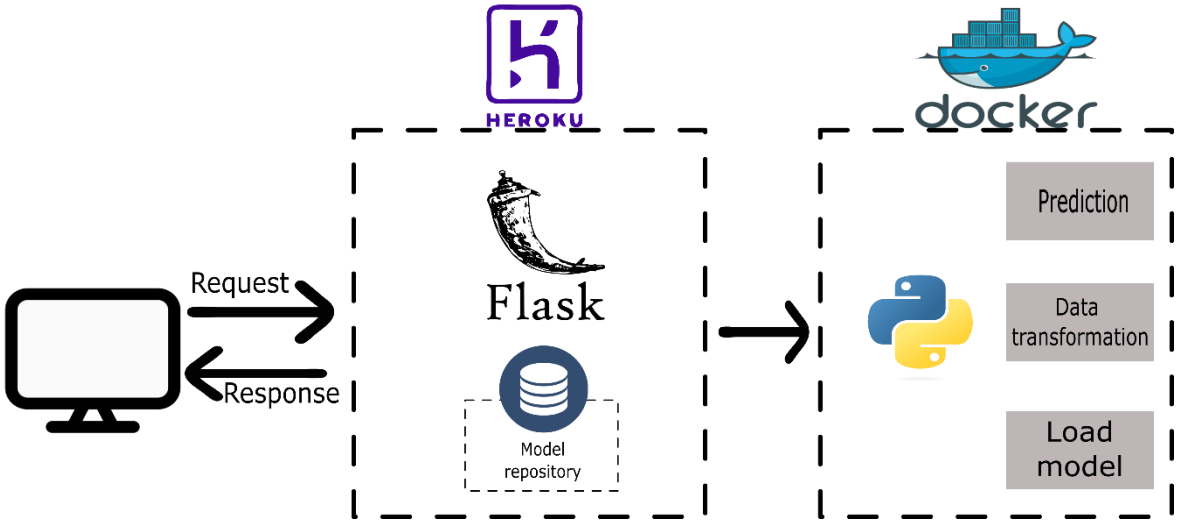
Feature importance



Conclusiones: La gráfica referente a la importancia de las variables, nos demuestra que la que mayor peso tiene al momento de realizar la predicción es la de duration, seguida por la de distance y year. Trip rout category tiene una importancia menor por lo que esas dos columnas podrían omitirse en un segundo entrenamiento.



Arquitectura despliegue



Próximos pasos y conclusiones generales

- Reducir variables para realizar la predicción.
- Utilizar otros modelos predictivos.
- No se consideraron algunas variables como las coordenadas para el modelo, se podría analizar la correlación de estas con el tipo de pase.
- Para las estaciones se hizo uso de una codificación numérica para cada estación, se podría probar una por One Hot encoding para contemplarlas como categorías.
- El uso del sistema ha ido a la baja, además de que algunos pases, como el anual casi no se solicita, un factor puede ser el clima que se vive en apocas de invierno, podría considerarse crear planes semestrales.