

Proyecto final

Predicción de lluvia



M. Tomás Martínez

Comisión: 42390

Profesor: Jorge Ruiz.

Tutor: Aldana Ruscitti

Kangaroo Crop

Nos encontramos al mando del área de Data de una de las mejores empresas en el sector agropecuario de Australia "Kangaroo Crops" en donde su mayores productos son el trigo, cebada, caña de azúcar y frutas.

Por lo tanto un estudio de la lluvia puede traer consigo numerosos beneficios para la empresa



Objetivo del Proyecto

Obtener los datos más influyentes, para así manejarlos y manipularlos para poder encontrar patrones, con el motivo de predecir la lluvias siguientes o de algún día determinado.

Crear un modelo que prediga la lluvia

Hipótesis

01.

¿Cuáles son los principales elementos para determinar si llueve o no?

¿En las ciudades ubicadas más al sur aumenta la probabilidad de que haya lluvias más abundantes

02.

¿En qué estación del año es más propenso de que llueve más?

¿Cuáles son los días que en promedio llueve más, los que se encuentran por debajo del promedio de la temperatura o los que están por encima?

03.

¿Cómo ha evolucionado la lluvia a lo largo del tiempo?

¿En promedio ha llovido más en los últimos 5 años?

MetaData

La base de datos de la empresa cuenta con aproximadamente 10 años de observaciones meteorológicas diarias de muchos lugares de Australia. En ellas podemos apreciar 145460 Filas y 23 columnas de datos.

Como parte de una base de datos que se especifica en la Lluvia, se encuentran en ellas las siguientes categorías:

- Fecha: Registro del día/mes/año.
- Lugar: Ciudades
- Lluvia: Cantidad de Lluvia en milímetro (mm)
- MinTemp: Grados (°C)
- MaxTemp: Grados (°C)
- Evaporacion: Cantidad en milímetros por día (mm/día)
- Luz Solar: Cantidad de vatio por metro cuadrado (W/m²)
- Direccion Viento: Dirección del Viento
- Veloc Viento: Velocidad en Km/h
- Direc Viento 9am: Dirección del Viento
- Humedad 9am: Cantidad en g/m³
- Humedad 3pm: Cantidad en g/m³
- Presion 9am: Pascal (Pa)
- Presion 3m: Pascal (Pa)
- Nubes 9am: Cantidad de Octas (0.0/8.0)
- Nubes 3am: Cantidad de Octas (0.0/8.0)
- Temp 9am: Grados (°C)
- Temp 3pm: Grados (°C)
- Lluvia Hoy: No o Yes
- Lluvia Mañana: No o Yes

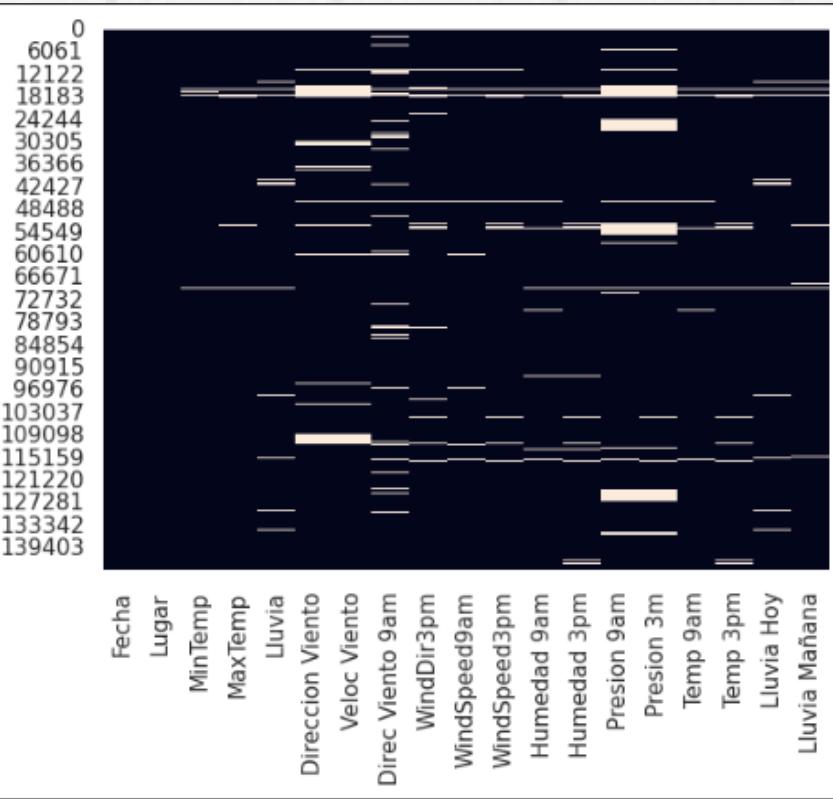


Proceso de Limpieza de Datos

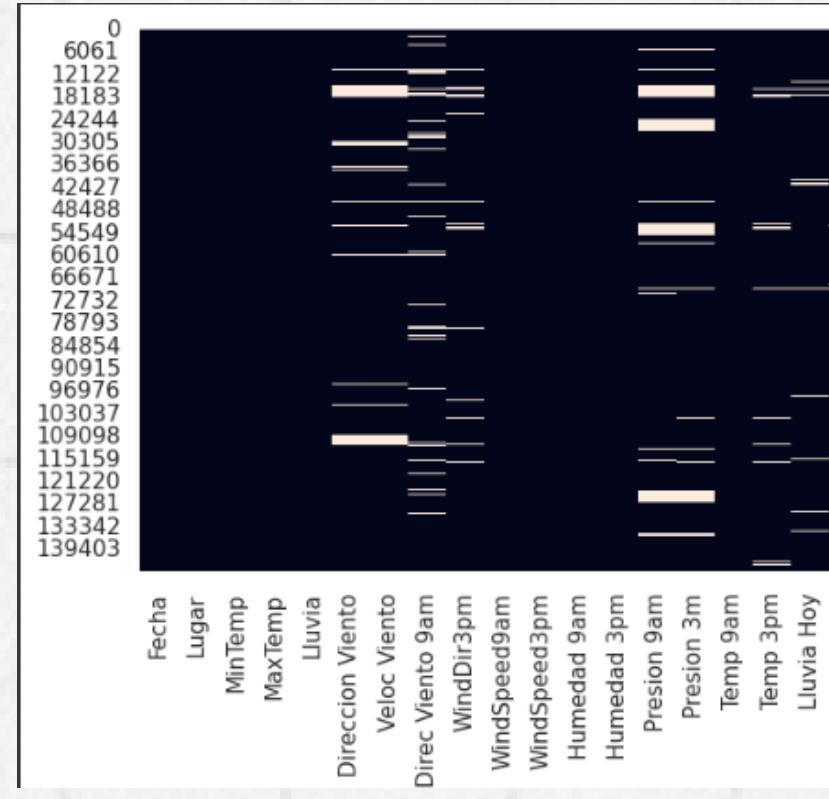
145460 rows x 23 columnas

130171 rows x 18 columnas

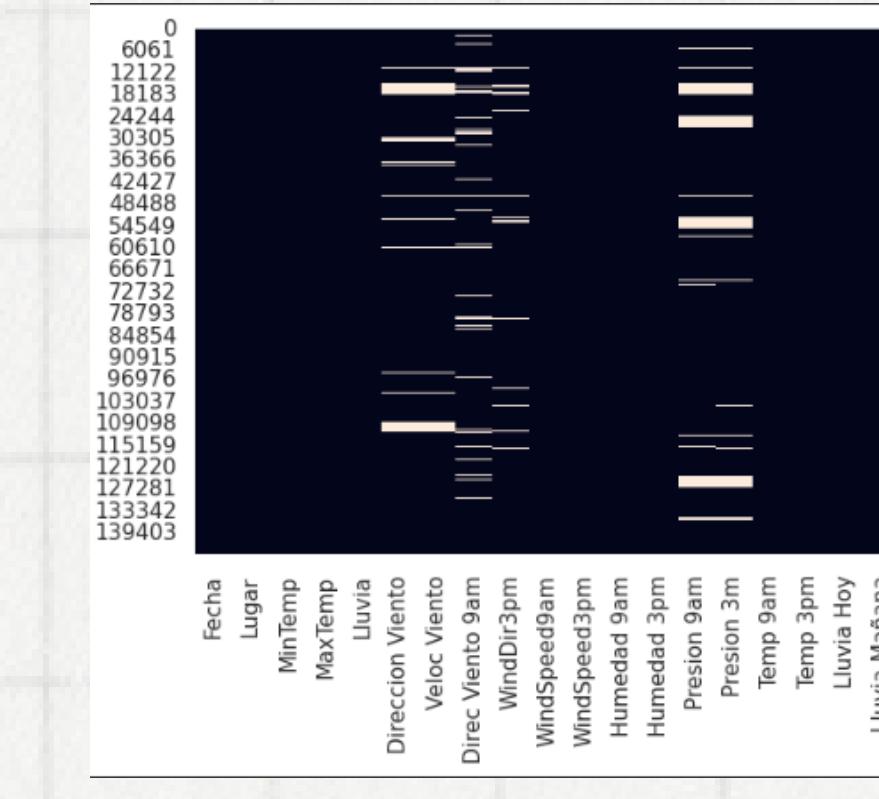
01



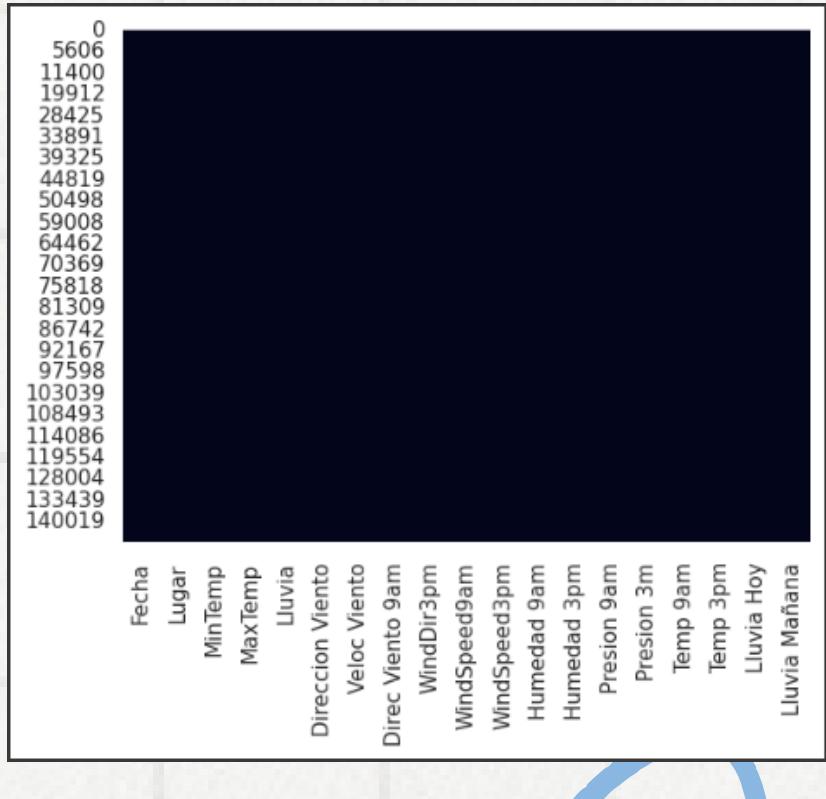
02



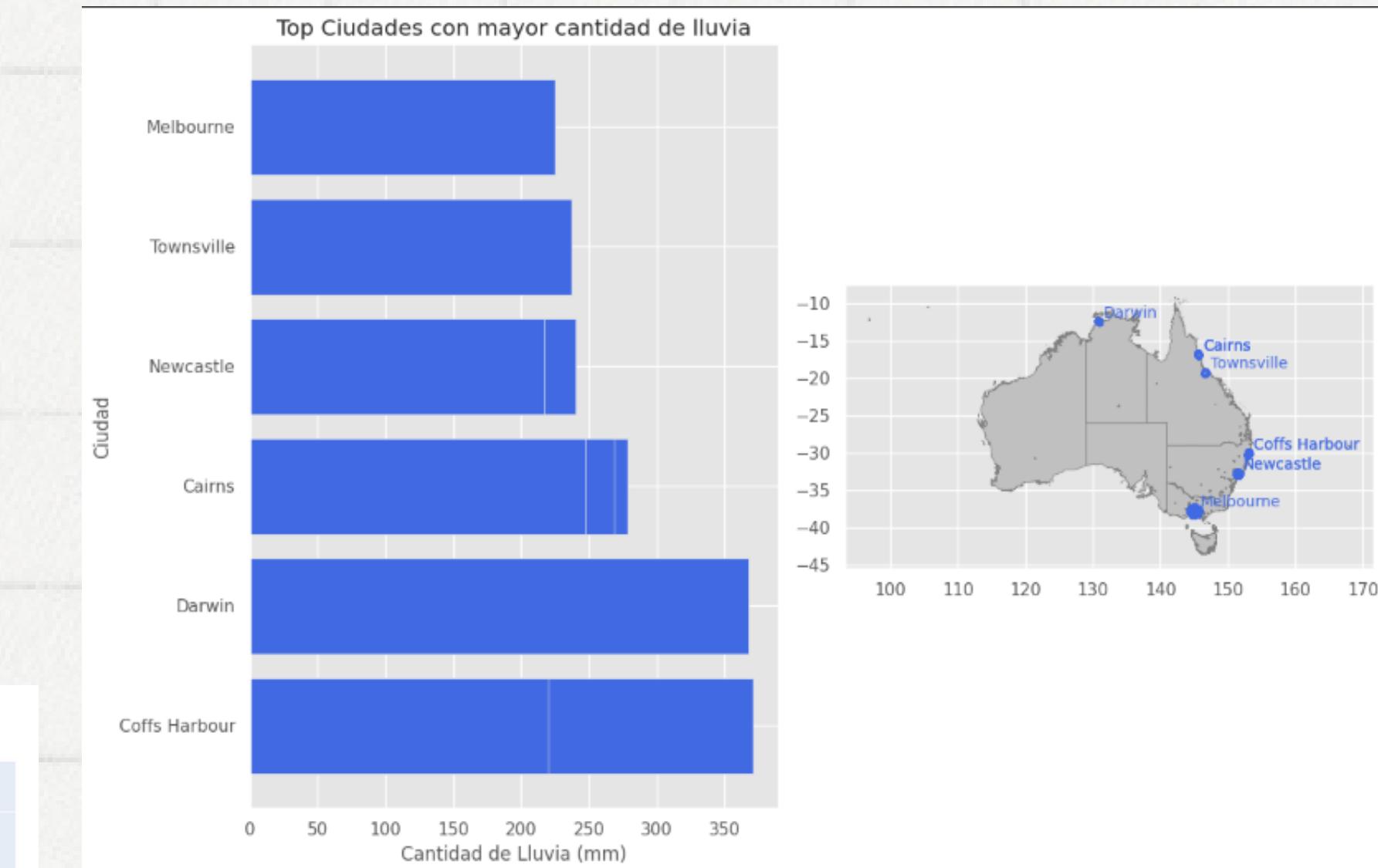
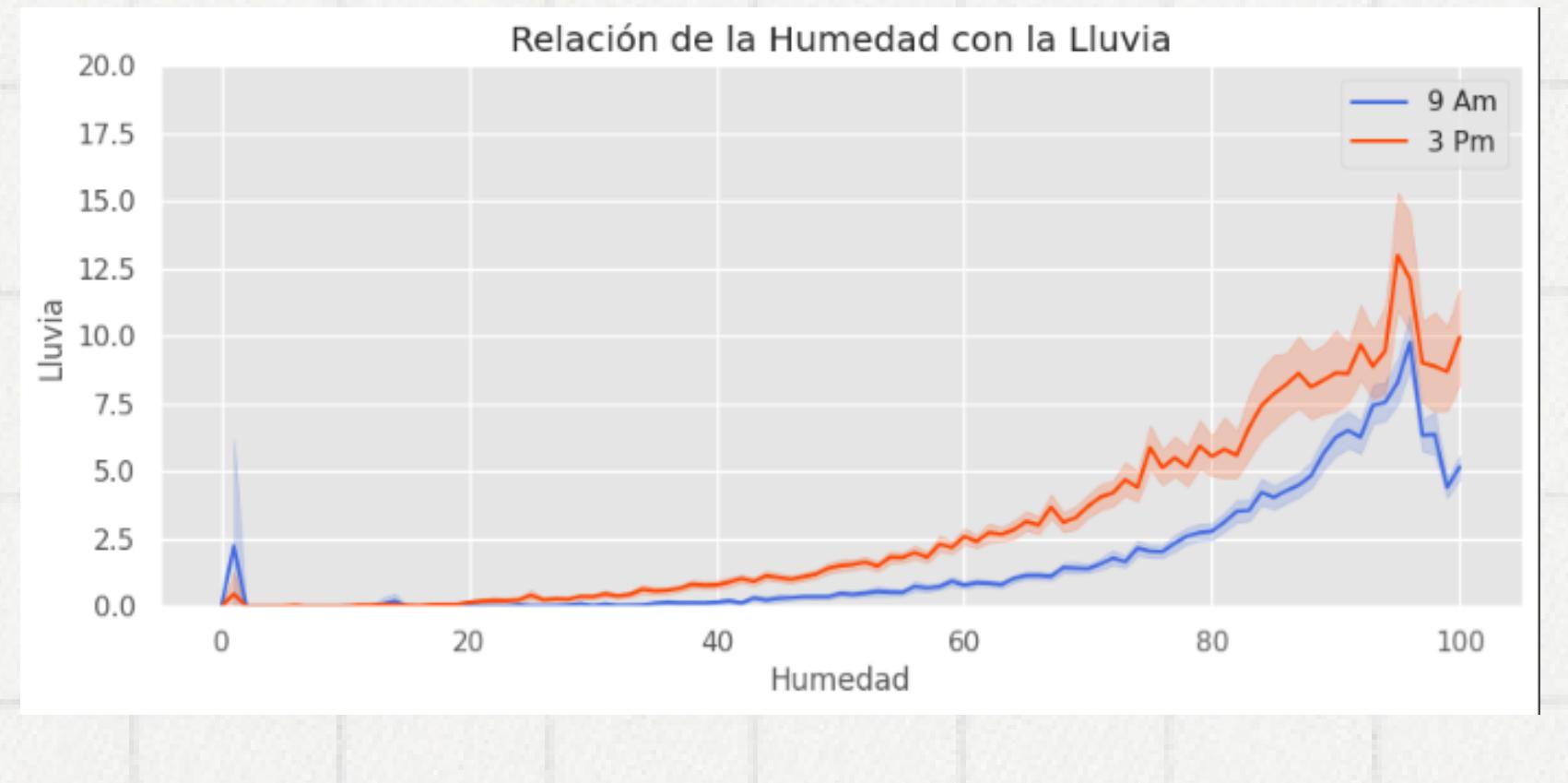
03

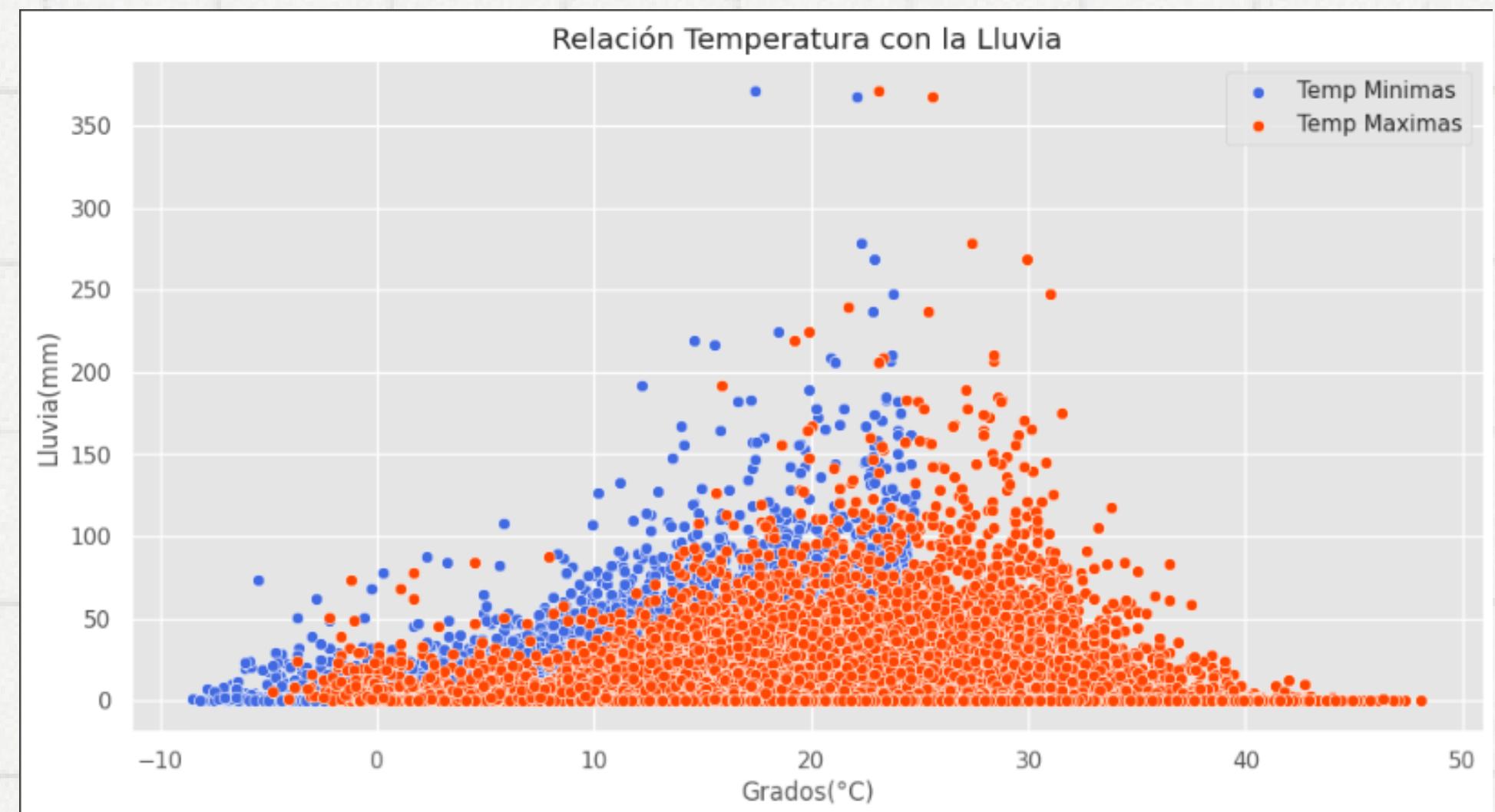
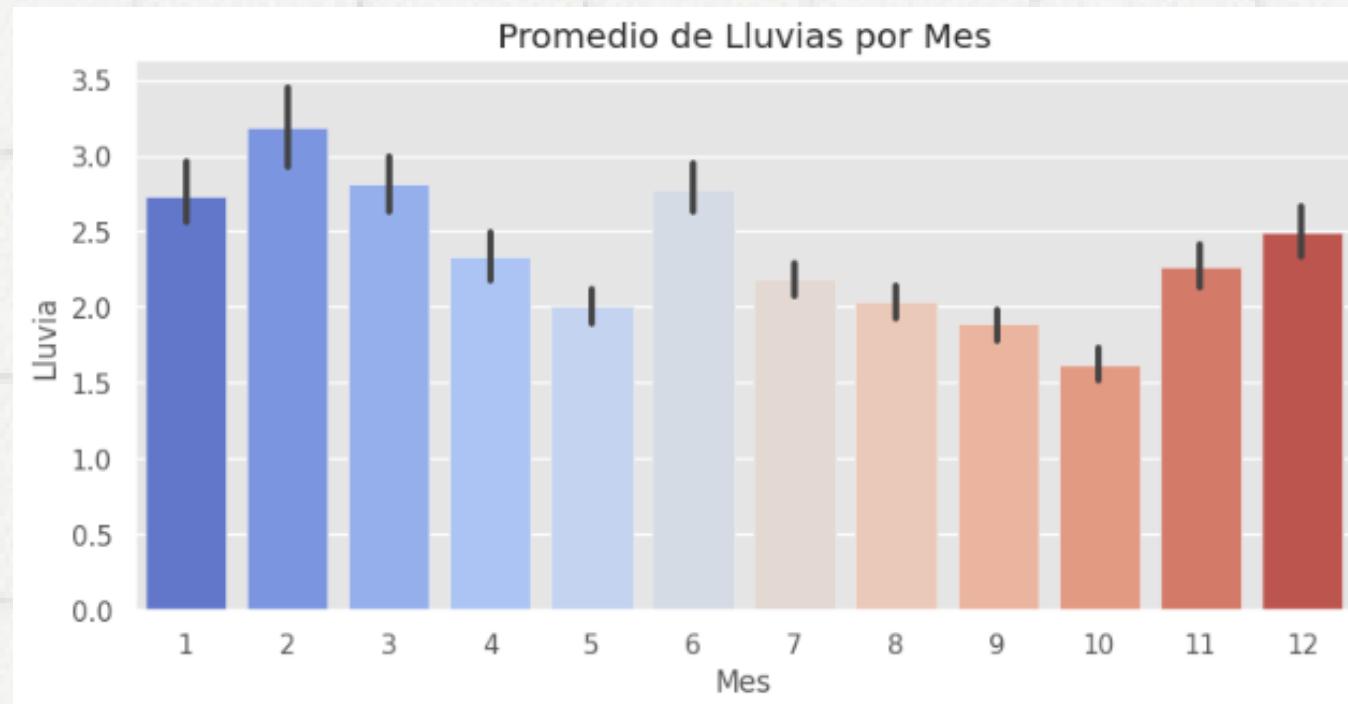
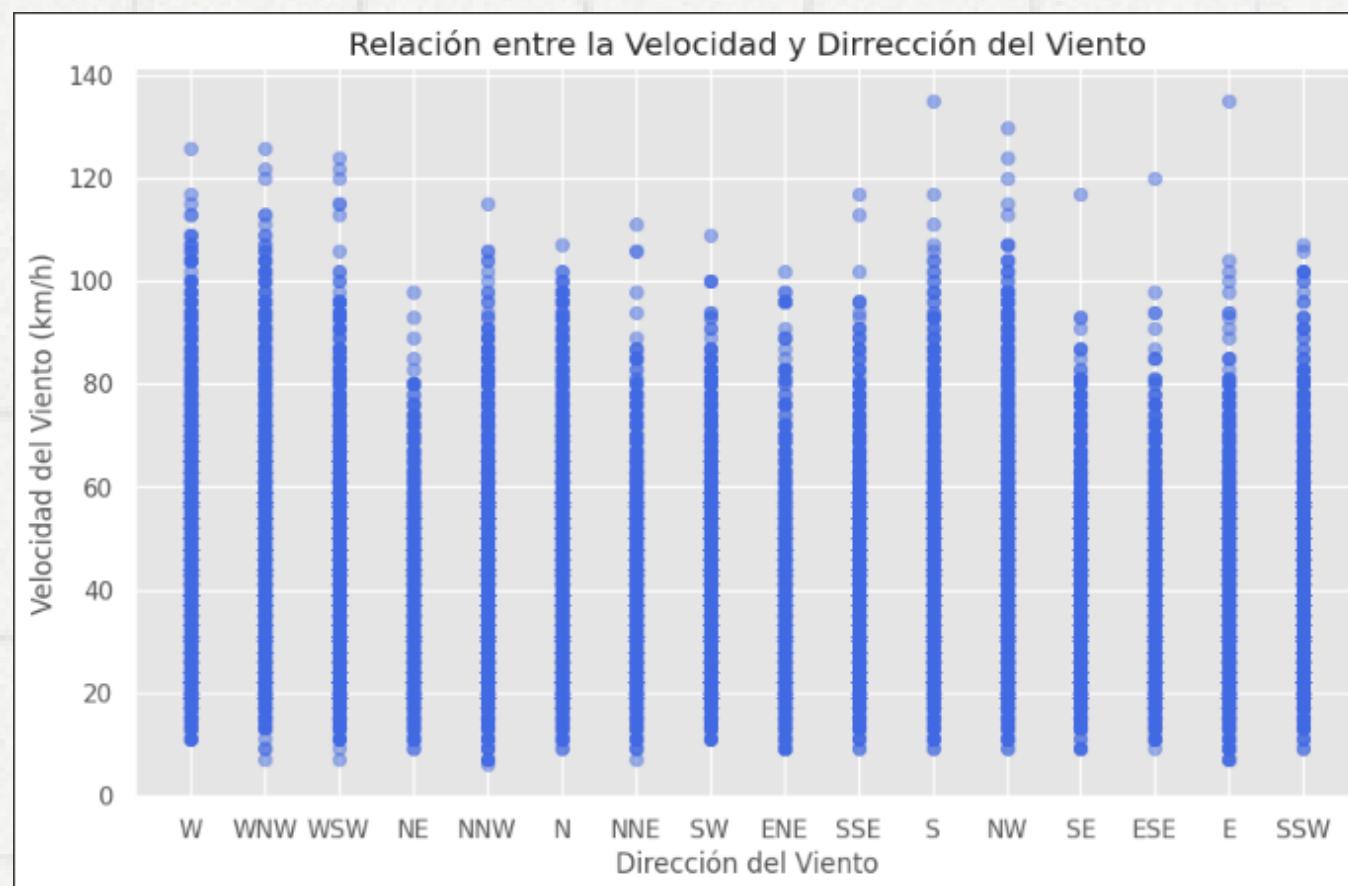


04

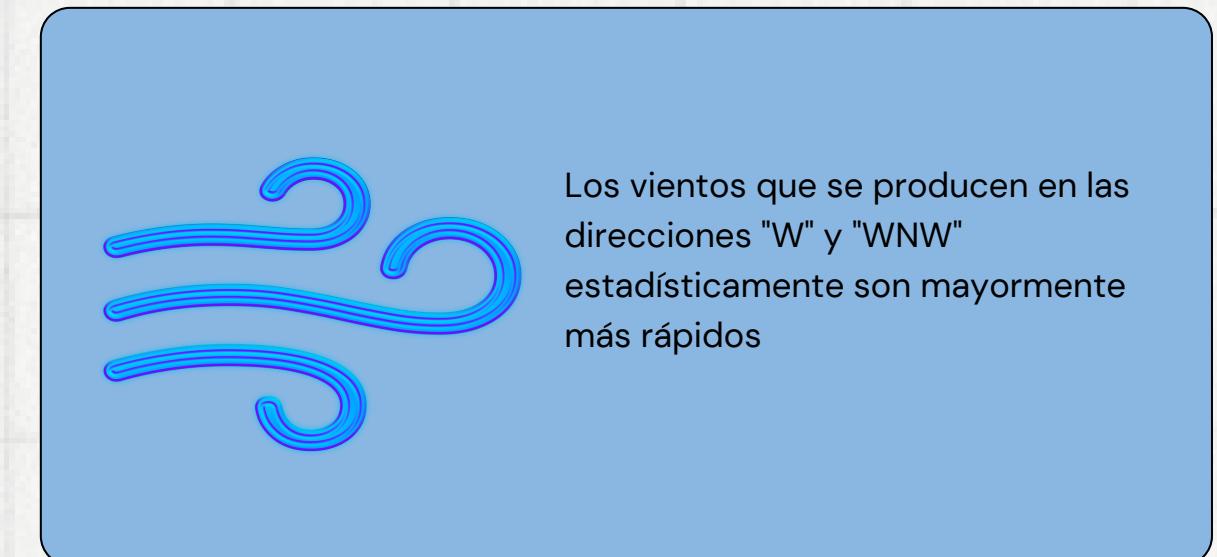
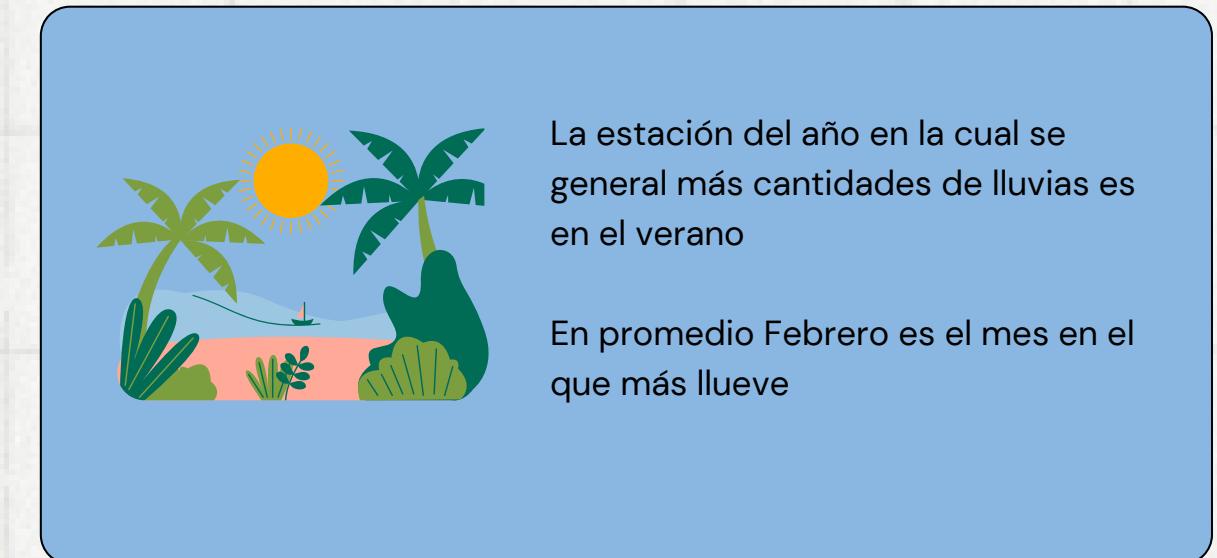
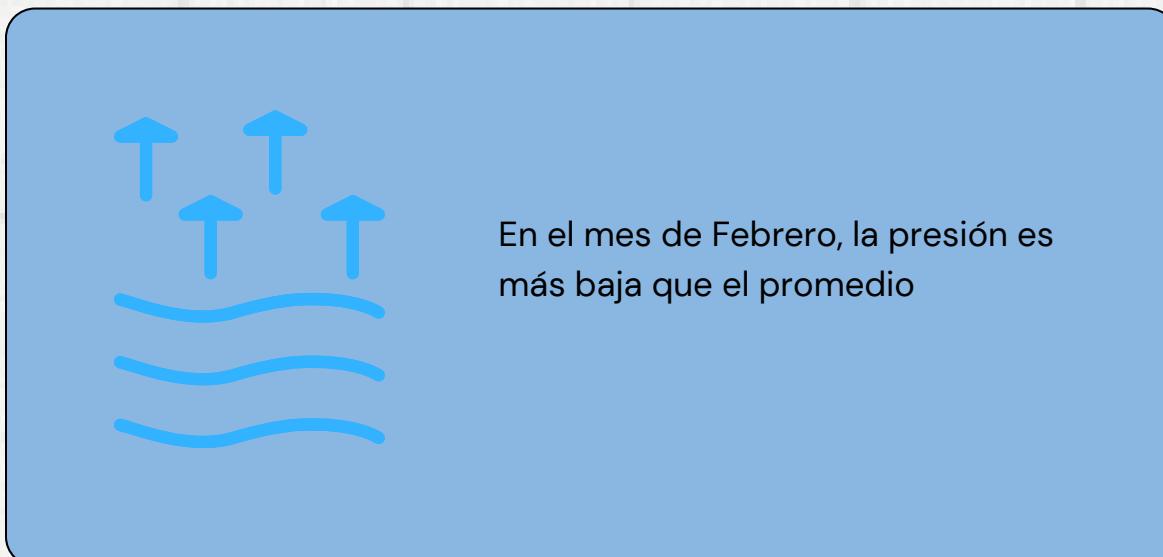
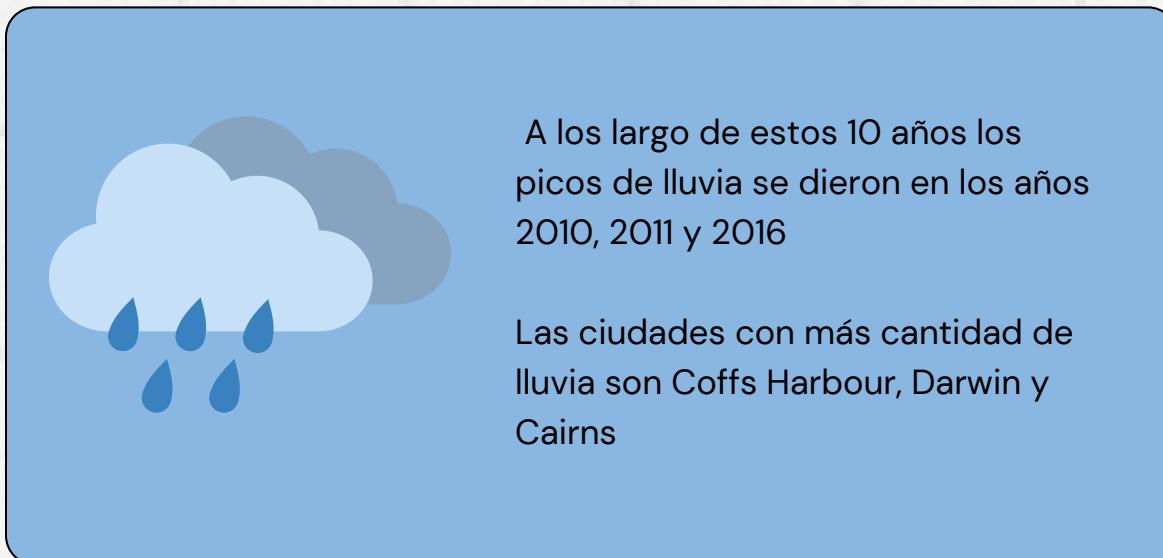
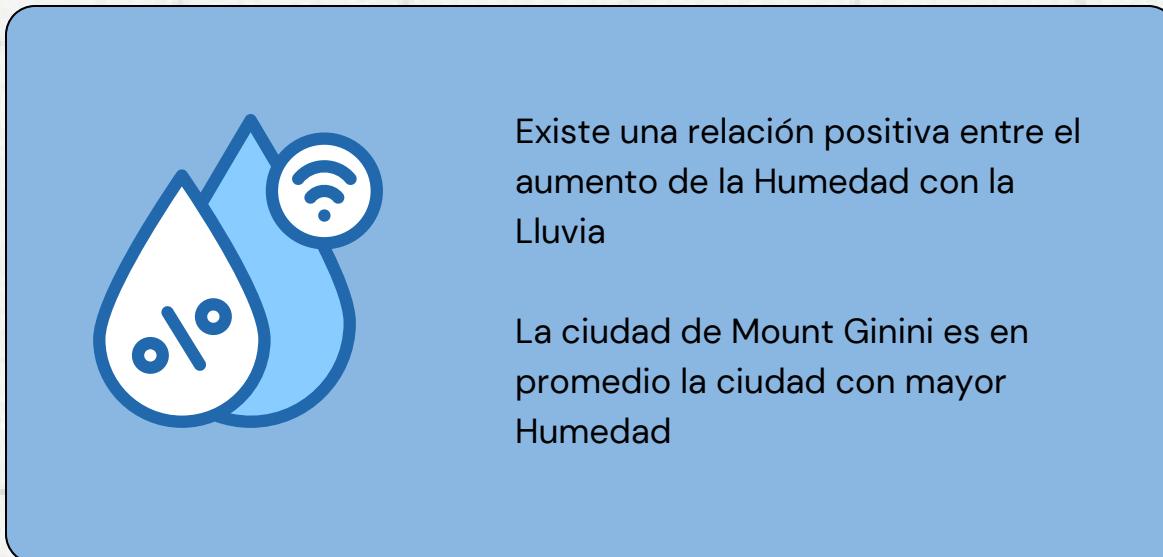


Gráficos y análisis estadísticos destacados





Principales Insights



Modelos de Machine Learning

Para este proyecto se van a utilizar Modelos de Clasificación dado que lo que se quiere saber es que si va a llover o no.

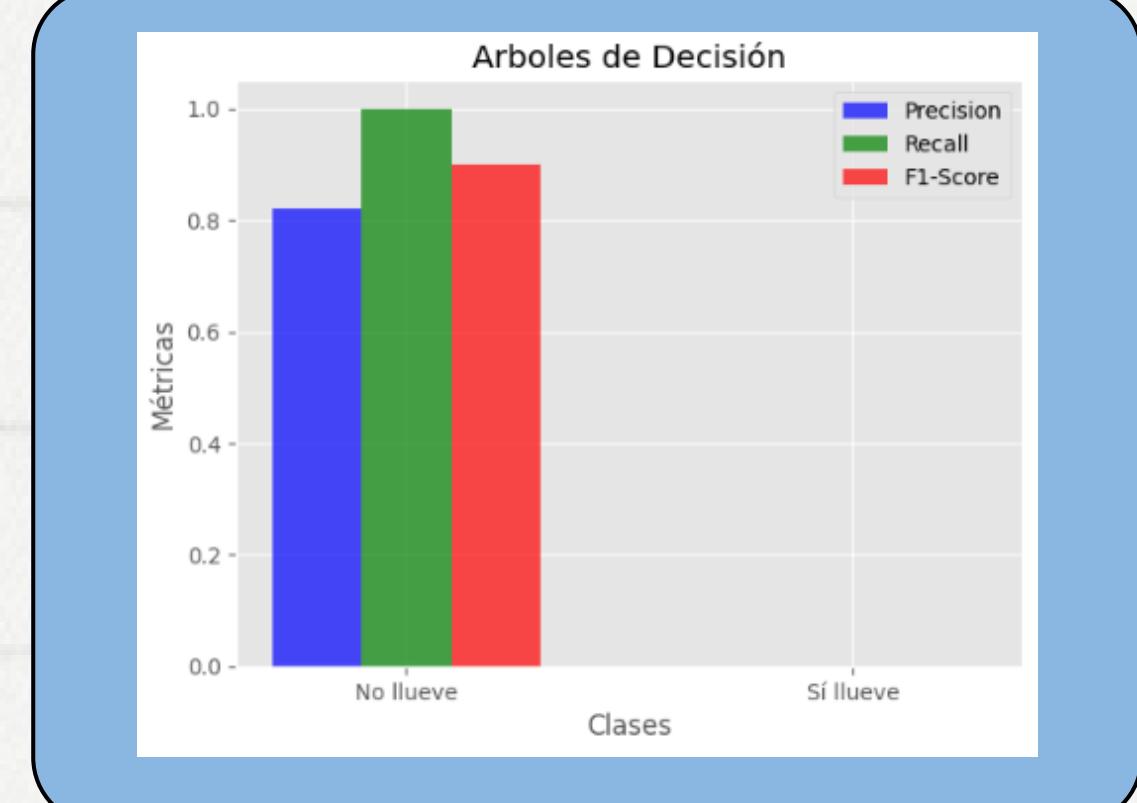
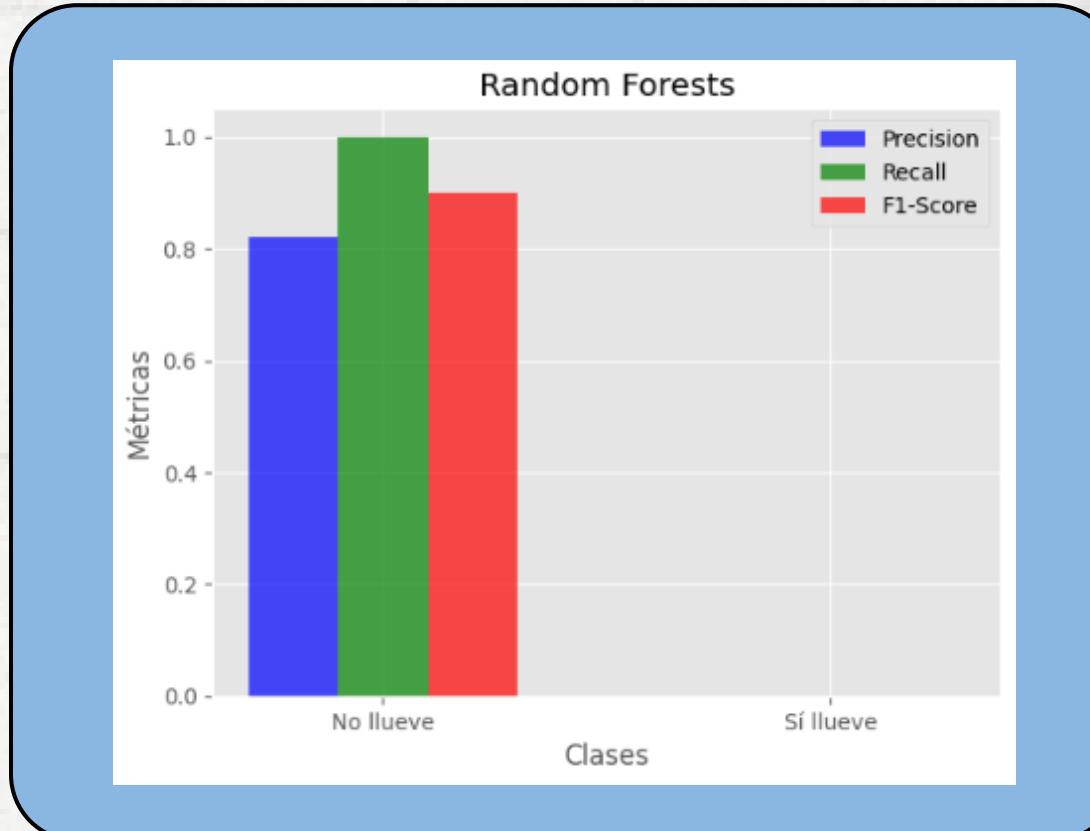
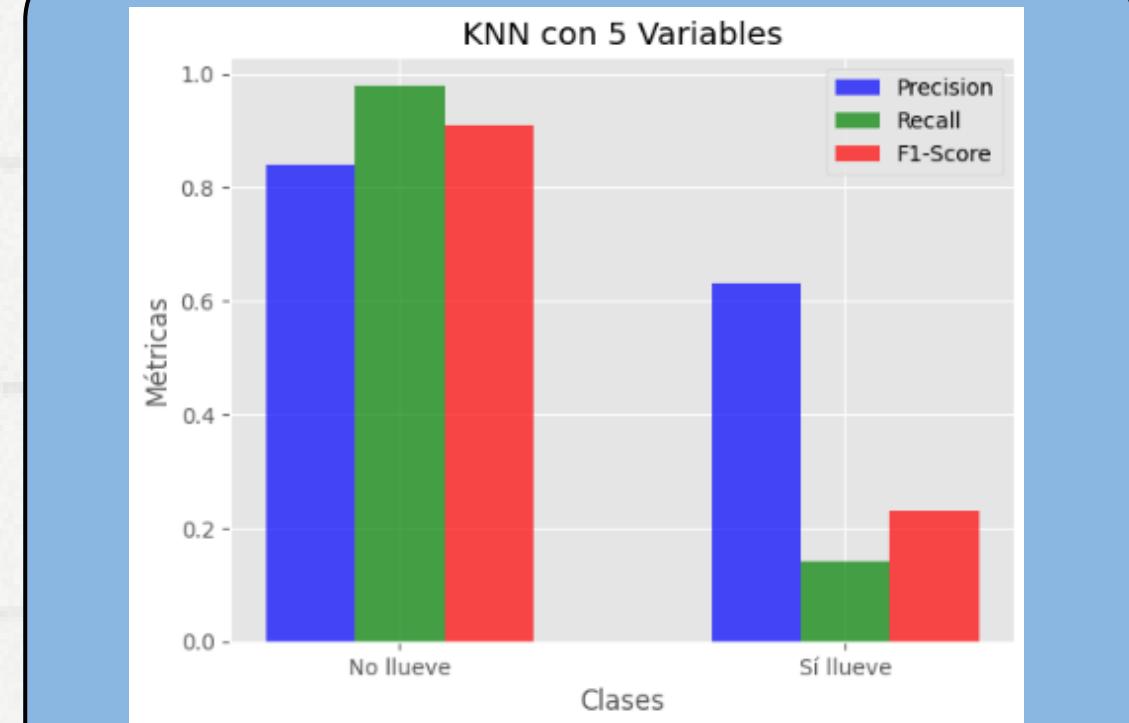
Y es por eso mismo que el Aprendizaje es Supervisado dado que tiene como objetivo predecir la respuesta que habrá en el futuro, gracias al entrenamiento del algoritmo con datos conocidos del pasado

Para eso, se opto por los siguientes Modelos:

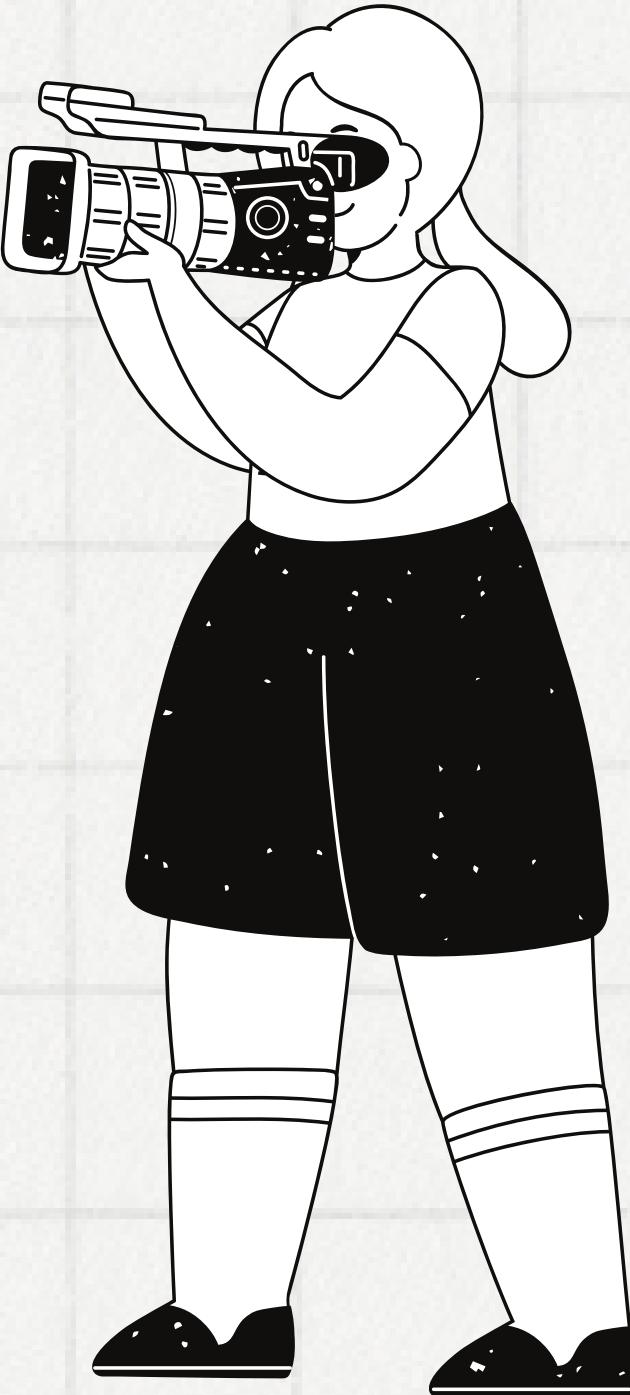
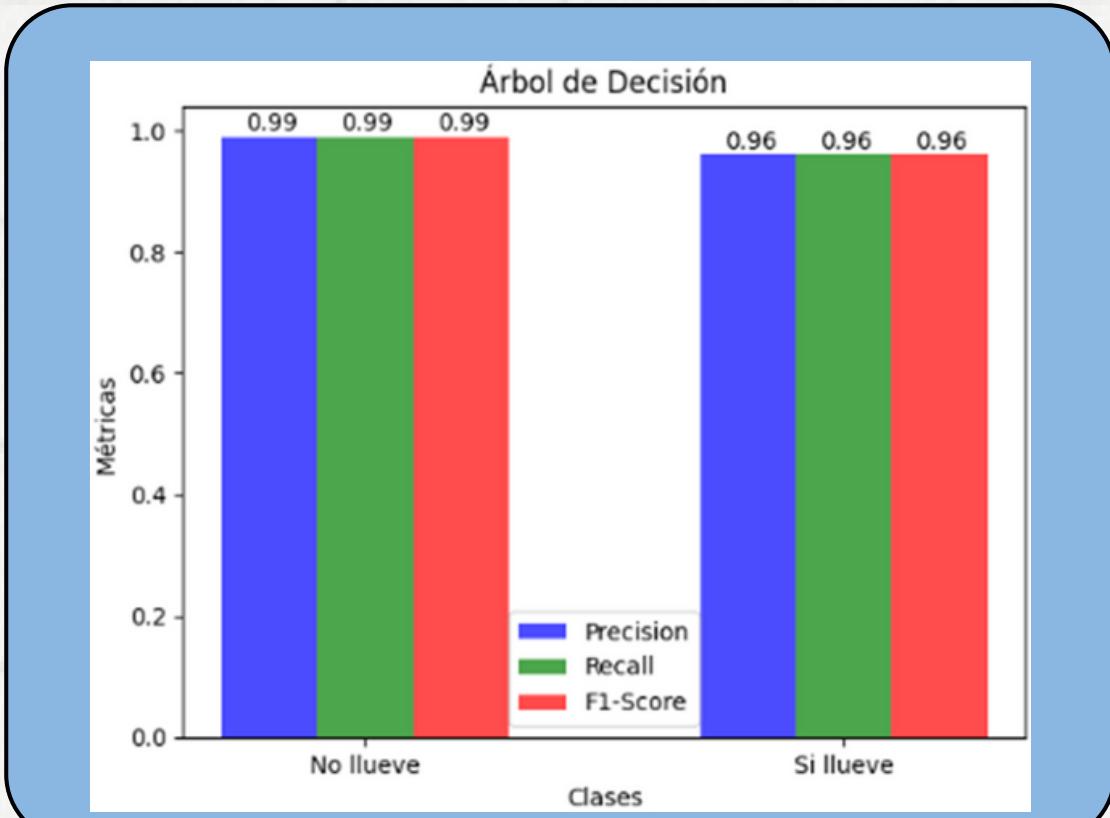
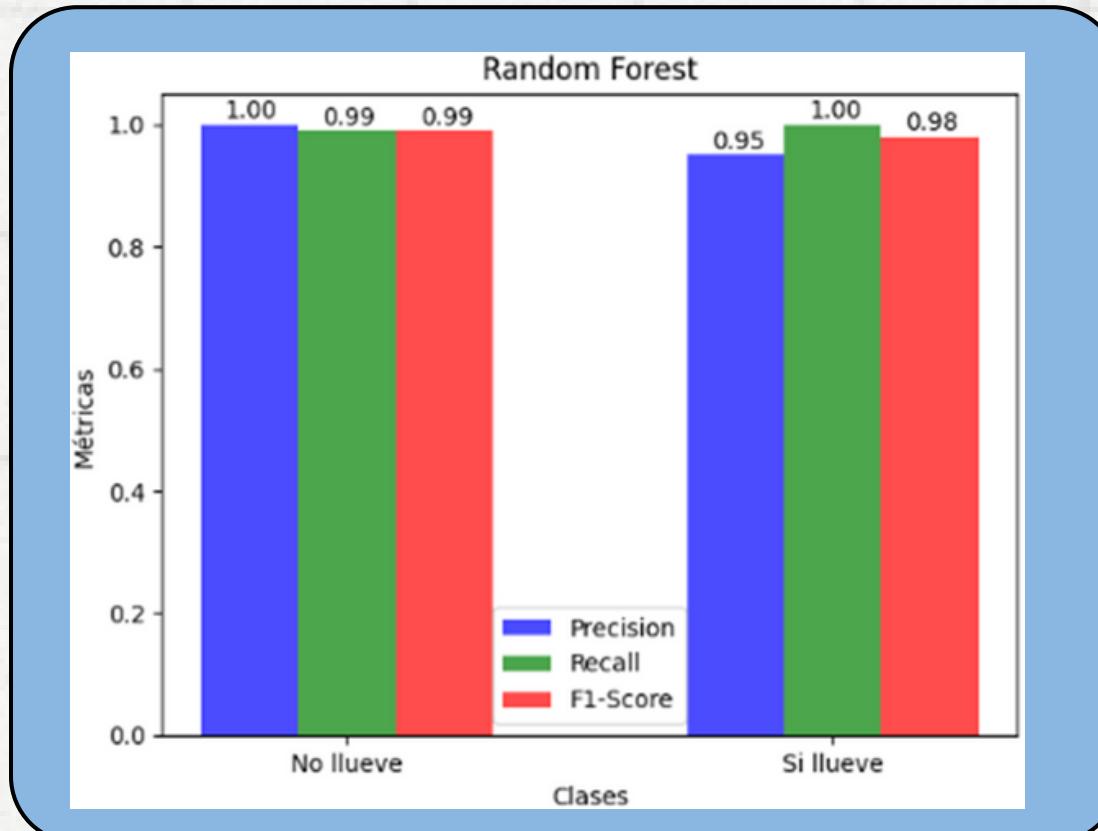
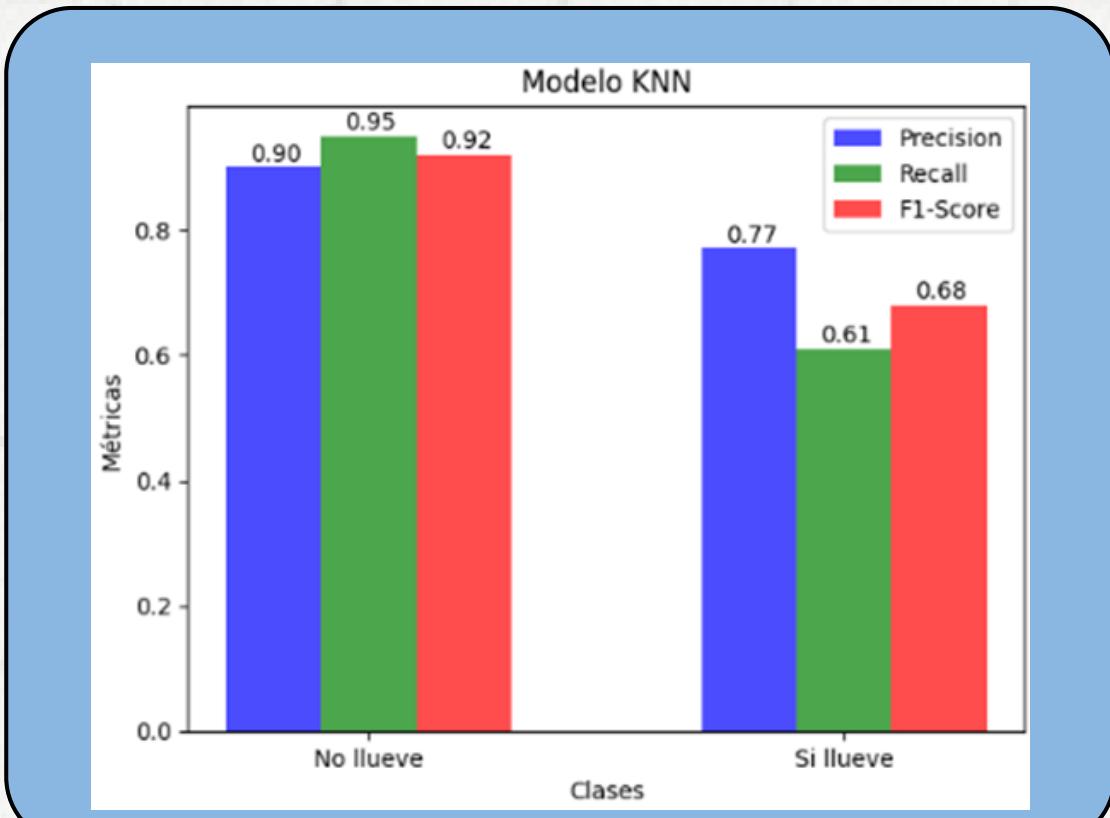
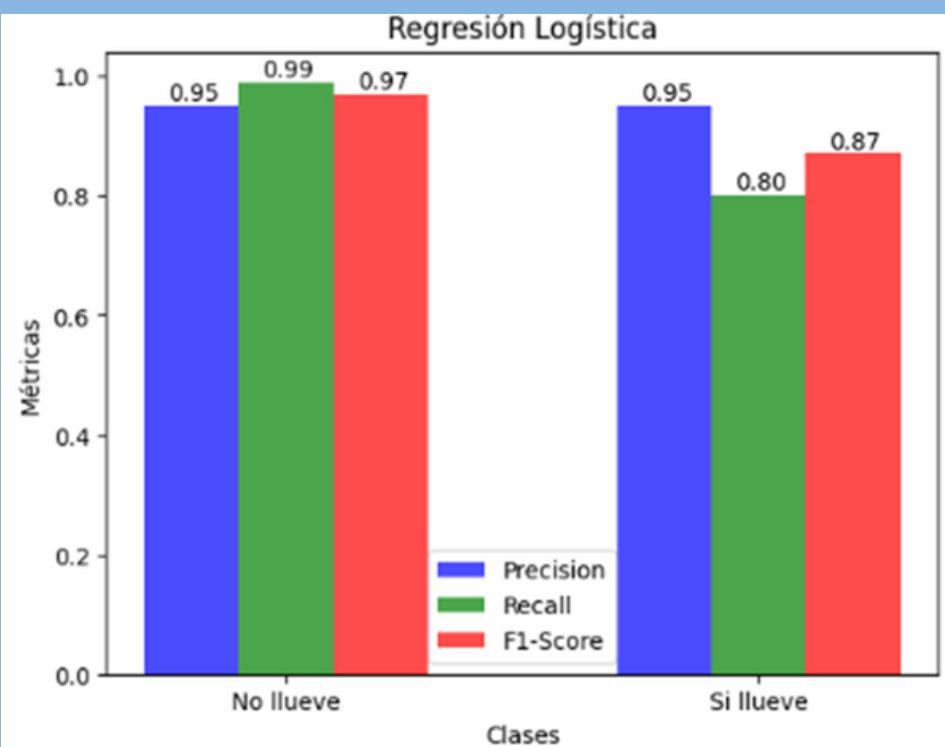
- Regresión Logística
- KNN (Vecinos Cercanos)
- Random Forests
- Arboles de Decisiones



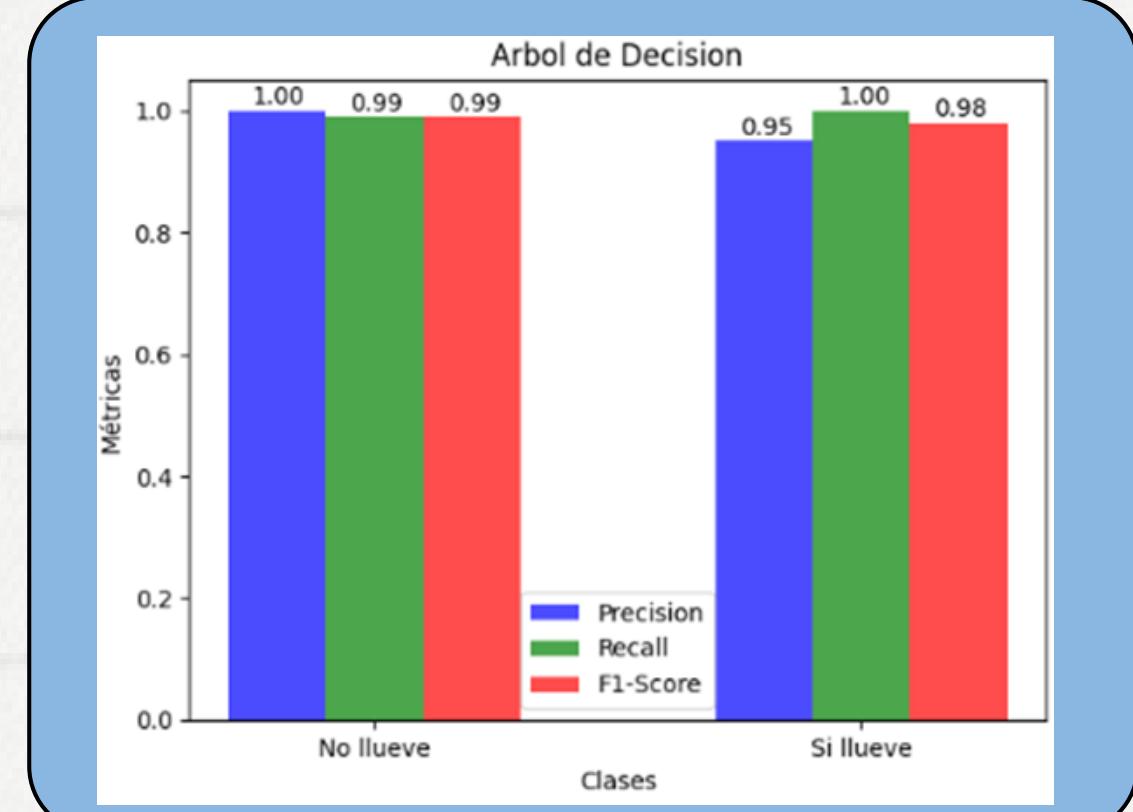
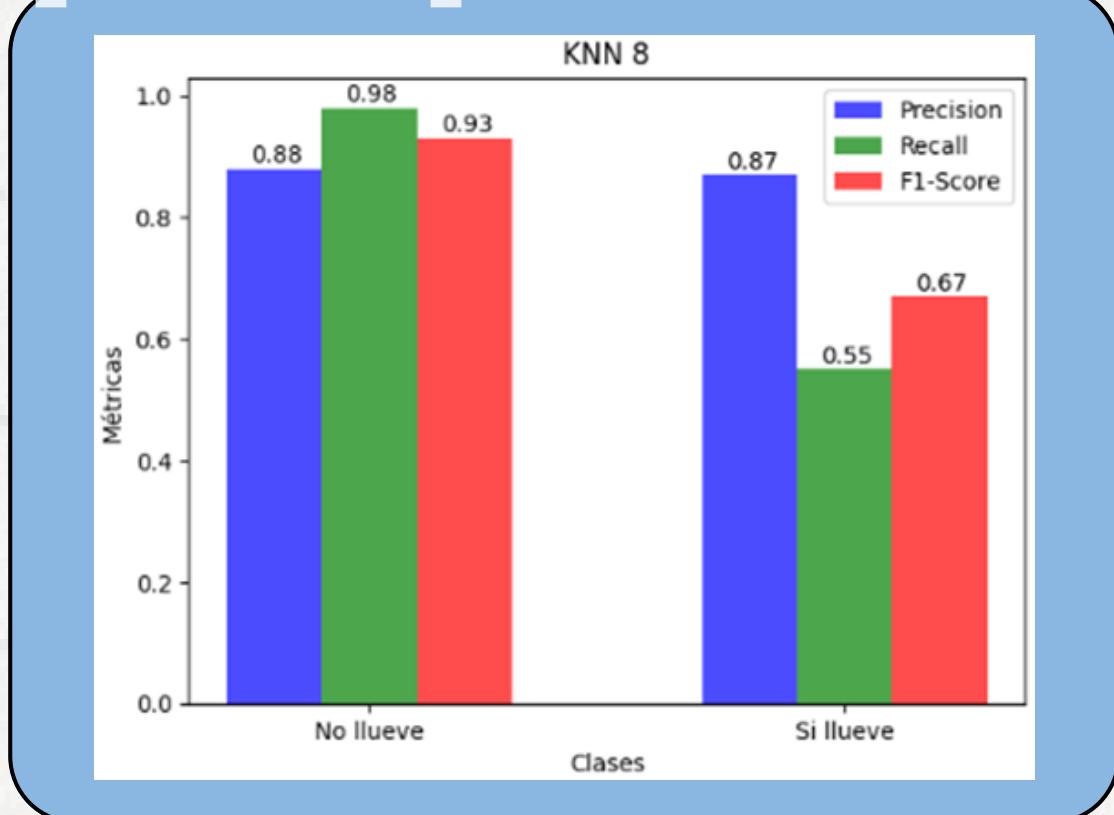
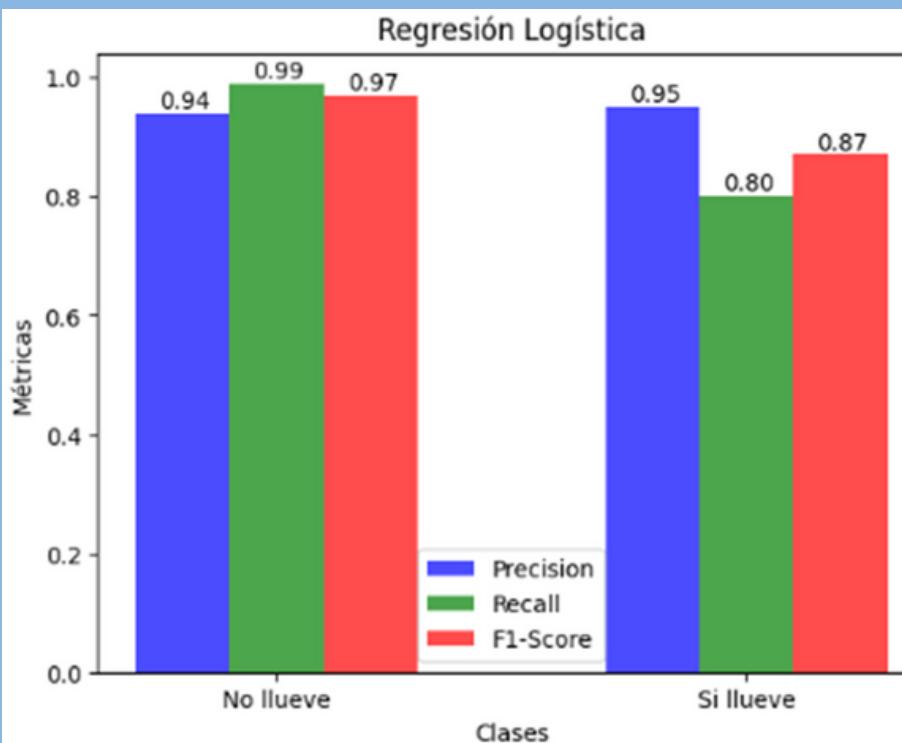
Primeros Modelos



Modelos Avanzados



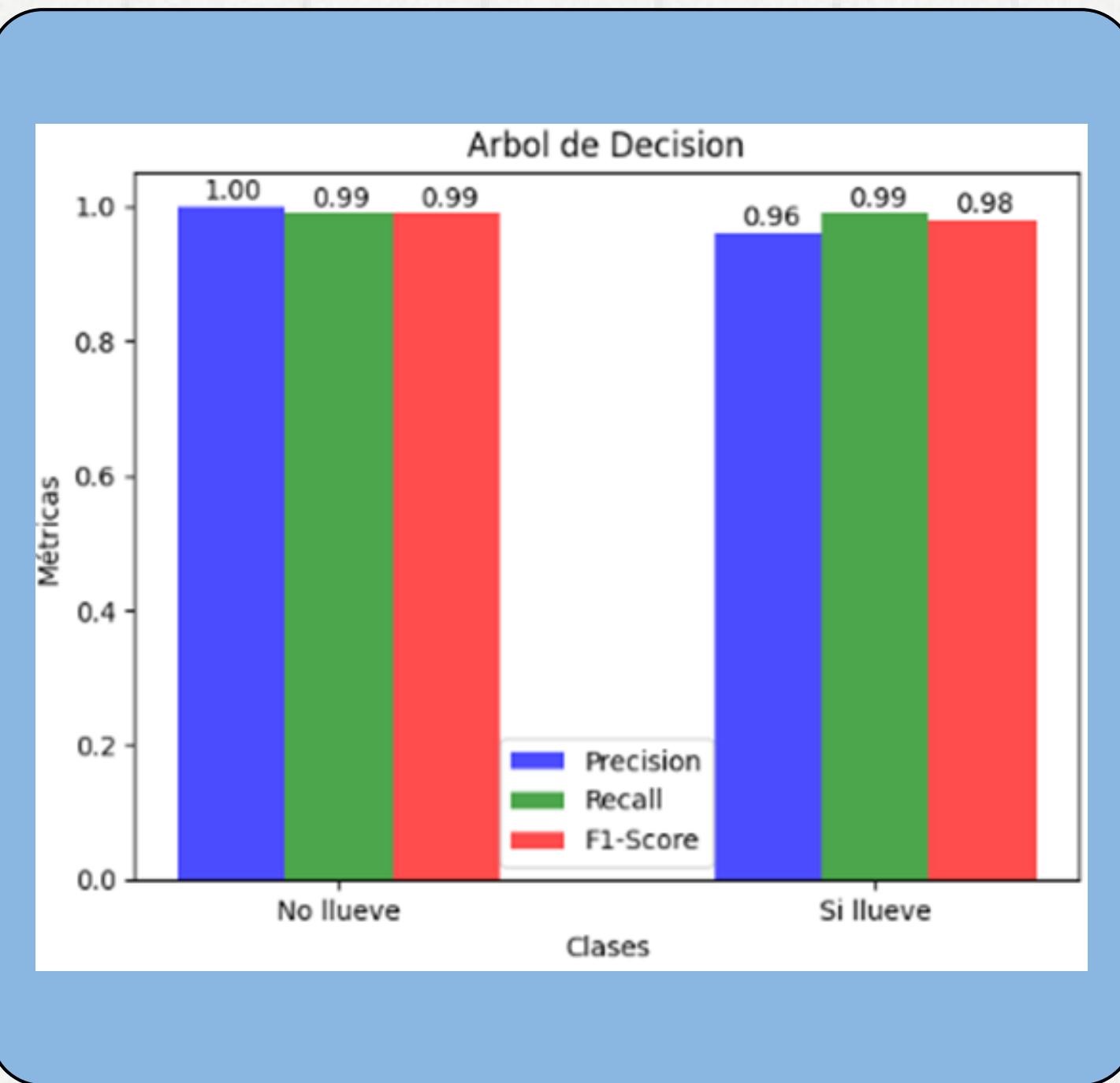
Con sus hiperparámetros



Mejores Modelos



Modelo con mejor Performance

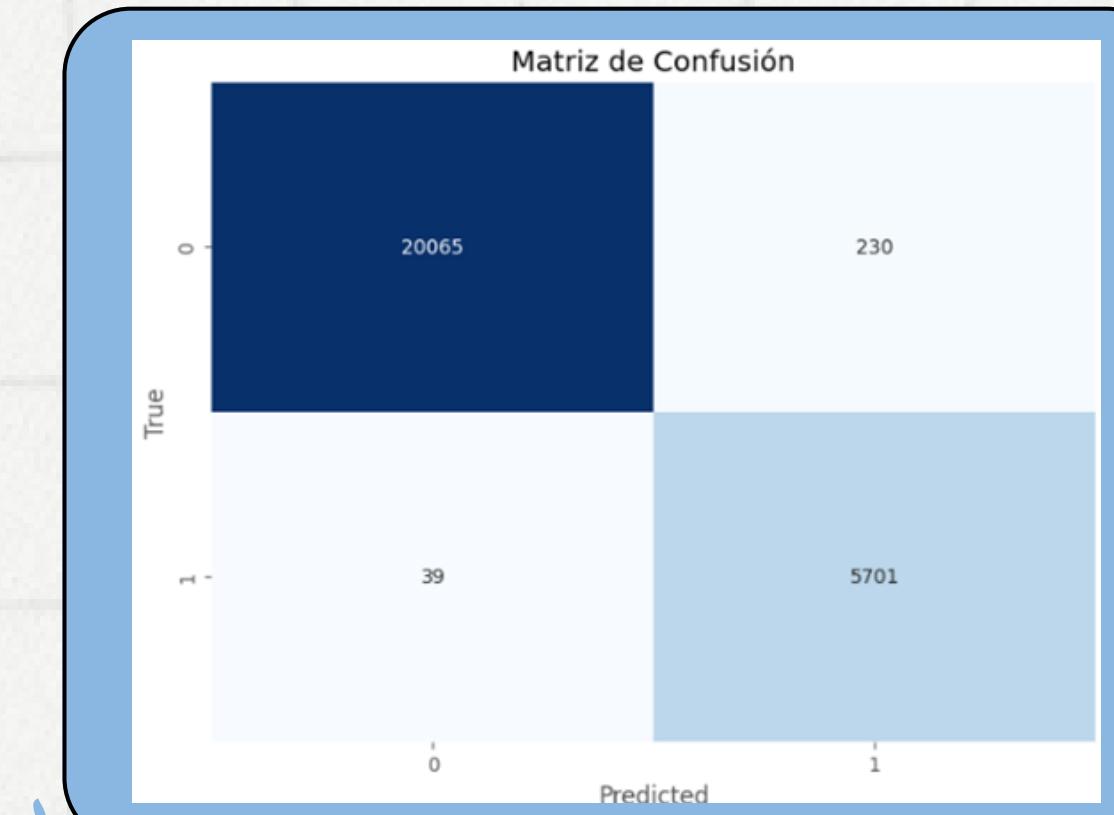
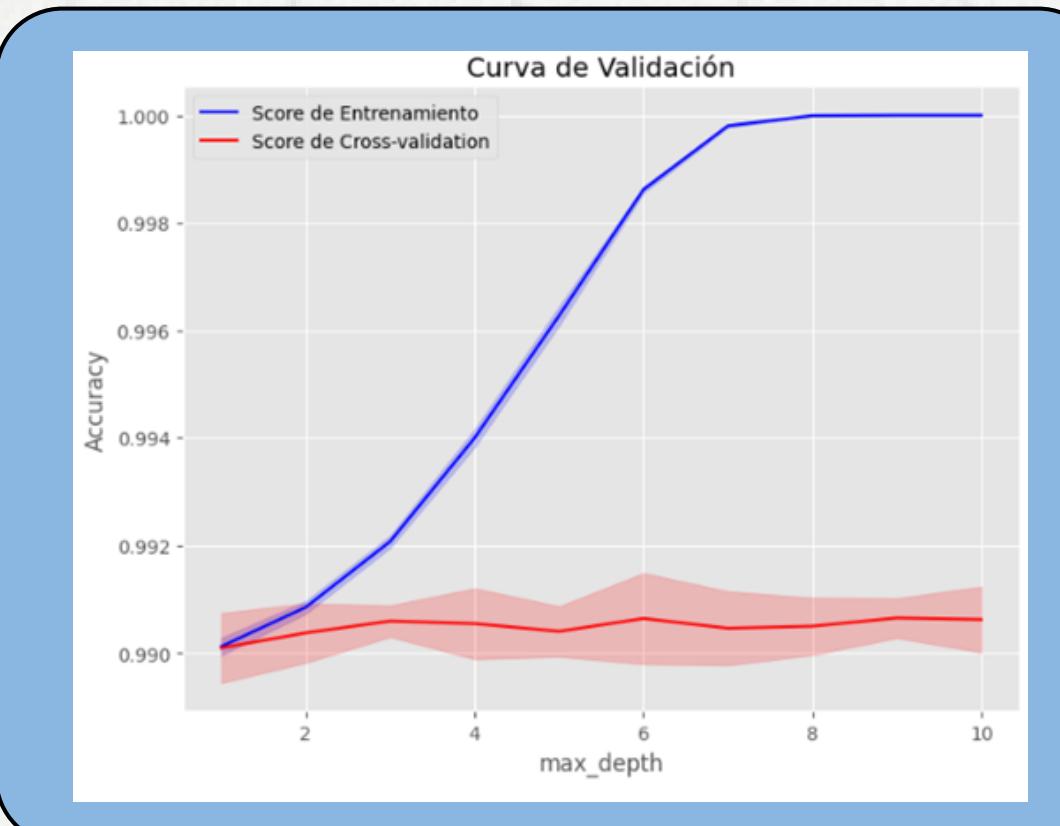
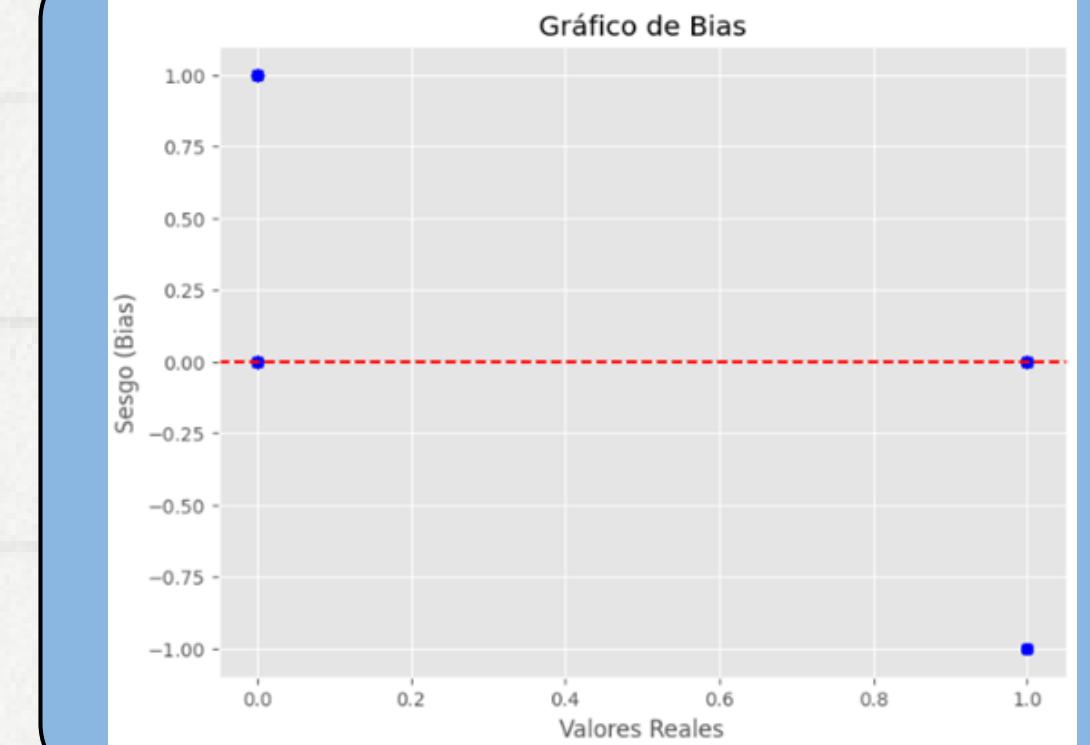
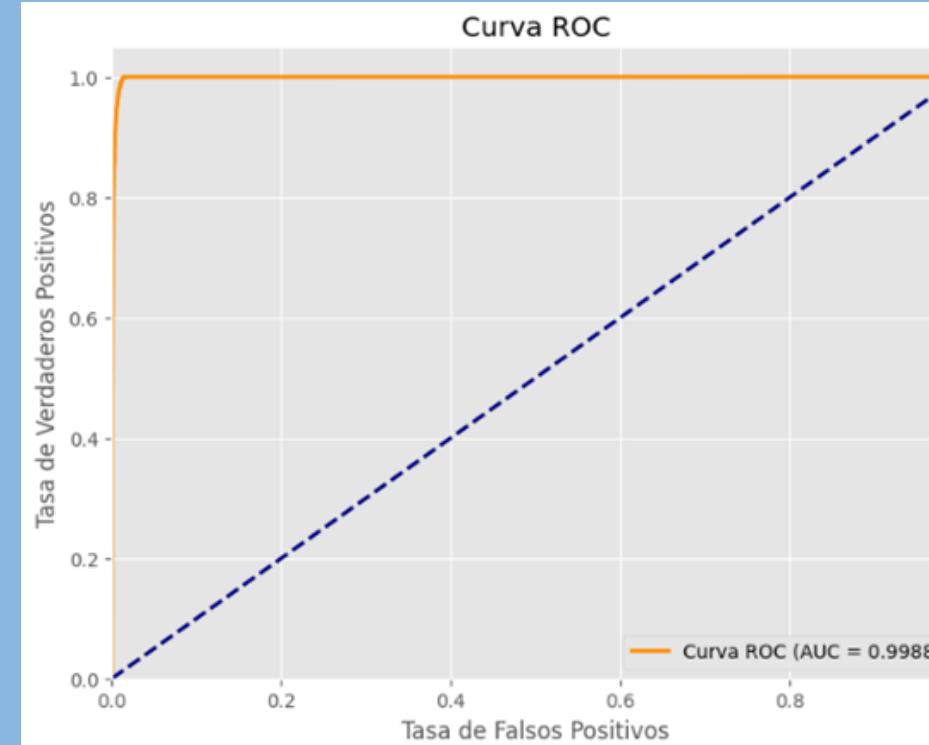


Los resultados obtenidos reflejan que el Árbol de Decisión es el Modelo que mejor performance tuvo de los distintos modelos y de sus distintas variables.



Como conclusión el equipo de Data trabajará de ahora en adelante para poder optimizar de mejor manera este modelo, con el objetivo de que sus resultados sean más precisos y poder generar más valor al modelo y al proyecto.

Evaluación del Modelo



Conclusiones

Entendimiento del problema:

En lo que respecta al objetivo del proyecto se llegó a un modelo robusto y que su performance es aceptable para aquello en lo que queremos predecir en este caso la lluvia, por lo tanto, las expectativas de nuestro jefe fueron alcanzadas, dados los requerimientos que él nos propuso se cumplieron.

Selección y construcción de características:

En este proceso se mejoró la capacidad predictiva del modelo al utilizar por ejemplo la estandarización de los datos, como así también tratar el desbalanceo de los datos lo cual fue el gran problema que se tuvo que atravesar para este proyecto ya que, naturalmente son más la cantidad de día que no llueve que los que si, y como consecuencia de esto repercute en nuestra dataset y proyecto, aunque se pudo mejorar y genera un conjunto de características óptimo y relevante, lo que conduce a modelos más precisos, interpretables y eficientes (Esto gracias a la técnicas como Smote, RandomOverSampler, Smotetoken). Además de pasar nuestros datos por una reducción de dimensionalidad como lo es PCA.

Exploración y preparación de datos:

En cuanto al EDA, se realizó varios gráficos que gracias a ello se descubrieron patrones, relaciones que tiene la lluvia con distintas determinaciones del clima como lo es la humedad, en que época de año llueve más, como es la evolución de la lluvia, no solo eso sino que también se trató varias preguntas y hipótesis con métodos estadísticos ya sea multivariados o no, y demás Insights que se puede ver en cierta parte del proyecto.

Sobre la limpieza y la detección de Outliers se eliminaron varias columnas de nuestro dataset, algunas se trabajaron con la moda y la mediana, en tanto los outliers no se trataron tanto ya que los mismo se determinaron que era un fiel reflejo de la realidad y en lo que en el día a día puede pasar por eso se los tomo como un dato más sin la consideración de que se lo tome como atípico.

Modelado y evaluación:

En este caso se tuvo una evolución muy buena de los modelos a lo largo del proyecto, significativamente gracias a nuestro Feature Engineering y al trato que se hizo en el mismo por lo anteriormente mencionado, pero no solo eso, se produjo buenos resultados en los distintos modelos con la validación simple, cruzada, hiperparámetros para cada modelo y métodos de ensamble. Por lo que pasamos de un modelo el cual no era ni apenas aceptable a uno en el cual se puede trabajar y desarrollar, aunque se puede mejorar todavía.

Conclusiones

Interpretación de resultados:

Mediante el desarrollo del proyecto se fue evaluando que le resultaba mejor a nuestro modelo, esto puede ser eliminar o no una variable que no influenciaba mucho en el modelo o lo empeoraba, balancear el dataset, reducir su dimensionalidad, que modelo se veía que andaba mejor, ese modelo con que parámetros y que hiperparámetros funcionaba más eficiente,etc.

Por lo que se tuvo la estrategia que al final del proyecto agrupar todos los métodos que se hicieron y fueron los más óptimos para combinar todo eso y generar nuestro modelo final con su mayor perfomance, y de alguna manera más sencilla y fácil de ver llevar todo el trabajo hacia una parte final del notebook el cual dé un cierre al proyecto.

Pasos a futuro

Para un futuro a nuestro modelo final le podemos realizar varias modificaciones, no solo en lo que respecta al modelo en sí, sino que también al contexto por ejemplo:

- Darle nuevos datos a probar y ver como funciona y se desempeña con esos datos.
- Estudiar como el modelo evoluciona a través del tiempo, por ejemplo en un año puede ser que la realidad haya cambiado o se generen nuevos paradigmas en lo que tiene que ver con clima (Quizá un año de muchas sequias continuas, o lluvias abundantes) y por lo tanto como consecuencia haga que el modelo no se adapte bien a las circunstancias por lo que se tendría que rever.
- Una vez llevado a producción ver como performa en la realidad si al final es aceptable o no su rendimiento, darle un tiempo para ver si funciona o no, tratar de tener un periodo de prueba en el cual se observe detenidamente como es su desarrollo
- Analizar y ver nuevamente como se llevó a cabo el progreso de Data Wrangling, ver que otros caminos se puede tomar, realizar una crítica constructiva del proceso en área de data, ver si tal medida era la más correcta o algo que no se tuvo en cuenta pueda ayudar aún más y de mejor manera el trabajo.

Pasos a futuro

Para un futuro a nuestro modelo final le podemos realizar varias modificaciones, no solo en lo que respecta al modelo en sí, sino que también al contexto por ejemplo:

- Tener en cuenta además los métodos utilizados en el cross validation, hiperparámetros, métodos de ensamble, dado que muchos de ellos son prueba y error para saber bien cual funciona mejor, tal vez comunicarse con un Machine Learning Engineer para poder optimar esto mucho mejor.
- Que se generen nuevas preguntas a partir de los datos o del modelo, ya sea en el área de data o por una necesidad de otra área de la empresa o nuestro jefe, dado que un trabajo de Data Science no tiene como tal un ciclo de inicio a fin, sino que es muy de ida y vueltas en las distintas etapas, puede darse que se generen nuevas necesidades, requisitos, situaciones en la realidad a medir y tener en cuenta, etc.
Por ejemplo, predecir cuanta cantidad puede llover en cierto mes.
- Conseguir más fuentes de datos ya sea dentro o fuera de la empresa, que tengan más años de análisis o más variables dentro de tu dataset para enriquecer el análisis.

**iMuchas
gracias por
su tiempo!**