**REPORTE TÉCNICO**

**INTRODUCCIÓN**

En Colombia, el Registro Único Nacional de Tránsito [(RUNT)](https://serviciosdetransito.com/index.php/noticias/139-que-es-y-como-funciona-el-runt), funciona como una gran base de datos centralizada en la que se almacena información sobre todos los vehículos del país, los conductores, los seguros de estos, las infracciones entre muchos más. El presente reporte asume el ejercicio de crear un modelo estadístico para predecir el número de vehículos registrados diariamente en el RUNT durante el año 2018.

**OBJETIVO**

Crear un modelo estadístico para predecir el número de vehículos registrados diariamente en el RUNT durante el año 2018, teniendo en cuenta cuenta la base de datos suministrada por el RUNT del 2012 al 2017

**BASE DE DATOS.**

La base de datos inicial, registro\_autos\_entrenamiento, en formato .XLSX, contiene los registros diarios de autos en el RUNT, desde las fechas 1/01/2012 hasta 31/12/2017. Se usa el lenguaje de Python principalmente en las librerías de PyCaret (<https://todobi.com/pycaret-paso-a-paso>) para el procesamiento de la base de datos. Inicialmente, la base de datos contiene 2192 observaciones y 2 variables. Las variables Fecha y Unidades.

**PRE-PROCESAMIENTO**

Se verifica cuantas columnas y filas tiene la base de datos y se verifica que no hayan datos nulos, además, se establecen 4 variables nuevas a partir de la fecha que son año, mes, día del mes y dia de la semana con esto se establece como días hábiles para registro de vehículos los días de la semana Lunes, Martes, Miércoles, Jueves y viernes.

Se identifica como variable explicativa la Fecha y las de respuesta son las unidades, esto es porque las unidades se predicen a partir de la fecha y no al revés, lo que significa que la variable independiente es la fecha.

Por último se crean columnas dummies para un mejor entrenamiento de los modelos

**MODELOS**

Para esta predicción se usan 4 modelos que se describirán a continuación, cada vez que se aplica un modelo, toma en cuenta los anteriores aplicados para mejorar la precisión de la predicción.

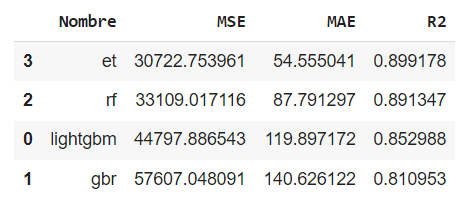
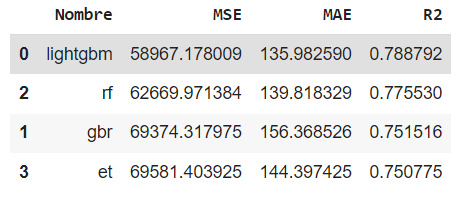
**Modelo Light Gradient Boosting Machine (lightgbm):** es una biblioteca de código abierto que proporciona una implementación eficiente y efectiva del algoritmo de aumento de gradiente. LightGBM amplía el algoritmo de aumento de gradientes añadiendo un tipo de selección automática de características, así como centrándose en aumentar los ejemplos con gradientes más grandes. Esto puede resultar en una dramática aceleración del entrenamiento y en una mejora del rendimiento predictivo.

**Modelo Gradient Boosting Regressor (gbr):** Es una forma de combinar múltiples modelos simples en un solo modelo compuesto. A medida que combinamos más y más modelos simples, el modelo final completo se convierte en un predictor más fuerte. El término "gradiente" en "aumento de gradiente" proviene del hecho de que el algoritmo utiliza el descenso de gradiente para minimizar la pérdida.

**Modelo Random Forest Regressor (rf):** Un modelo Random Forest está formado por un conjunto de [árboles de decisión](https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html) individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generados mediante [bootstrapping](https://en.wikipedia.org/wiki/Bootstrapping_(statistics). La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

**Modelo Extra Trees Regressor (et):** este modelo lleva la aleatoriedad de Random Forest un paso más allá. Además de considerar un subconjunto de las características predictivas para cada uno de los árboles a crear, a la hora de escoger una característica y un valor de corte (threshold) para dividir cada nodo, en lugar de escoger el threshold que mejor divida cada característica, se genera un valor de corte aleatorio para cada característica planteada, encogiéndose como regla de división el mejor de ellos.

Para finalizar se realiza la comparación de todos los modelos aplicados presentados a continuación:

* Para los datos de entrenamiento y la comparación de los modelos entrenados y mejorados:  
  
* Para los datos de validación e igualmente la comparación de los modelos entrenados y mejorados:  
  

**CONCLUSIONES**

* De la gráfica de registros de vehículos por día del mes, de los días 28, 29 y 30 de cada mes, se observa un incremento significativo de registros de vehículos.
* En la gráfica de registros de vehículos por meses, se observa un aumento significativo al finalizar cada año, esto posiblemente relacionado con las fiestas decembrinas.
* Se nota que el modelo 'et' tiene una caída notable luego de aplicar la validación en cuanto al R2 que es nuestra métrica de interés. El modelo 'gbr' no se destaca mucho en comparación con los demás, mientras que el modelo 'rf' se mantiene como segundo en ambas fases, y el 'lightgbm' es el que menor disminución tiene luego de aplicar la validación así que decidiremos implementar este.
* Se observa que la predicción final para el año 2018 se ajusta al comportamiento que ha tenido regularmente el registro de vehículos en el histórico de datos mostrando un incremento en los días finales de cada mes y un incremento significativo finalizando en año lo que corresponde a un comportamiento cíclico en el número de vehículos registrados por año