**Máster en Big Data y Data Science**

*Análisis para el mantenimiento preventivo en autobuses*

Trabajo Fin de Máster

Alumno: Gómez Ramírez, Daniel

Dirección: Valencia

Tutor Trabajo Fin de Máster: Peralta Martín-Palomino, Arturo

Edición octubre 2021 a marzo 2022

Índice

[Resumen 5](#_Toc97639008)

[1. Introducción 5](#_Toc97639009)

[2. Objetivos 7](#_Toc97639010)

[3. Marco teórico 8](#_Toc97639011)

[4. Desarrollo del proyecto y resultados 9](#_Toc97639012)

[4.1. Metodologia 9](#_Toc97639013)

[4.2. Resultados 9](#_Toc97639014)

[5. Conclusiones 12](#_Toc97639015)

[6. Referencias 13](#_Toc97639016)

*¡¡¡INFORMATIVO!!! (quitar)*

*\*Actualizar tabla de contenidos*

*\*Hacer uso de salto de página para los nuevos apartados.*

*\*Hacer uso de los estilos definidos: Normal, Título1, Título2, Título3.*

Índice de figuras

[Figura 1. Arquitectura Jupyter Cliente - Servidor. Fuente: https://www.paradigmadigital.com/dev/jupyter-data-science-aplicada/ 4](#_Toc6399225)

*¡¡¡INFORMATIVO!!! (quitar)*

*Sólo necesario si la actividad tienen una gran cantidad de figuras y un extensión considerable. Para actividades "cortas" no es necesario.*

*\*Actualizar índice de tablas (Referencias / Insertar tabla de ilustraciones)*

Índice de tablas

[Tabla 1. Operaciones matemáticas utilizadas en el estudio realizado. Elaboración propia. 3](#_Toc6399146)

# Resumen

En el presente TFM se analizará un dataset con datos sobre diferentes reparaciones en piezas de autobuses. Dependiendo de la pieza del autobús, la reparación se lleva a cabo dentro de un periodo de tiempo. Se pretende realizar una predicción de ese tiempo estimado para así, ahorrar costes y prevenir problemas futuros.

Además, se apoyará dicho análisis con un cuadro de mando realizado en Power BI

# Introducción

En el presente TFM se analizará un conjunto de datos, suministrados por la empresa Softour Sistemas. Su problema actual es que se necesita reducir gastos, y una línea en la que están trabajando y en la cual podrían ahorrar costes es en la de las reparaciones de los autobuses. Actualmente disponen de mantenimiento correctivo y de mantenimiento preventivo, y están interesados en incluir mantenimiento predictivo, utilizando los datos históricos de los que disponen.

Lo que supondría que en vez de arreglarlos cuando se estropean, se revisará el autobús previo a sufrir una avería y no tendrían que ir, con tanta frecuencia, a taller de forma irremediable por una avería, lo que podría afectar a otras piezas, las cuales no habrían sufrido deterioro si se hubiera hecho un mantenimiento antes de averiarse, con el consecuente ahorro en coste de piezas y tiempo de reparación y reducción de tiempo de inmovilización del vehículo, lo que aumenta las horas de producción.

Para lograrlo se pretende implementar un modelo que prediga en función del conjunto de datos facilitado, cuanto tiempo tardará un autobús en averiarse, en concreto cuando el motor sea el motivo.

Actualmente disponen de un conjunto de datos sobre diferentes reparaciones en piezas de autobuses. Dependiendo de la pieza del autobús, las reparaciones se efectúan al cabo periodo de tiempo. En un principio se contempló buscar la solución planteando de un problema de regresión, intentando predecir el número días en el cual un autobús iba a volver a taller. Pero debido a que los datos de los que se dispone tienen pocas observaciones y uno de los dispositivos no había registrado correctamente las características medibles, no se obtuvo unos buenos resultados, ya que el R2 de los modelos que se probaron, rondaban en 0.2, con lo que se decidió plantearlo como un problema de clasificación.

Se puede consultar las distintas pruebas en el anexo el problema de regresión: [*enlace*](https://github.com/SoftDani/Notebooks/blob/main/TFM/TFM_560_480_v5.ipynb)

*https://github.com/SoftDani/Notebooks/blob/main/TFM/TFM\_560\_480\_v5.ipynb*

Para poder predecir, si la avería es por el motor, para poder tener una cantidad de datos razonables, por encima de 1000 observaciones, se han de reducir las características, en concreto, las cuales, uno de los dispositivos no ha registrado datos.

Con ayuda de criterio experto se han creado unos rangos de tiempos con los cuales se puede obtener información que puede aportar valor, creando 3 grupos en los cuales determinan el tiempo desde que se produjo una avería en cualquier otra subárea, ‘de 0 a 2 meses’, de 2 a 6 meses’, ‘más de 6 meses’ ya que, a los 6 meses un autobús ha de pasar la ITV.

Para el procesado de datos y creación de los modelos se ha utilizado Python, que es un lenguaje de programación de alto nivel, que no es necesario compilar para ejecutar.

# Objetivos

Descripción y contextualización de la actividad.

¿Que se pide realizar? (puede ser las mismas cuestiones de la actividad colgadas por el profesor en el campus)

*¡¡¡INFORMATIVO!!! (quitar)*

# Marco teórico

Describir de manera clara el método empleado para cumplir los objetivos.

Los métodos, procedimientos, materiales y equipos utilizados deben describirse con suficiente precisión y detalle como para que el trabajo sea reproducible por otros profesionales del área.

Deben mencionarse las razones por las cuales fue seleccionada la metodología empleada.

Limitaciones, si las hay, de la metodología.

*¡¡¡INFORMATIVO!!! (quitar)*

# Desarrollo del proyecto y resultados

## Metodologia

## Resultados

Los resultados en un principio se pretendían evaluar, con el parámetro de f1-score que combina la precisión que es la proporción de verdaderos positivos entre todos los resultados positivos, ‘VP/(VP+FP)’, con la sensibilidad, que es la proporción de casos positivos entre el total de positivos reales ‘VP/(VP+FN) ‘, en una sola métrica, ya que lo ideal es tener un valor alto de precisión y de sensibilidad, obteniendo de este modo un buen valor de f1-score.

De este modo se puede ver que la sensibilidad nos indica numéricamente cuando el modelo se decanta por una de las clases y la precisión nos indica numéricamente como acierta dentro de esa clase.

En nuestro caso particular, esto no se daba, con lo que en vez de elegir el modelo con el mejor parámetro de f1-score, se ha observado que para anteponer la seguridad no se debía seleccionar el modelo con mejor f-1 score, se ha escogido el modelo en el cual la precisión en los casos en los que mas tardaba el autobús en ir a taller ‘entre 2 y 6 meses’ y ‘ más de 6 meses’ fuera alta, aunque la sensibilidad fuera baja, y por el contrario en el caso en los que menos tardaba el autobús en ir a taller ‘entre 0 y 2 meses’ se ha priorizado una sensibilidad alta, aunque la precisión fuera baja.

Por tanto al haber realizado varios modelos, y comprobar que la precisión de los modelos con mejor f1-score, no es muy alta en las clases que más tiempo tarda el autobús en ir a taller, se ha buscado un modelo que anteponga la seguridad de los autobuses a la precisión, esto quiere decir, que se ha priorizado que prediga con mas frecuencia el rango de tiempo entre ‘entre 0 y 2 meses’ aunque sea erróneamente, ya que es mas seguro para el autobús al predecir que habrá una avería antes de lo que en realidad sucedería, ya que al contrario sería catastrófico, al averiarse antes de haber sido revisado.

El mejor resultado que hemos obtenidos con esta premisa ha sido con un modelo Random Forest, sin balancear los datos.

Tabla

Descripción generada automáticamente

Como se puede observar tenemos una sensibilidad baja para las dos primeras clases, pero una precisión relativamente alta, actualmente los mejores resultados de precisión, y para la tercera clase tenemos una sensibilidad alta, aunque una precisión algo más baja, priorizando de esta forma la seguridad, quedando la matriz de confusión de este modo, teniendo en cuenta que los valores predichos se representan en el eje de las x (abscisas) y los reales en el eje y (ordenadas):

Imagen que contiene Patrón de fondo

Descripción generada automáticamente

Se puede ver como el modelo se decanta más por la clase 3, ‘de 0 a 2 meses’ aunque falla mas ya que un 45% de los valores predichos se reparten entre la clase ‘de 2 a 6 meses’ y ‘más de 6 meses’, siendo el error mayor en la clase 2 entorno a un 28% y en la clase 1 entorno a un 17%, lo que favorece al modelo ya que la primera y la segunda clase son contiguas en lo que a tiempo se refiere.

Por otro lado, el modelo selecciona menos las clases 2,’ de 2 a 6 meses’, pero cuando los selecciona acierta más, en un 80%, en el caso de fallar lo hace a nuestro favor ya que predice más la clase 1, siendo esto positivo desde el punto de vista de la seguridad.

Como contrapunto en el caso de que el modelo prediga la clase 1, ‘más de 6 meses’ algo poco frecuente como muestra la sensibilidad, acierta en un 60% decantándose también más por la clase 2 contigua en el tiempo con ella entorno a un 26% y entorno a un 13% a la clase 3, que es la que menos antepone la seguridad.

No obtenemos los valores deseados, pero si podemos decir que el modelo antepone la seguridad de los autobuses en el caso de errar.

# Conclusiones

Conclusiones que se derivan de forma lógica del trabajo realizado.

En el caso de incluirse reflexiones de carácter menos objetivo, indirectas o colaterales al trabajo, deberán justificarse.

Pueden incluirse sugerencias y trabajos futuros sobre la aplicación del trabajo y sus resultados.

*¡¡¡INFORMATIVO!!! (quitar)*

# Referencias

Sí se han utilizado referencias bibliográficas externas en alguna de las etapas de desarrollo de trabajo. Por un lado, las referencias deberán ser incluidas en el texto, en un formato corto mostrando año y apellido de autor, por otro lado, se debe incluir esta misma cita en el apartado de referencias.

La normativa a seguir para hacer referencias será APA en formato de paréntesis.

https://normasapa.com/como-hacer-referencias-bibliografia-en-normas-apa/

*¡¡¡INFORMATIVO!!! (quitar)*