Asignatura

TRABAJO FIN DE MÁSTER

PROYECTO: *Recomendador de precios de ventas de vehículos de segunda mano*

CURSO 2023-2024

MÁSTER UNIVERSITADO EN GESTION Y ANALISIS DE GRANDES VOLUMENES DE DATOS: BIG DATA

UEMC universidad online

UEMC Universidad Europea Miguel de Cervantes

I. CONCEPTO DEL TFM

El trabajo de fin de máster debe reflejar la aplicación de los conocimientos aprendidos durante el máster y, por tanto, el grado de implicación, precisión y rigurosidad en su desarrollo también deben ser altos. Se miden tanto la obtención de competencias profesionales como competencias genéricas que se verán reflejadas en su capacidad de realizar el documento escrito como en su defensa.

II. MODALIDAD DE TFM

El trabajo fin de máster consistirá en un proyecto individual y original de análisis y clasificación de muestras de malware disponibles para desarrollar modelos predictivos que ayuden a mejorar sistemas de detección de malware.

III. DESCRIPCIÓN DEL CASO

- **Antecedentes**

Dentro de las tecnologías de las tecnologías de machine learning para aprendimiento supervisado, el análisis de regresión nos permite estimar el valor de una variable dependiente de valor continuo (numérico) a partir de un conjunto de variables predictoras. Hay distintas etapas que debemos abordar en un análisis de regresión para ver la viabilidad de construir un modelo que explique nuestra variable dependiente y que sea capaz de predecir futuros valores con suficiente precesión. Habitualmente comenzaremos con un análisis de individual de las variables para continuar con un análisis multivariado.

En la etapa final, existen diferentes modelos que podemos testear para resolver nuestro problema desde regresiones lineales simples, regresiones múltiples, polinómicas, árboles de decisión o modelos de ensambling o incluso redes

neuronales regresoras.

Hay muchos ejemplos del mundo real donde se pueden utilizar análisis de regresión desde estimar cual va a ser la temperatura ambiental, el precio futuro de una acción, el valor de una propiedad, o el tiempo que vamos a tardar en vender un artículo.

- **Problema a resolver**

En este TFM queremos desarrollar una herramienta que le permita estimar, a potenciales vendedores de vehículos de segunda mano el precio de venta por el cual deberían anunciar sus coches mediante un aprendizaje supervisado. Trabajaremos con datos etiquetados con el precio al que se venden o se están vendiendo una serie de coches, analizaremos las variables a nuestra disposición, algunas categóricas otras numéricas, explicaremos la relación de las variables entre sí (análisis de correlación) y con la variable dependiente (precio del vehículo) y construiremos un modelo de regresión que nos permita predecir el precio, calcularemos las métricas principales que expliquen el rendimiento de nuestro modelo (r2, rmse, mape).

El alumno podrá trabajar con R, Python o Spark para la preparación de datos, análisis de los mismos y modelado.

- **Objetivos**

Los objetivos a alcanzar en este trabajo son acercar al alumno a lo que es un problema habitual que puede tener que resolver un Data Scientist como parte de su actividad laboral en todas sus etapas, desde recolección de datos, preparación, análisis y modelado.

Buscamos familiarizar al alumno con el análisis univariable y multivariable orientado al objetivo de entender una variable continúa.

Se busca también que el alumno trabaje con distintos modelos de regresión, que entienda las distintas métricas que nos permiten conocer la bondad del ajuste de los modelos y que gane confianza en la aplicación de los mismos a casos de uso reales.

- **Datos**

Respecto a los datos para el trabajo, hay dos posibilidades, para aquellos alumnos que se muestran cómodos con lenguajes de programación como Python, recomendamos que sean ellos mismos quienes consigan los datos utilizando una librería de scraping: Scrapy (scrapy.org) crawleando cualquiera de los portales públicos donde se anuncian vehículos de segunda mano, como, por ejemplo:

https://www.coches.net/segunda-mano/

https://www.autohero.com/es/search/

https://www.cazoo.co.uk/cars/bmw/

En función del tamaño del dataset, el alumnado podrá elegir una marca (i.e. Renault, BMW, …), deberá utilizar scrapy para descargar la información disponible de los vehículos de esa marca en la web elegida y prepara un dataset incluyendo al menos (modelo, versión, kilómetros, tipo de combustible y precio). Trabajará las siguientes etapas con este dataset que ha recopilado.

Para aquellos alumnos que prefieran saltarse esta etapa de recolección de datos, podrán utilizar el dataset de kaggle de vehículos usados basado en listados de craiglist.org:

https://www.kaggle.com/austinreese/craigslist-carstrucks-data