

Universidad Galileo de Guatemala.
Post grado en Análisis y Predicción de Datos.
Maestría en Ciencia de Datos.
Product Development.
FISIC.



Machine Learning as a Service.

Galileo
UNIVERSIDAD

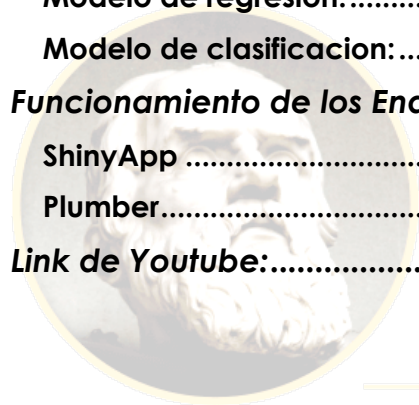
La Revolución en la Educación

Francisco Antonio Morales Hernández
Mydelin Stephanie Valladares Lima

1900 2196
1900 8841

Tabla de contenido

Explicacion de la data:	3
Regresión:.....	3
Clasificación:	3
Explicación de los modelos:	4
Modelo de regresión:.....	4
Modelo de clasificacion:.....	4
Funcionamiento de los Endpoints:	5
ShinyApp	5
Plumber.....	5
Link de Youtube:	5



Galileo
UNIVERSIDAD
La Revolución en la Educación

Explicacion de la data:

Regresión:

Para realizar el modelo de regresion, utilizamos un dataset simple que nos muestra la probabilidad de ingreso a cierta universidad, en base a ciertos parametros de evaluación.

- Variable a Predecir

Chance of Admit: Probabilidad de ser admitido en una Universidad.

- Features

GRE Score: El Graduate Record Examinations (GRE) es una prueba estandarizada y adaptativa. El examen tiene como objetivo medir el razonamiento verbal, razonamiento cuantitativo, la escritura analítica y habilidades de pensamiento crítico que se han adquirido a lo largo de un extenso período de tiempo y que no están relacionados con campo específicos de estudio. Es decir, el GRE es una prueba generalista.

TOEFL Score: Es una prueba estandarizada de dominio del idioma inglés, específicamente orientada a hablantes no nativos de este idioma, en donde se evalúa la comprension que tiene el alumno sobre esta lengua no nativa.

CGPA: Por sus siglas en inglés significa "Cumulative Grade Point Average", y hace referencia al punteo acumulado en cierto periodo de tiempo escolar (por ejemplo durante el pregrado, la secundaria, la primaria, una maestría, etc)

Clasificación:

Para la clasificación utilizaremos el dataset de Iris, en el cual se busca la clasificación de cierto tipo de flor en basea sus caracterisicas.

- Variable a predecir

Species: Tipo de especie (Iris-virginica, Iris-setosa, Iris-versicolor).

- Features

Sepal.Length: Largo de la sepa.
Sepal.Width: Ancho de la sepa.
Petal.Length: Largo del pétalo
Petal.Width: Ancho del pétalo

Nota: Tomar en cuenta que el fin del presente trabajo es el deployment y modelos para un ambiente de producción por lo que no se hace mucho énfasis en la sofisticación de los modelos.

Explicación de los modelos:

Modelo de regresión:

Regresión polinomial de primer grado: El modelo de regresión es el más simple de todos, siendo una regresión polinomial y haciendo uso de la función lm incorporada en R.

Regresión polinomial de primer grado: También implementé un modelo de regresión polinomial de grado dos. Dado que la complejidad del polinomio no es tan grande, se esperaría que aun siendo de grado dos, no se tenga overfitting, sin embargo, para evaluar el modelo a mayor detalle se deben realizar otro tipo de evaluaciones que van más allá de los objetivos del presente trabajo.

Modelo de clasificación:

Clasificación utilizando Support Vector Machine: Este algoritmo construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta.

En pocas palabras, dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos) pertenece a una categoría o a la otra.

Clasificación utilizando K nearest neighbors:

Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

En el reconocimiento de patrones, el algoritmo k-nn es usado como método de clasificación de objetos (elementos) basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos. k-nn es un tipo de aprendizaje vago (lazy learning), donde la función se aproxima solo localmente y todo el cómputo es diferido a la clasificación.

Funcionamiento de los Endpoints:

ShinyApp

En la ShinyApp unicamente se muestra el deployment de las estadísticas para cada uno de los modelos utilizados.

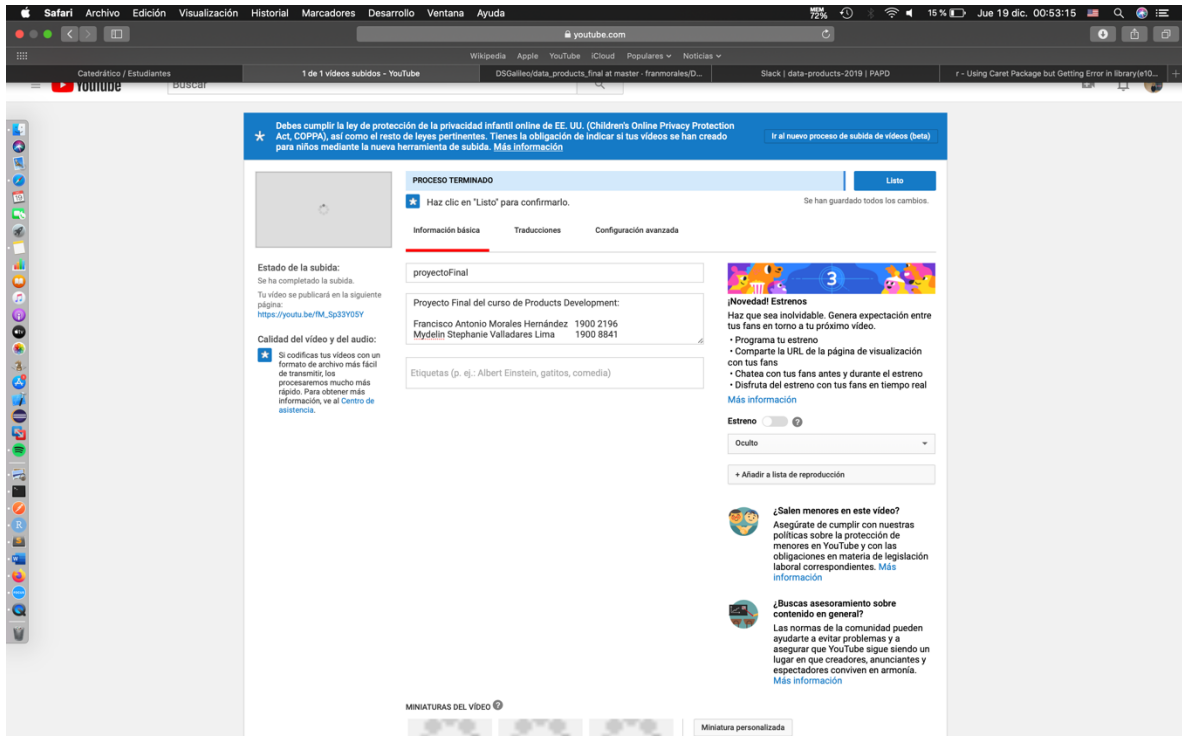
Plumber

En plumber existe una interacción entre el usuario y los modelos construidos. El usuario puede ya sea ingresar los parametros manualmente o correr un script (archivo json) desde postman para poder probar la efectividad de los modelos.

Link de Youtube y Repositorio de Github:

https://github.com/franmorales/DSGalileo/tree/master/data_products_final

https://youtu.be/fM_Sp33Y05Y



Galileo
UNIVERSIDAD
La Revolución en la Educación