Laboratorio de construcción de software

Tercera entrega - Trabajo Práctico Inicial

Universidad Nacional de General Sarmiento

Comisión 1 segundo semestre - año 2023

Docentes

Juan Carlos Monteros

Francisco Orozco De La Hoz

Leandro Dikenstein

Alumnos

Ignacio Barrientos

Lucrecia Verón

Cambio de objetivo

En un inicio para este trabajo inicial se eligió como tema el dengue, y se tenía como objetivo predecir qué zonas del país tenían más casos de esta enfermedad. Durante la investigación e implementación de los modelos de machine learning para este problema se han presentado algunos inconvenientes, estos fueron la escasez de datos, también se pensó otro objetivo como predecir qué provincias eran más propensas a que haya más dengue, pero esto involucra otros tipos de datos como datos ambientales y geográficos entre otros. Finalmente se ha decido cambiar de tema por Covid-19, con el objetivo de predecir la probabilidad de casos confirmados de Covid, ya que contamos con más datos sobre el tema, los cuales son más consistentes y se podrá trabajar mejor.

Modelo elegido y re-entrenamiento de modelo

Se han evaluado dos modelos, **regresión logística** y **regresión lineal**, de estos dos hemos elegido el modelo de **regresión logística** para realizar el entrenamiento, ya que contamos con datos categóricos y este modelo es más adecuado con este tipo de datos.

En el entrenamiento y reentrenamiento del modelo se utilizó la biblioteca Scikit-learn para poder utilizar la regresión logística.

En primer lugar se separaron en un dataframe los datos que vamos a utilizar, los cuales ya fueron limpiados previamente. Se utilizaran las columnas que nos proporcionan la información de un caso confirmado o descartado y fecha diagnostico, tambien se añadieron los datos sobre el sexo y edad de la persona, y provincia en la que reside.
Antes de realizar el entrenamiento a estos datos categóricos no numéricos se los codifica con el método **LabelEncoder** y convertirlos en datos numéricos.

Para poder realizar el entrenamiento y así analizar los resultados, creamos la instancia del modelo, los datos de entrenamiento y prueba con las instancias **x_entrenamiento, x_prueba, y_entrenamiento e y_prueba**, luego se utilizó el método **.fit()** pasando por parámetro ambas instancias de entrenamiento para poder entrenar el modelo con nuestros datos.
Con el modelo entrenado, mediante el método **.predict()** se hizo la predicción pasándole por parámetro la instancia de prueba de la variable independiente, y su resultado se guarda en la instancia de prueba de la variable dependiente.
Finalmente se evalúa la exactitud del modelo con el método **accuracy_score()** pasando como parámetros las instancias de prueba y entrenamiento de la variable dependiente.

Este modelo tiene una precisión de 0.68 con nuestros datos, por lo tanto todavía hay errores que corregir.