Estudio de Contaminantes en la Estación Obispado por Análisis Discriminante

Samuel Méndez Villegas, Salette Noemi Villalobos, Paola Montserrat Vega Ortega, Ethan Enrique Verduzco Pérez

Dra. Olivia Carrillo Gamboa y Dr. Daniel Otero Fadul

Resumen

El análisis discriminante es un algoritmo de Machine Learning el cual desarrolla un modelo de categorización para las diferentes variables. En el caso de los datos del Gobierno del Estado de Nuevo León se llevó a cabo el cálculo de pertenecer a uno de los tres niveles del contaminante Ozono, definidos de acuerdo con las normativas gubernamentales. El modelo es de clasificación, el cual supone distribuciones sobre cada variable de entrada. El análisis se llevó a cabo con el lenguaje de programación Python y la librería de scikit-learn. El fin de utilizar este análisis es poder explicar las diferencias de los grupos dentro de los datos, además de verificar que tan bien clasifican correctamente los objetos para conocer si están dentro de la norma para una buena calidad de aire.

Introducción

El análisis discriminante da a conocer a qué grupo pertenece cada supuesto. También, es usado para poder seleccionar qué variables son necesarias para la clasificación [2].
Los datos analizados están divididos en 15 categorías que representan los contaminantes de Nuevo León en la estación centro, perteneciente a la zona del Obispado. Dentro del análisis se trabajó con 9646 datos. El proceso se realizó enfocado en la variable de O3 de acuerdo a la normatividad Nacional para mantener una buena calidad de aire.

Metodología

Para poder desarrollar el análisis discriminante se hizo uso de la librería de Machine Learning de scikit-learn ”LinearDiscriminantAnalysis”[3]. Se definieron como variables independientes los 14 contaminantes de la base de datos excluyendo el O3 que fue tomado como la variable dependiente. El proceso se inició definiendo el dataset y tratando los valores erróneos para poder realizar un mejor análisis de clasificación.
En un primer instante, los datos de ozono no estaban discretizados, por lo tanto, dadas las especificaciones de la norma Nacional para poder respirar sin problemas el aire, la cual es de 95 ppb (partes por billón) se le asignó una clasificación a los datos. Se formaron 3 grupos, donde 1 representa una cantidad normal de ozono, 2 elevado y 3 sobrepasa la norma. Con ello, se pudo realizar el ajuste del modelo y definir la varianza explicada. Asimismo, se transformaron las variables para poder obtener una mejor predicción.

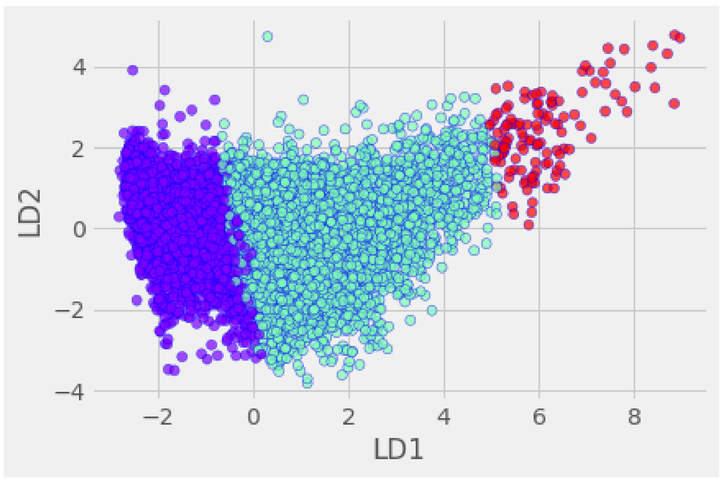

Fig. 1: Visualización de clusters.

Resultados

Una vez obtenidas las predicciones se proyectó la manera en la que los discriminantes influyen en la formación de grupos. Dicha representación se observa en la Figura 1. Se puede ver que hay un grupo conteniendo la mayoría de registros (valores elevados en azul claro) y que son mínimos los casos donde el ozono rebasa el límite establecido por la norma (puntos en color rosa). Por otro lado, se encontró que los valores normales del ozono (en morado) cuentan con menos relevancia dentro del discriminante 1, en contraste con las otras dos categorías antes explicadas. Sin embargo, las 3 categorías contienen valores dispersos dentro del segundo discriminante. Asimismo, la varianza explicada del primer y segundo discriminante fue de 0.96 y 0.035 respectivamente. Por lo tanto, podemos identificar que el primer discriminante tiene mayor poder de discriminación y de predicción en la clasificación. Finalizando, se probó el nivel de clasificación del modelo, donde se asignó correctamente el 93.6 % de los datos a partir de la función discriminante.

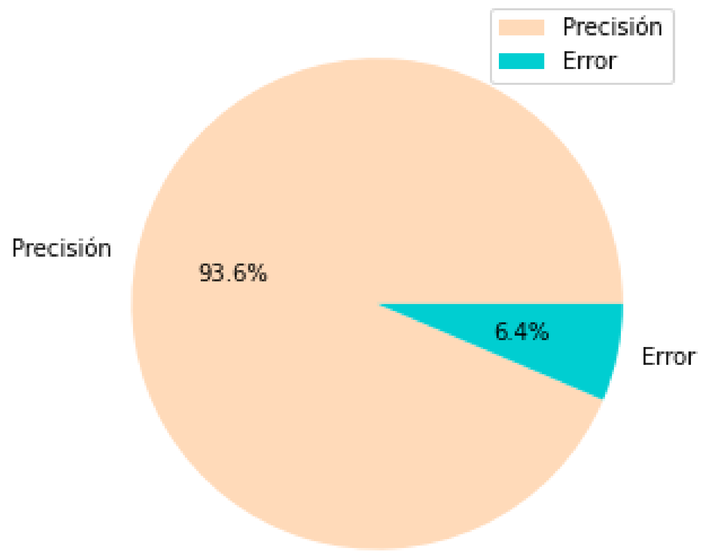

Fig. 2: Predicciones acertadas.

F1 score	Precision	Recall	Accuracy
0.927719	0.935612	0.926913	0.926913

Fig. 3: Medidas de evaluación.

La Figura 3 hace contraste de los valores predichos y actuales, de esta manera, se puede argumentar que el análisis fue bastante bueno [1].

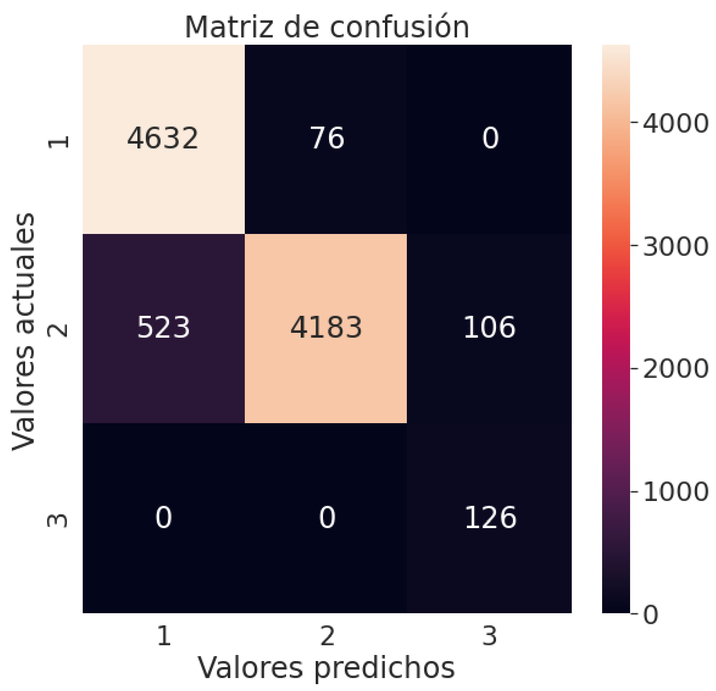

Fig. 4: Resultados de evaluación del modelo.

Conclusión

El proceso de selección de variables discriminantes puede variar, y por lo tanto, la aplicación de pruebas estadísticas. Al obtener la regla de decisión con el algoritmo, y obtenido el porcentaje de acierto en el pronóstico, se puede concluir que este proceso puede ser aplicado en ciertas investigaciones relacionadas con las variables estudiadas.
En definitiva, los resultados de la investigación son satisfactorios, ya que se obtuvo un porcentaje elevado de clasificaciones satisfactorias para el nivel de ozono.

Referencias

[1] Bharathi, M. Confusion Matrix for Multi-Class Classification. (2021). Analytics Vidhya.

[2] Mercedes Torrado-Fonseca, Vanesa Berlanga-Silvente (2013). Análisis discriminante. Revista d’Innovació i Recerca en Educació.

[3] Maklin, Cory. K-Linear Discriminant Analysis In Python. Recuperado de: https://towardsdatascience.com/linear-discriminant-analysis-in-python-76b8b17817c2