UNIVERSIDAD CENTRAL DEL ECUADOR|

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

MINERÍA DE DATOS

Nombre: Daniel Alejandro Morocho **Paralelo:** S.I.8 - 001

Fecha: 12 de julio de 2023

Tarea de Reto Predictivo"

Ejercicio 4: RETO PREDICTIVO. ***En este ejercicio usted tiene una tabla de datos SegurosV2019.csv con información sobre fraudes en seguros, esta tabla tiene 12 variables y 6413 casos, se trata de predecir la variable Fraude que indica si hubo o no fraude. Este ejercicio es un verdadero reto predictivo ya que se trata de un problema muy desbalanceado, se tienen 6146 no fraudes y a penas 267 fraudes, esto hace que sea muy difícil el aprendizaje para cualquier modelo predictivo.***

Para este ejercicio usted recibe además el archivo SegurosNuevos500V2019VE.csv en el cual la variable Fraude viene con un NA para todos sus registros. El reto consiste en predecir para este archivo los valores de la variable Fraude, para esto haga lo siguiente:

- ***Determine cuál de los modelos estudiados en el curso funciona mejor para estos datos, debe calibrar los modelos (seleccionar los mejores parámetros), por ejemplo, para Árboles debe determinar La Profundidad Máxima etc... etc... También debe seleccionar las mejores variables predictoras (que podrían ser todas).***
- ***Realice las predicciones de los individuos nuevos usando la opción Predicción Individuos Nuevos de predictoR.***

Solución:

Después de subir los datos, comprobamos que el problema está desbalanceado. (El no representa el 99.17% de los datos y el sí el 0.83% de los mismos). Entonces el modelo puede presentar dificultades porque necesitamos saber acerca de los casos de 'fraude'.

Para evaluar qué tan bien generaliza el modelo se consideran todas las variables para desarrollar los modelos de predicción. La distribución de datos para aprendizaje será de un 70% y de un 30% de los datos para prueba. Porque esta distribución se usa por lo general cuando hay y conjunto de datos grandes y se necesita una cantidad grande para entrenar y probar el modelo.

Al estar frente a un problema tan desbalanceado, necesitamos evaluar bien el rendimiento del modelo en los datos que representan la minoría (en este caso, la variable 'Fraude'), con esta distribución nos podemos asegurar de tener los datos suficientes de la variable 'Fraude' en el conjunto de los datos prueba.

Después se comparará con otra distribución parecida, para observar como se comportan los modelos,

K-vecinos más cercanos (KNN)

Empezaremos la predicción con el modelo de K-vecinos más cercanos (KNN) y todos los kernels disponibles.

Usaremos el K máximo 66, el que nos sugiere el software (se consideran 66 vecinos para la toma de decisiones). Y comenzaremos a evaluar los resultados.

Al analizar los resultados, se observa que el mejor modelo es el K vecinos más cercanos-optimal que tiene una precisión global del 96.83%, una precisión negativa del 99.617% y una precisión positiva del 18.462%.

Como el problema es desbalanceado se considera una distribución del 60% de los datos para aprendizaje y un 40% de los datos para prueba. Utilizando un K máximo de 62 (automático).

Y los resultados son los siguientes

Al analizar los resultados, se observa que el mejor modelo es el K vecinos más cercanos-biweight que tiene una precisión global del 96.673%, una precisión negativa del 99.549% y una precisión positiva del 16.092%.

Como el problema es desbalanceado se consideró una última distribución: 80% de los datos para aprendizaje y un 20% de los datos para prueba. Utilizando un K máximo de 71 (automático).

Y los resultados son los siguientes

Al analizar los resultados, se observa que el mejor modelo es el K vecinos más cercanos-gaussian y triangular que tienen una precisión global del 97.068%, una precisión negativa del 99.59% y una precisión positiva del 25.581%.

Al analizar los resultados con diferente distribución de datos de aprendizaje y prueba, se aprecia que se obtuvo mejores resultados con una distribución de 80% a 20% respectivamente. Y los modelos que mejor rendimiento tuvieron son: K vecinos más cercanos-gaussian y triangular.

Bayes

Se obtiene una precisión positiva de 30.23% lo que representa el número más alto empleado en los modelos.

Por lo tanto es el mejor método para analizar este problema desbalanceado.