Desarrollo de un clasificador Bayes Naive y una aplicación con datos del flujo vehicular en autopistas de Buenos Aires

Romero Avila, L.[1], Salas Morales, H.[1,2], Martin, R.[1], and Rossi, P.[1]

[1] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales
Intendente Güiraldes 2160 - Ciudad Universitaria, Buenos Aires, Argentina
C1428EGA - Tel. (+54) 11 5285-7400
https://exactas.uba.ar
[2] Universidad de los Andes, Mérida, Venezuela
Av. 3 Independencia entre calles 23 y 24 - Mérida, Venezuela
CP 5101 - Tel. (+58) 274-2401111
http://www.ula.ve/

luisr.romeroa@gmail.com
hecsalms@gmail.com
rodrigosmartin88@gmail.com
pau_cha@yahoo.com.ar

Resumen. Se desarrolló un modelo de clasificación para identificar los días laborables a partir del flujo vehicular en estaciones de peaje, considerando los registros de 2019 de las estaciones Illia y Alberti. Cada observación consistió en siete covariables: cuatro variables dicotómicas que identificaron cinco bloques horarios, una variable dicotómica para el sentido de circulación, una variable dicotómica para la estación de peaje, la cantidad de vehículos livianos y la cantidad de vehículos pesados, contabilizados en ambos casos para hora reloj, sentido de circulación y estación. Se definieron diez casos de estudio, considerando cada bloque horario y estación, entrenándose un clasificador de Bayes Naive que implementó la regla óptima de Bayes para la decisión de la variable respuesta. Las covariables que contabilizaron el flujo vehicular fueron modeladas como variables aleatorias continuas, estimándose su densidad a través del estimador no paramétrico de Rosenblatt-Parzen basado en núcleos gaussianos, cuya ventana se determinó por convalidación cruzada en diez iteraciones, buscando minimizar el error de clasificación. Cada uno de los estimadores finales se comparó con un estimador de regresión logística sin regularización, obteniendo un menor error de clasificación en el estimador de Bayes Naive en ocho de los diez casos estudiados.

Palabras clave: Aprendizaje Supervisado · Naive Bayes · Estimación no paramétrica · Flujo Vehicular.

1 Introducción

1.1 Justificación

Este trabajo se desarrolló en el marco de la materia optativa **Seminario Elemental de Estadística con R**, donde fue requerida la aplicación de conceptos de estimación no paramétrica de densidad, clasificación basada en la Regla Óptima de Bayes y regresión logística en un contexto práctico. Aunque se fundamentó principalmente en las clases impartidas, se buscó respaldar las afirmaciones con la bibliografía adecuada.

La motivación principal de este trabajo fue la aplicación de los conocimientos adquiridos, sin tener la intención de resolver un problema real específico. De hecho, se reconoce la existencia de herramientas potencialmente más performantes para abordar la temática en cuestión. Sin embargo, dada la problemática recurrente del tráfico vehicular en la ciudad de Buenos Aires, donde las autopistas juegan un papel cada vez más importante [1], se consideró interesante examinar esta temática en el marco indicado anteriormente.

1.2 Objetivos específicos

- Desarrollar el código fuente de un clasificador de Bayes Naive basado en la regla óptima de Bayes, que permita determinar las probabilidades posteriores de registros con covariables continuas y discretas.
- Evaluar el desempeño del clasificador propuesto para la identificación de días laborarles a partir del flujo vehicular en autopistas, considerando los registros de 2019 de las estaciones de peaje Illia y Alberti de la Ciudad de Buenos Aires.
- Comparar el desempeño del clasificador propuesto en la evaluación anterior con una implementación reconocida del modelo de regresión logística.

2 Propuesta de solución

2.1 Preprocesamiento de datos

El dataset utilizado es de acceso público y se encuentra disponible en la página web del Gobierno de la Ciudad Autónoma de Buenos Aires [2]. Consiste en los registros del año 2019 de todas las estaciones de peaje de las autopistas sobre las que la Ciudad ejerce jurisdicción y se compone de 812.153 observaciones en un registro tabular con información sobre fecha, hora de inicio de la contabilización del flujo vehicular, hora de finalización de la contabilización, tipo de vehículo, estación de peaje, sentido de circulación, forma de pago y el flujo vehicular en cantidad de vehículos por hora reloj, llamadas respectivamente, `fecha`, `hora_inicio`, `hora_fin`, `estacion`, `sentido`, `forma_pago` y `cant_pasos`.

Se observó que la cantidad de registros por cada hora reloj no era constante y que no había registros de horarios para los cuales el flujo vehicular fuese cero. De esta forma, al estar también incluidos los vehículos que no abonaron peajes

en los registros, por estar como un posible valor de la forma de pago “NO COBRADO”, “EXENTO”, “INFRACCION” y “T. DISCAPACIDAD”, se decidió asumir que la ausencia de registros para una combinación fijada de `hora`, `fecha`, `tipo_vehiculo`, `sentido` y `forma_pago`, equivalía a un flujo vehicular igual a cero.

Posteriormente, se seleccionaron los registros correspondientes a las estaciones de peaje de Illia y Alberti, objeto de este estudio. Se descartartó la variable `forma_pago` por no considerarse relevante y se determinó la cantidad de registros faltantes, obteniendo los valores mostrados en la Tabla 1.

Tabla 1. Cantidad observada de registros faltantes

`estación`	`tipo_vehículo`	`sentido`	Registros faltantes (cualquier fecha y hora)
ILLIA	LIVIANO	CENTRO	6
ILLIA	LIVIANO	PROVINCIA	0
ILLIA	PESADO	CENTRO	8
ILLIA	PESADO	PROVINCIA	10
ALBERTI	LIVIANO	CENTRO	6
ALBERTI	LIVIANO	PROVINCIA	3
ALBERTI	PESADO	CENTRO	275
ALBERTI	PESADO	PROVINCIA	195

Seguidamente, se cambió el diseño del dataset, considerando la variable relativa a la cantidad de pasos de vehículos y livianos en un único registro. Así, para `fecha`, `hora_fin`, `sentido` y `estacion` se crearon las variables `cant_pasos_livianos`, que tomó los valores de `cant_pasos` cuando `tipo_vehiculo` corresponde a “LIVIANOS” y `cant_pasos_pesados` que tomó los valores de `cant_pasos` cuando `tipo_vehiculo` corresponde a “PESADOS”. Las Tablas 2 y 3, muestran un extracto del dataset antes y después del rediseño, respectivamente, considerando un ejemplo con uno de los registros faltantes.

Tabla 2. Diseño original del dataset

	`fecha`	`hora_fin`	`estación`	`tipo_vehículo`	`sentido`	`cant_pasos`
17	2019-01-01	2	ALBERTI	LIVIANO	CENTRO	125
18	2019-01-01	2	ALBERTI	LIVIANO	PROVINCIA	332
19	2019-01-01	2	ALBERTI	PESADO	PROVINCIA	2

Dado que se pretende clasificar los días como laborables o no laborables a partir del flujo vehicular, se hizo necesario construir la variable respuesta a partir de la fecha de cada uno de los registros, asignando el valor 1 si la observación en cuestión era un día laborable y 0 en caso contrario. Se estableció que los sábados

Tabla 3. Configuración del dataset luego del rediseño

	fecha	hora_fin	estación	sentido	cant_pasos_livianos	cant_pasos_pesados
17	2019-01-01	2	ALBERTI	CENTRO	125	0
18	2019-01-01	2	ALBERTI	PROVINCIA	332	2

y domingos no eran días laborables a los efectos de la determinación del valor de la variable respuesta. También, se consideró el calendario oficial de feriados y días no laborables del año 2019, según lo establecido por el Poder Ejecutivo Nacional [3]. Es así que se consideró no laborable a los feriados nacionales y días no laborables del mencionado calendario oficial, con excepción de los días no laborables establecidos para la religiones judía e islámica y otras minorías, que fueron categorizados como laborables, excepto en los casos que coincidían con sábado o domingo.

La información contenida en `hora_fin` se agrupó en cuatro variables categóricas dicotómicas o variables *dummies*, que representaron cinco bloques horarios: madrugada, mañana, mediodía, tarde y noche, denominándose `Horario_MANANA`, `Horario_MEDIODIA`, `Horario_TARDE` y `Horario_NOCHE`. Los valores correspondientes a la categoría madrugada se obtuvieron cuando las cuatro variables categóricas anteriores eran igual a cero.

Se descartó la variable `fecha`, mientras que las variables `sentido` y `estación` se transformaron en variables categóricas como sigue:

- Se definió la variable `Estacion_ILLIA`, que tomó el valor 1 si `estacion` correspondía a "ILLIA" y 0 si `estacion` correspondía a "ALBERTI".
- Se definió la variable `sentido_PROVINCIA`, que tomó el valor 1 si `sentido` correspondía a "PROVINCIA" y 0 si `sentido` correspondía a "CENTRO".

La estructura final del dataset consistió en 7300 observaciones y 9 variables en total, una de las cuales era la variable respuesta. Las variables descriptoras fueron entonces: `Estacion_ILLIA`, `sentido_PROVINCIA`, `Horario_MANANA`, `Horario_MEDIODIA`, `Horario_NOCHE`, `Horario_TARDE`, `cant_pasos_livianos` y `cant_pasos_pesados`.

2.2 Fundamentación de la implementación desarrollada

Se define $(\mathbb{N})_2 = \{0, 1\} \subset \mathbb{N}$. Sea $\aleph_i = [x_{i1}, \ldots, x_{ip}]$ la i-ésima realización de un vector aleatorio $\mathbb{X} = [X_1, \ldots, X_p] \in (\mathbb{N})_2^p$ e $Y_i \in (\mathbb{N})_2$ la variable respuesta asociada a la realización anterior. A los fines de aplicar la regla óptima de Bayes [4] se quiere deducir si

$$\mathbb{P}(Y_i = 1 | \mathbb{X} = \aleph_i) \geqslant \mathbb{P}(Y_i = 0 | \mathbb{X} = \aleph_i) \tag{1}$$

Para ello, sea en general $y \in \{0, 1\}$ una realización de Y_i. Entonces, por Regla de Bayes [5] se conoce que:

$$\mathbb{P}(Y_i = y | \mathbb{X} = \aleph_i) = \frac{p_{(\mathbb{X}|Y_i=y)}(\aleph_i)\mathbb{P}(Y_i = y)}{p_{\mathbb{X}}(\aleph_i)} \tag{2}$$

Pero, asumiendo independencia de las covariables resulta que la probabilidad conjunta puede ser escrita como:

$$p_{(\mathbb{X}|Y_i=y)}(\mathbb{x}_i) = \prod_{j=1}^{p} p_{(X_j|Y_i=y)}(x_{i,j}) \tag{3}$$

esta es la Hipótesis Ingenua, característica del modelo de Bayes Naive [6]. Luego, dado que la probabilidad puntual condicional a $y \in \{0,1\}$ es:

$$p_{(X_j|Y_i=y)}(x_{i,j}) = \frac{\mathbb{P}\left(\{X_j = x_{i,j}\} \cap \{Y_i = y\}\right)}{\mathbb{P}(Y_i = y)} \tag{4}$$

se deduce que es posible, por teorema de la probabilidad total [5], determinar la probabilidad marginal:

$$p_{\mathbb{X}}(\mathbb{X}_i) = \mathbb{P}(Y_i = 0) \prod_{j=1}^{p} p_{(X_j|Y_i=0)}(x_{i,j}) + \mathbb{P}(Y_i = 1) \prod_{j=1}^{p} p_{(X_j|Y_i=1)}(x_{i,j}) \tag{5}$$

conociendo las probabilidades a priori y $\mathbb{P}\left(\{X_j = x_{i,j}\} \cap \{Y_i = y\}\right), 1 \leqslant j \leqslant p$.

Hasta acá se ha asumido que las covariables X_j son variables aleatorias discretas. Si alguna resulta ser una variable aleatoria continua, basta sustituir $p_{(X_j|Y_i=y)}(x_{i,j})$ por la densidad $f_{(X_j|Y_i=y)}(x_{i,j})$[7]. Así, resulta posible determinar una estimación de las probabilidades posteriores indicadas en (1), estimando no-paramétricamente las probabilidades a priori, conjuntas y marginales.

En efecto, si se define $\mathcal{I}_y = \{i : y_i = y\}$ el conjunto formado por los índices tales que su correspondiente observación de la variable respuesta es igual a $y \in \{0,1\}$ e $\mathbb{x}^* = [x_1^*, \ldots, x_p^*]$ un nuevo vector de realizaciones de $\mathbb{X}$; en el caso de las covariables con distribuciones discretas, como $p_{(X_j|Y_i=y)}(x_j^*)$ es tal que $X_j, x_j^* \in (\mathbb{N})_2$, se puede utilizar el estimador:

$$\widehat{\mathbb{P}}\left(\{X_j = x_j^*\} \cap \{Y_i = y\}\right) = \begin{cases} \frac{1}{\#\mathcal{I}_y} \sum_{i \in \mathcal{I}_y} x_{i,j}, & x_j^* = 1 \\ 1 - \frac{1}{\#\mathcal{I}_y} \sum_{i \in \mathcal{I}_y} x_{i,j}, & x_j^* = 0 \end{cases} \tag{6}$$

para estimar $\mathbb{P}\left(\{X_j = x_j^*\} \cap \{Y_i = y\}\right)$ pues es un estimador consistente por Ley de los Grandes Números [8]. Por este mismo motivo, es posible utilizar como estimador de las probabilidades a priori:

$$\widehat{\mathbb{P}}\left(Y = y\right) = \frac{1}{\#\mathcal{I}_y} \sum_{i \in \mathcal{I}_y} x_{i,j} = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_{i,j}, & y = 1 \\ 1 - \frac{1}{n} \sum_{i=1}^{n} x_{i,j}, & y = 0 \end{cases}$$

donde n es la cantidad total de observaciones con las que se entrena el modelo, siendo que $n = \#\mathcal{I}_1 + \#\mathcal{I}_0$.

Luego, como para estimar cada una de las densidades de las variables aleatorias continuas se puede utilizar el estimador no paramétrico de la densidad de Rosenblatt-Parzen basado en núcleos K gaussianos [8]:

$$\widehat{f}_{(X_j|Y_i=y)}(x_j^*) = \frac{1}{(\#\mathcal{I}_y)h} \sum_{k \in \mathcal{I}_y} K\left(\frac{x_j^* - x_{k,j}}{h}\right) \tag{7}$$

donde $x_{k,j}$ son los valores que toma la variable X_j en $\left[x_j^* - h, x_j^* + h\right]$, una vecindad reducida a los casos en $k \in \mathcal{I}_y$, es decir, los casos en los que $Y_k = y$; se deduce entonces que es posible estimar las probabilidades puntuales, conjuntas, marginales y posteriores, considerando las ecuaciones (4), (3), (5), (2), respectivamente, aplicando el principio *plug-in* [8]. Así, resulta razonable que utilizando el modelo dado en (8) puede predecirse la variable respuesta de $\mathbb{x}^*$ a partir de una estimación de la Regla Óptima de Bayes.

$$g(\mathbb{x}^*) = \begin{cases} 1, & \widehat{\mathbb{P}}(Y = 1 | \mathbb{X} = \mathbb{x}^*) \geqslant \widehat{\mathbb{P}}(Y = 0 | \mathbb{X} = \mathbb{x}^*) \\ 0, & \texttt{otherwise} \end{cases} \tag{8}$$

En principio, en la ecuación (7) la variable h no tiene un valor definido, por lo que resulta ser un hiperparámetro del modelo. De hecho, si se considera más de una variable predictora como variable continua, existen tantos hiperparámetros como variables continuas. En el presente se refiere a dichos hiperparámetros como ventana o ancho de ventana indistintamente, dado el nombre que suele utilizarse en el contexto del estimador de núcleos de Rosenblatt-Parzen. A diferencia de las variables discretas, cuyas probabilidades se calcularon por funciones propias, las estimaciones de densidades fueron calculadas por medio de la función `density` de la biblioteca `stats` de `R` [9], considerando que si se quiere utilizar como ventana h_0 es preciso tomar un ancho $h = h_0/R$, siendo $R = \sqrt{3} - 3.06 \cdot 10^{-6}$.

2.3 Aplicación del modelo y evaluación

Exploratoriamente se decidió separar los registros en casos o grupos de estudio según el bloque horario y la estación de peaje. Para la implementación, se obtuvieron diez casos de estudio y diez modelos a entrenar, correspondientes a cinco bloques horarios y dos estaciones, utilizando como variables predictoras `sentido_PROVINCIA`, `cant_pasos_pesados` y `cant_pasos_livianos`. Las variables `Horario_MANANA`, `Horario_MEDIODIA`, `Horario_NOCHE`, `Horario_TARDE` y `Estacion_ILLIA` se utilizaron únicamente para identificar el caso de estudio.

Las variables predictoras `cant_pasos_pesados` y `cant_pasos_livianos` se modelaron como variables aleatorias continuas, definiendo para cada una el hiperparámetro h referido al ancho de la ventana. Se llevó a cabo una evaluación mediante convalidación cruzada en diez *folds* o iteraciones [4] para determinar los valores óptimos de h a partir de una grilla dada, buscando minimizar el error L de clasificación:

$$L(\mathbb{X}) = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g(\mathbb{x}_i)\right)^2 = \frac{1}{n} \sum_{i=1}^{n} I_{\{Y_i \neq g(\mathbb{x}_i)\}} \tag{9}$$

donde n es la cantidad de observaciones correspondientes al dataset $\mathbb{X}$ e $I(\cdot)$ es la función indicadora del evento $(\cdot)$. Se decidió utilizar esta métrica pues resulta conocido que la regla óptima de Bayes minimiza en promedio este error [4].

La evaluación utilizó un diseño factorial [10] para determinar la combinación óptima de los hiperparámetros de una doble grilla inicial centrada en las ventanas de Silverman [11] correspondientes a la totalidad de observaciones de la

estación Illia, la cual fue ajustada para incluir valores cercanos a cero. Fue realizado igualmente un análisis gráfico complementario que tuvo como finalidad observar el comportamiento del error de clasificación versus los valores del ancho de ventana para determinar un rango sobre el cual refinar los resultados.

El conjunto total de los registros se dividió en un dataset de entrenamiento $\mathbb{X}_{\texttt{train}}$ y un dataset de testeo $\mathbb{X}_{\texttt{test}}$ con el 70 y el 30% de los registros totales, respectivamente. No obstante, para determinar los valores finales de los hiperparámetros a utilizar en cada caso de estudio, se realizó la convalidación cruzada en diez iteraciones con el 85% de las observaciones del conjunto de entrenamiento correspondiente a cada caso de estudio $\left(\mathbb{X}_{\texttt{cv}}^{(k)}, 1 \leqslant k \leqslant 10\right)$, repitiendo la estrategia descrita en el párrafo anterior. Luego, se evaluó el error de clasificación tanto en este conjunto de datos como en el 15% restante $\left(\mathbb{X}_{\texttt{v}}^{(k)}, 1 \leqslant k \leqslant 10\right)$, a los fines de determinar si la diferencia de la métrica entre ambos conjuntos de datos era mayor al 10%. En aquellos casos donde la diferencia era mayor al 10%, se consideró que el modelo estaba sobreajustado y se decidió utilizar el par de ventanas donde se alcanzaba el siguiente mínimo local.

Finalmente, se calculó el error de clasificación para el conjunto de prueba $\mathbb{X}_{\texttt{test}}$ mediante el ensamble de los diez modelos previamente entrenados. Los resultados obtenidos se compararon con la probabilidad a priori de la clase negativa, que representa el error de clasificación del modelo simplificado que asigna a todos los registros la clase mayoritaria. También, se comparó este modelo simplificado con el modelo de Bayes Naive resultante de restringir los registros a los casos de estudio más efectivos, pues se consideró que para discriminar si un día perteneciente a un conjunto de registros dado era laborable o no, bastaba seleccionar el caso de estudio más efectivo de cada estación, esto es, el bloque horario en el que se obtuviera un menor error de clasificación.

2.4 Comparación del desempeño del modelo

Aunque el modelo de regresión logística es paramétrico [4], se eligió utilizarlo como criterio de comparación ya que fue uno de los modelos de clasificación más comentados en la asignatura en la que se contextualizó el presente trabajo. Concretamente se utilizó el modelo `glm` de la biblioteca `Stats` de `R` [12]. Para cada caso de estudio, se entrenó un modelo de regresión logística utilizando los mismos conjuntos de datos $\mathbb{X}_{\texttt{cv}}^{(k)}$ que se utilizaron para determinar los hiperparámetros del ancho de ventana y que equivalían al 85% de los registros del dataset de entrenamiento. Se calculó el error de clasificación para los conjuntos de entrenamiento del modelo de regresión, así como para los conjuntos $\mathbb{X}_{\texttt{v}}^{(k)}$, que comprendían el 15% restante de las observaciones de los diez casos de estudio. A los fines de la comparación de los dos modelos, se consideró que el modelo no paramétrico de Bayes Naive tenía un error de clasificación menor que el modelo de regresión logística si los errores de clasificación de los conjuntos $\mathbb{X}_{\texttt{v}}^{(k)}$, eran menores que los errores de clasificación de los mismos conjuntos utilizando el modelo de regresión logística.

3 Resultados y conclusiones

3.1 Búsqueda de hiperparámetros

Como se explicó anteriormente, para la determinación de los hiperparámetros de ancho de ventana se partió de un entorno centrado en la ventana de Silverman para cada una de las covariables `cant_pasos_livianos` y `cant_pasos_pesados` considerando los registros de la estación Alberti de todos los bloques horarios, determinándose los valores de h donde se minimiza el error de clasificación mediante convalidación cruzada. En la Tabla 4 se muestra un resumen sobre los parámetros de la doble grilla inicial y los resultados de error de clasificación obtenidos.

Tabla 4. Parámetros utilizados para la evaluación de la doble grilla inicial de ventanas de `cant_pasos_livianos` y `cant_pasos_pesados` y mínimo error encontrado

Livianos	Ventana de Silverman	1575.3
	Extremo izquierdo de la grilla	275.3
	Extremo derecho de la grilla	2675.3
	Paso de la grilla	200
	Ventana que minimiza el error	475.3
Pesados	Ventana de Silverman	65.3
	Extremo izquierdo de la grilla	20.3
	Extremo derecho de la grilla	120.3
	Paso de la grilla	10
	Ventana que minimiza el error	40.33
Error	0.2497	

Dado el valor de ancho de ventana $h_{\texttt{min}} = 475.3$ donde se alcanza el mínimo correspondiente a la covariable `cant_pasos_livianos` y de la consideración de la Figura 1, donde se observa que el error de clasificación es en general creciente en función de h; se decidió utilizar una grilla para los casos de estudio tal que su valor máximo fuera $h = 1000$ y ampliar el extremo izquierdo a valores más cercanos a cero, pues no se observan los intervalos de decrecimiento característicos del sobreajuste.

Por su parte, en cuanto respecta a la covariable `cant_pasos_pesados`, al obtenerese $h_{\texttt{min}} = 40.33$ y observarse la Figura 2 que el comportamiento del error de clasificación respecto al valor del ancho de la ventana no pareciera estar bien definido, se decidió no reducir el extremo derecho de la grilla utilizada para cada uno de los casos de estudio, ampliando el extremo izquierdo a valores más cercanos a cero, pues nuevamente no parece observarse los intervalos de decrecimiento característicos.

De esta forma, en cada caso de estudio se decidió estudiar el comportamiento del error de clasificación en el intervalo $[100, 1000]$ para el ancho de ventana de `cant_pasos_livianos` y $[5, 120]$ para el ancho de ventana de `cant_pasos_pesados`.

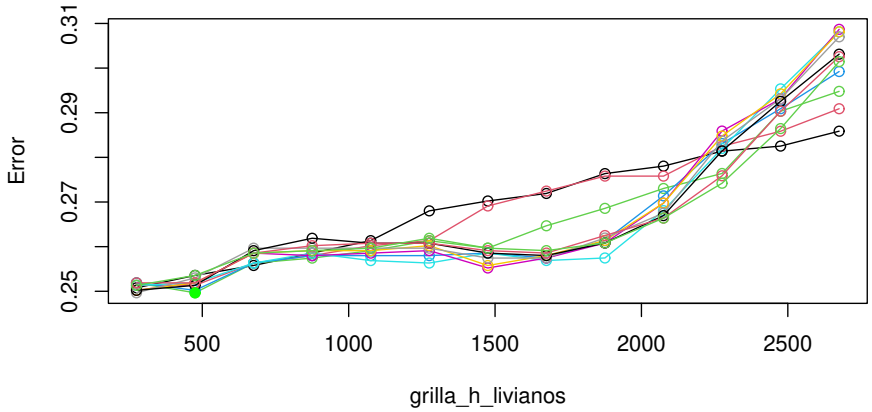

Fig. 1. Error de clasificación en función del ancho de ventana correspondiente a vehículos livianos de la estación Illia en todos los horarios, mostrados en curvas de nivel para ancho de ventana de vehículos pesados. El ancho de ventana se mide cantidad de vehículos por hora. Un punto sólido verde indica el mínimo local seleccionado.

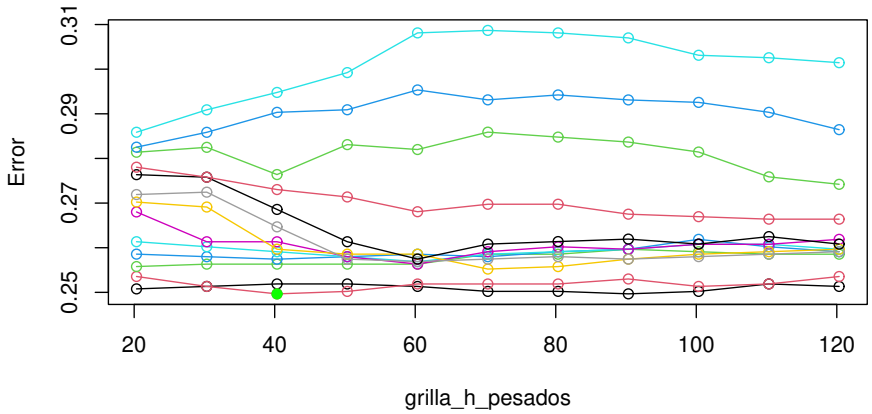

Fig. 2. Error de clasificación en función del ancho de ventana correspondiente a vehículos pesados de la estación Illia en todos los horarios, mostrados en curvas de nivel para ancho de ventana de vehículos livianos. El ancho de ventana se mide en cantidad de vehículos por hora. Un punto sólido verde indica el mínimo local seleccionado.

La Tabla 5 presenta las ventanas del estimador de la densidad de Rosenblatt-Parzen que minimizan el error de clasificación para cada uno de los casos de estudio, luego de su determinación por convalidación cruzada.

Tabla 5. Resultados de las ventanas que minimizan el error para cada caso de estudio

Estación	Variable	Mañana	Mediodía	Tarde	Noche	Madrugada
ILLIA	`cant_pasos_livianos`	400	945	400	950	815
	`cant_pasos_pesados`	30	15	105	950	815
ALBERTI	`cant_pasos_livianos`	400	725	600	400	30
	`cant_pasos_pesados`	55	20	45	105	3

Puede verse que los casos correspondientes a los bloques horarios Mediodía y Noche de la estación Illia presentaron un ancho de ventana cercano al extremo superior de la grilla de `cant_pasos_livianos`. Por ello, se observó gráficamente el comportamiento del error de clasificación en un entorno del máximo de la grilla. En la Figura 3 se muestran las curvas de nivel del error de clasificación versus el ancho de ventana para el bloque mediodía correspondiente a `cant_pasos_livianos`, mientras que en la Figura 4 muestra lo propio para la variable `cant_pasos_pesados`.

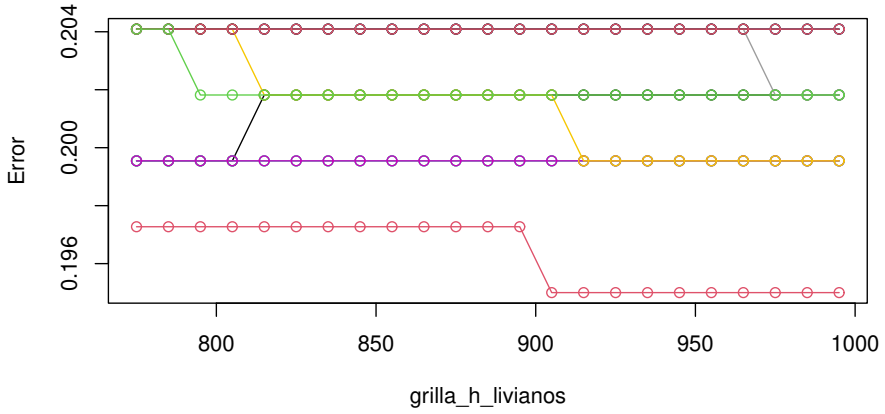

Fig. 3. Error de clasificación en función del ancho de ventana correspondiente a vehículos livianos de la estación Illia en el bloque horario Mediodía, mostrados en curvas de nivel para ancho de ventana de vehículos pesados. El ancho de ventana se mide en cantidad de vehículos por hora.

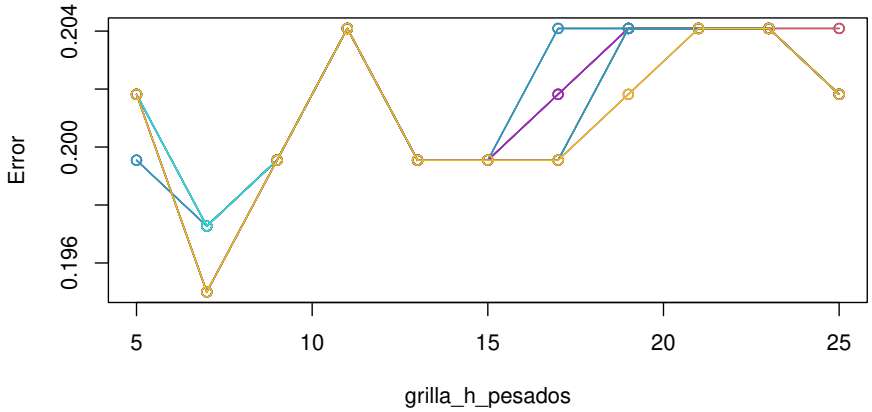

Fig. 4. Error de clasificación en función del ancho de ventana correspondiente a vehículos pesados de la estación Illia en el bloque horario Mediodía, mostrados en curvas de nivel para ancho de ventana de vehículos livianos. El ancho de ventana se mide en cantidad de vehículos por hora.

Se observó que fijado el valor del ancho de ventana de `cant_pasos_pesados`, el error de clasificación tiende a mantenerse para valores superiores del ancho de ventana de `cant_pasos_livianos`. No obstante, en la Figura 4 puede observarse que, si en cambio se fija el ancho de ventana de `cant_pasos_livianos`, el error de clasificación sí tiene mínimos claros. Tomando el mínimo local $h_{\texttt{pesados}} = 7$, se notó que los valores de $h_{\texttt{livianos}} \in \{905, 915, \ldots, 985, 995\}$ tenían un mismo valor de error de clasificación. Por tanto, se seleccionó el valor medio de $h_{\texttt{livianos}} = 945$, notándose que al evaluar el error en los dataset $\mathbb{X}_{\texttt{cv}}^{(1)}$ y $\mathbb{X}_{\texttt{v}}^{(1)}$ la diferencia de los errores era de 0.0795. Así, al ver que existía otro mínimo local en $h_{\texttt{pesados}} = 15$ y $h_{\texttt{livianos}} = 945$, exploratoriamente se decidió evaluar nuevamente la diferencia de los errores entrenando al modelo con estos anchos de ventana, obteniendo una diferencia de 0.0252. Un procedimiento similar se siguió para el caso correspondiente al bloque horario Noche, siendo la Figura 5 correspondiente al comportamiento del error de clasificación fijado `cant_pasos_pesados`.

3.2 Evaluación del desempeño del modelo

Definidos los valores de las ventanas, se determinó el error de clasificación para los conjuntos $\mathbb{X}_{\texttt{cv}}^{(k)}$ y $\mathbb{X}_{\texttt{v}}^{(k)}$, $1 \leqslant k \leqslant 10$, mostrándose los resultados obtenidos en la Tabla 6.

Realizado el ensamble de los diez modelos, se comparó el error de clasificación del modelo evaluado en $\mathbb{X}_{\texttt{train}}$ y $\mathbb{X}_{\texttt{test}}$ y la probabilidad a priori de la clase negativa. En la Tabla 7 se muestran los valores obtenidos.

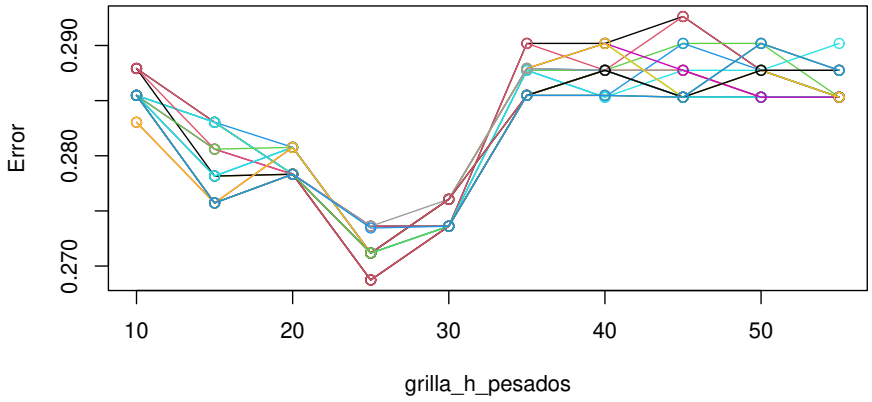

Fig. 5. Error de clasificación en función del ancho de ventana correspondiente a vehículos pesados de la estación Illia bloque horario Noche, mostrados en curvas de nivel para ancho de ventana de vehículos livianos. El ancho de ventana se mide en cantidad de vehículos por hora.

Tabla 6. Error de clasificaión del modelo de Bayes Naive para cada caso de estudio

Estación	Dataset	Mañana	Mediodía	Tarde	Noche	Madrugada
Illia	$\mathbb{X}_{\texttt{cv}}$	0.2005	0.1927	0.1778	0.2639	0.2289
$1 \leqslant k \leqslant 5$	$\mathbb{X}_{\texttt{v}}$	0.1974	0.2179	0.1899	0.3562	0.2125
Alberti	$\mathbb{X}_{\texttt{cv}}$	0.2084	0.1903	0.1844	0.2602	0.1923
$6 \leqslant k \leqslant 10$	$\mathbb{X}_{\texttt{v}}$	0.24	0.2000	0.2405	0.3013	0.2703

Tabla 7. Comparación del error del modelo de Bayes Naive con un modelo trivial

Dataset	Error	$\widehat{\mathbb{P}}(Y_i = 0)$
$\mathbb{X}_{\texttt{train}}$	0.2564	0.3272
$\mathbb{X}_{\texttt{test}}$	0.2643	0.3416

Se observa que el modelo propuesto tiene un mejor desempeño que el clasificador trivial que le asigna la clase mayoritaria a todas las observaciones. No obstante, si se toma el subconjunto de observaciones de los casos para los cuales el modelo desarrollado tiene un mejor desempeño, se obtienen los resultados indicados en la Tabla 8.

Tabla 8. Resultados del Modelo de Bayes para la determinación de los días laborables

Dataset	Restricción	Error	$\widehat{\mathbb{P}}(Y_i = 0)$
$\mathbb{X}_{\texttt{train}}$	Estación Illia Tarde	0.1776	0.3195
$\mathbb{X}_{\texttt{test}}$	Estación Illia Tarde	0.2040	0.3632
$\mathbb{X}_{\texttt{train}}$	Estación Alberti Mediodia	0.2256	0.3271
$\mathbb{X}_{\texttt{test}}$	Estación Alberti Mediodia	0.1970	0.3434

De estos valores, puede evidenciarse que la implementación del modelo en los bloques horarios específicos indicados tiene un mejor desempeño que su implementación general sin la restricción horaria, por lo que si el objetivo fuese determinar si un día es laborable a partir del flujo vehicular y se cuenta con todos los registros horarios, basta tomar las observaciones del horario Tarde en el caso de la estación Illia y las observaciones del horario mediodía de la estación Alberti para obtener el mejor desempeño del modelo.

3.3 Comparación del desempeño del modelo desarrollado con regresión logística

Finalmente, la Tabla 9 muestra los valores del error de clasificación obtenidos a partir del modelo de regresión logística entrenado con los datasets $\mathbb{X}_{\texttt{v}}^{(k)}, 1 \leqslant k \leqslant 10$, según se describió anteriormente. Se incluye también una comparación de los mismos con los correspondientes al modelo de Bayes Naive, de forma tal que dicha comparación vale 1 si el error $L(\mathbb{X}_{\texttt{v}}^{(k)})$ del modelo de Bayes Naive es menor que el equivalente al modelo de regresión logística y 0 en caso contrario.

Tabla 9. Resultados del error de clasificación de cada caso de estudio implementando un modelo de regresión logística.

Estación	Dataset	Mañana	Mediodía	Tarde	Noche	Madrugada
Illia	$\mathbb{X}_{\texttt{cv}}$	0.2238	0.1995	0.1889	0.2857	0.2244
$1 \leqslant k \leqslant 5$	$\mathbb{X}_{\texttt{v}}$	0.2239	0.2308	0.2025	0.3288	0.2250
Comparación BN		1	1	1	0	1
Alberti	$\mathbb{X}_{\texttt{cv}}$	0.2100	0.1947	0.1911	0.2771	0.2163
$6 \leqslant k \leqslant 10$	$\mathbb{X}_{\texttt{v}}$	0.24	0.2250	0.2405	0.1644	0.2973
Comparación BN		1	1	1	0	1

Se observó que el modelo de Bayes Naive presenta un error de clasificación menor respecto a Regresión Logística en ocho de los diez casos de estudio.

3.4 Conclusiones

A lo largo del presente, se introdujo un modelo de clasificación que asume independencia de las variables predictoras, basado en la Regla Óptima de Bayes y en el estimador no paramétrico de la densidad de Rosenblatt-Parzen, conceptos estudiados en la materia donde se enmarcó el presente. Se considera destacable que la implementación desarrollada permite combinar variables predictoras categóricas y continuas y que no asume una distribución para las segundas, a diferencia de los reconocidos modelos `naiveBayes` de la biblioteca `e1071` [13] de `R` que asume distribución normal de los datos continuos; o las familias de modelos de Bayes Naive de la biblioteca `Scikit-learn` [14] de `Python`, que tiene la distribución normal como única distribución continua disponible para el modelado.

De esta forma, basándose en los registros de 2019 de la estación Illia y Alberti de la ciudad de Buenos Aires, se concluye que es posible determinar los días laborables utilizando el flujo vehicular en autopistas con un error de clasificación de aproximadamente del 20% mediante el uso del clasificador de Bayes Naive desarrollado. En comparación, se observó que un clasificador trivial que asigna la clase mayoritaria (es decir, los días laborables) a todas las observaciones presenta un menor rendimiento, y que un modelo de regresión logística sin regularización presentó un rendimiento inferior en 8 de 10 casos de estudio.

References

1. Blanco, J., San Cristóbal, D.: Reestructuración de la red de autopistas y metropolización en Buenos Aires. Revista Iberoamericana de Urbanismo **8**(7), 73-88 (2012)
2. Flujo Vehicular por Unidades de Peaje AUSA, `https://data.buenosaires.gob.ar/dataset/flujo-vehicular-por-unidades-de-peaje-ausa`. Last accessed 10 May 2023
3. Feriados 2019: Conocé el calendario completo, `https://www.argentina.gob.ar/noticias/feriados-2019-conoce-el-calendario-completo`. Last accessed 10 May 2023
4. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning with application in R. Springer Science & Bussiness Media, New York, EE.UU. (2013)
5. Lock, R., et al.: Unlocking the power of data. John Wiley & Sons, Inc., New Jersey, EE.UU. (2013)
6. Murphy, K.: Machine Learning A Probabilistic Perspective. The MIT Press, Cambridge, Massachusetts, EE.UU. (2012)
7. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. 2nd edn. Springer Science & Business Media, EE.UU.
8. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer Science & Business Media, EE.UU. (2004)

9. Density: kernel density estimation, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/density. Last accessed 10 May 2023
10. Hernández, R., Fernández, C., Baptista, M.: Metodología de la investigación. 5ta edn. McGraw-Hill & Interamericana Editores, S.A. DE C.V, México D.F., México. (2010)
11. Scott, D.: Multivariate density stimation. 2nd edn. John Wiley & Sons, Inc., New Jersey, EE.UU. (2004)
12. glm: fitting generalized linear models - RDocumentation, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm. Last accessed 10 May 2023
13. naiveBayes function - RDocumentation, https://www.rdocumentation.org/packages/e1071/versions/1.7-13/topics/naiveBayes. Last accessed 28 Jun 2023
14. 1.9. Naive Bayes - scikit-learn 1.1.2. documentation, https://scikit-learn.org/stable/modules/naive_bayes.html#naive-bayes. Last accessed 28 Jun 2023