Universidad Nacional de La Matanza

DIIT
Departamento de Ingeniería e Investigaciones Tecnológicas

Ciencia de datos

Métricas

+

Naive Bayes

Agenda

- Repaso
  - Caso smokers
  - Importancia de variables y KNN
  - Algunas métricas de evaluación de clasificadores
  - Threshold
- Complejidad y tamaño de la muestra
  - ¿Tengo suficientes datos para el fenómeno que intento modelar?
  - Curvas de aprendizaje
- El problema de las clases raras
- Naive Bayes
  - Conceptos generales del clasificador
  - Naive Bayes con titanic
  - Naive Bayes en clasificación de texto
- Clasificación multiclase
  - Transformamos en un problema de clasificación binario

DIIT

Caso Smokers

Repaso y breve análisis

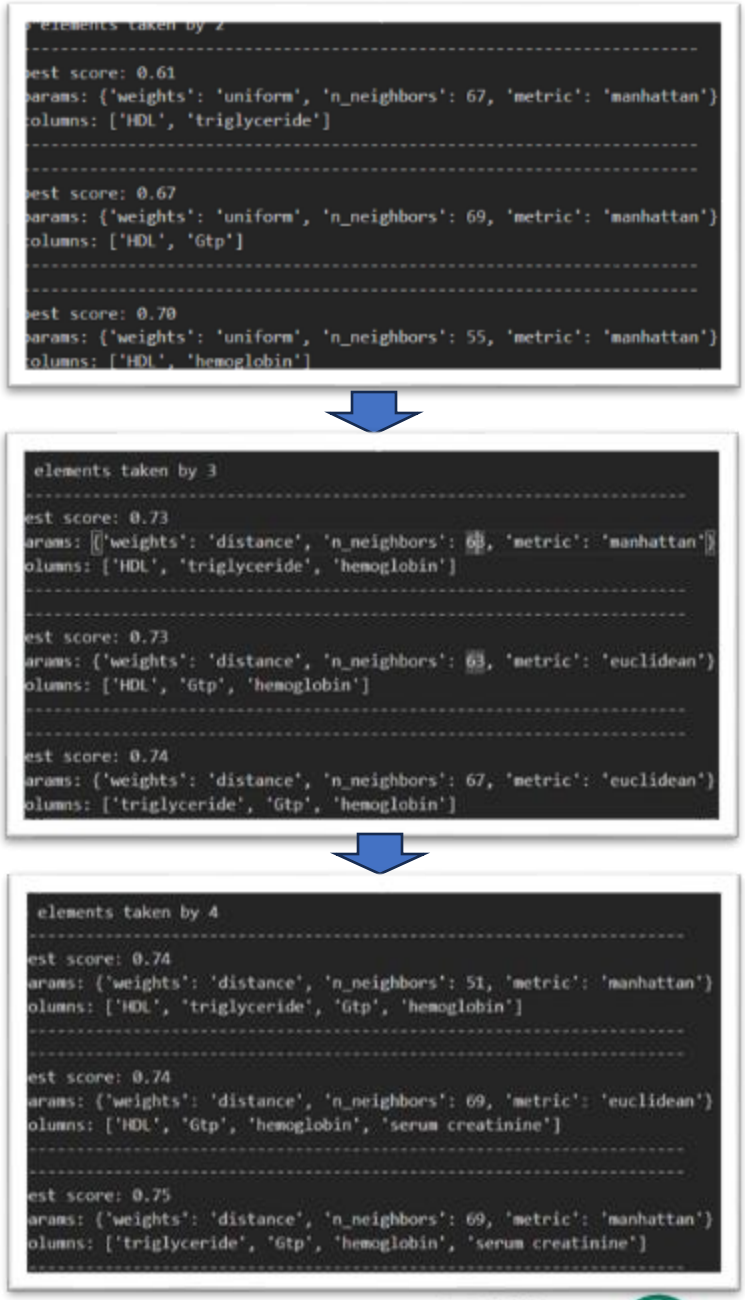
Máximo accuracy logrado con KNN = 0.75 – 0.77

Correlación y clasificación

- Dos variables predictoras fuertemente correlacionadas sugieren que ambas presentan la misma información y son candidatas a:
  - Eliminar alguna de ambas
  - Realizar una fusión manual (Ej: BMI)
  - Realizar una fusión mediante alguna herramienta matemática como (PCA).
- La correlación entre predictora y target puede ser una buena noticia, pero la no correlación no indica que la variable no sea predictora, pueden existir fuertes asociaciones no lineales.
- **Las variables pueden tener asociaciones no lineales, la valiosa herramienta de análisis de correlación no es determinante en la capacidad predictiva.**
- CorrelationVsClasification.ipynb

Caso smokers

- Respaldando la teoría: se comprueba que más columnas no ofrecen mejores resultados.
- Se analizan vía búsqueda exhaustiva la mejor combinación de columnas
- Los mejores resultados se obtienen con 3 o 4 columnas como máximo.
- A medida que se suman columnas no se obtiene un mejor score.
- ¿Qué hago si obtengo un score similar utilizando 3 y 5 columnas?
  - Principio de parsimonia o “Navaja de Ockham”
  - “En igualdad de condiciones, la explicación más simple suele ser la más probable”
- ¿Akaike Information Criterion (AIC)?
  - Equilibrio entre ajuste y simplicidad

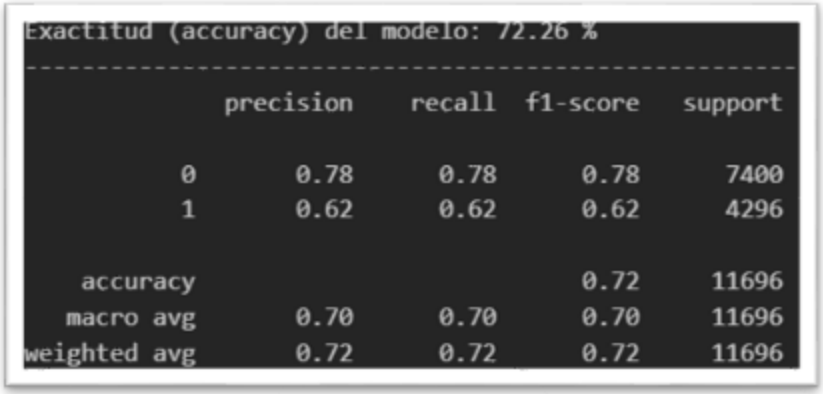


Métricas de evaluación

Mas allá de la exactitud

Métricas (Caso smokers)

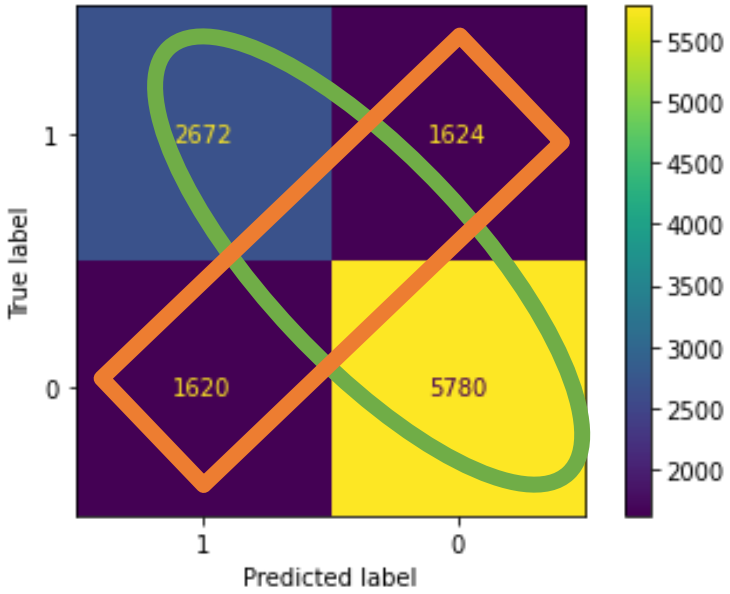
- Tenemos una exactitud del 72%
- Pero la precisión para la clase “1” es 0.62
- El recall para la clase “1” es 0.62
- Tengo buena capacidad (0.78) para decir quien no fuma, pero ¿Es eso lo que me interesa?



Exactitud (accuracy) del modelo: 72.26 %

	precision	recall	f1-score	support
0	0.78	0.78	0.78	7400
1	0.62	0.62	0.62	4296
accuracy			0.72	11696
macro avg	0.70	0.70	0.70	11696
weighted avg	0.72	0.72	0.72	11696

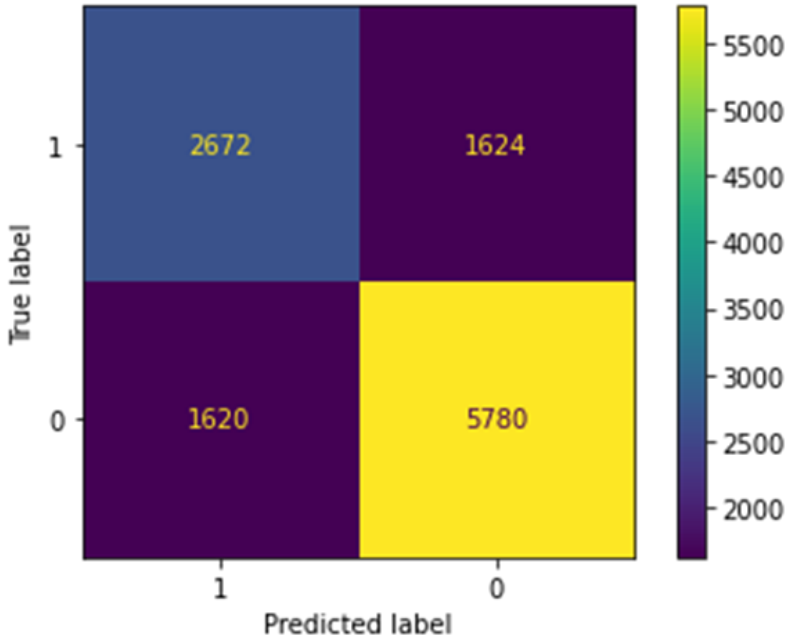
Clasificación – Matriz de confusión



Métrica	Calculo	Valor
Accuracy (Exactitud)	(TP+TN)/ (Todos)	0.72
Precisión(Precision)	TP/(TP+**FP**)	2672/(2672+1620)=0,62
Sensibilidad(Recall)	TP/(TP+**FN**)	2672/(2672+1624)=0,62
Especificidad	TN/(TN+FP)	5780/(5780+1620)=0,78

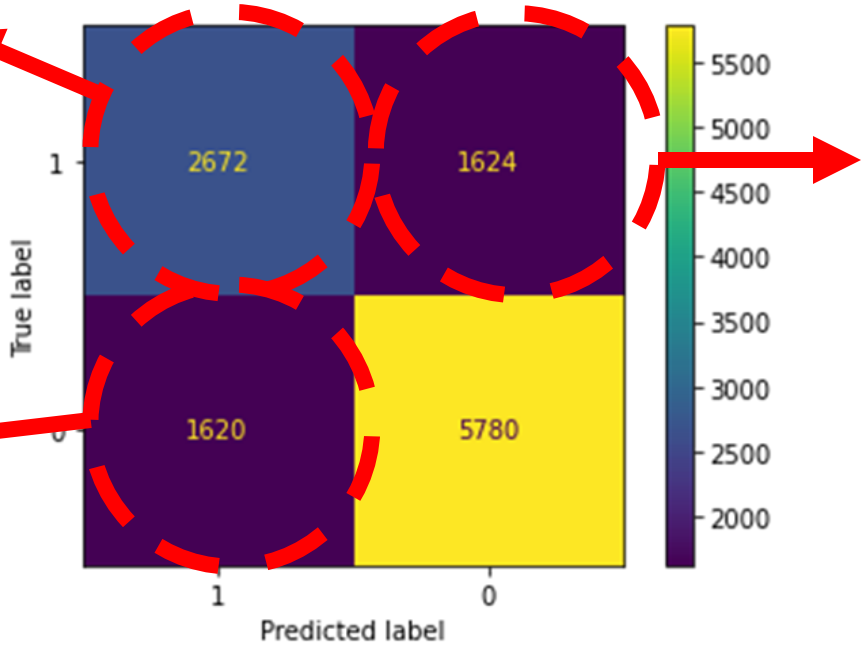
¿Qué puntuación deseo priorizar?

- Los “1” significan que el correo es spam, quiero atrapar tanto spam como pueda, no es grave que un spam llegue a la bandeja de entrada, pero sí que un correo valido se etiquete como spam.(Precisión-EFP)
- Los “1” significan una enfermedad grave, no se me pueden escapar, no quiero falsos negativos (Recall-EFN)
- Los “1” significan que la persona va a la cárcel, bajo ningún concepto quiero un falso positivo (Especificidad)



Modifico el threshold o punto de corte según mi objetivo

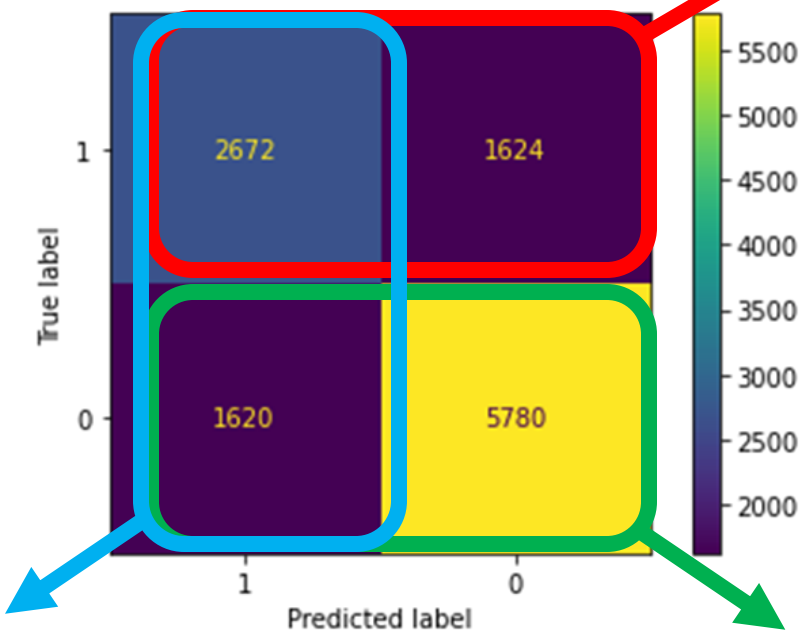
(Precisión) Quiero capturar todos los positivos posibles, no me preocupan los falsos negativos



(Recall) Me preocupa esto porque no quiero que nadie se vuelva a la casa tranquilo cuando tiene una enfermedad grave.

(Especificidad) Me preocupan los falsos positivos, porque implica que por ejemplo un inocente vaya a la cárcel. Por otro lado, esta métrica no me interesa si no me preocupa indicar una enfermedad cuando en realidad no lo hay

Modifico el threshold o punto de corte según mi objetivo



$$\text{Recall} = \frac{TP}{TP + FN}$$

(Enfermedad)

$$\text{Precision} = \frac{TP}{TP + FP}$$

(SPAM)

$$\text{Specificity} = \frac{TN}{TN + FP}$$

(Carcel)

Caso smokers

- 01_Metrics_Smokers.ipynb
- Quiero “atrapar” tantos fumadores como sea posible.
- Los no fumadores no son de mi interés Estoy pensando en una campaña publicitaria muy costosa para **retener** fumadores. Los falsos positivos me representan dinero invertido en alguien que no fuma*.
- Debo intentar mejorar la precisión para la clase de interés “1=Fumador” que se encuentra en 0.62
- Con esfuerzo lo logramos llevar hasta 0.69, sin embargo, la campaña es cara, los recursos limitados y el clasificador aún no tiene la fuerza esperada.
- ¿Qué más puedo hacer?
- ¿Como puedo mejorar una métrica (a costa de otra) cuando tengo el mejor modelo posible?

* La ciencia de datos también puede (y se usa) para fines poco éticos.

Exactitud (accuracy) del modelo: 72.26 %

	precision	recall	f1-score	support
0	0.78	0.78	0.78	7400
1	0.62	0.62	0.62	4296
accuracy			0.72	11696
macro avg	0.70	0.70	0.70	11696
weighted avg	0.72	0.72	0.72	11696

	precision	recall	f1-score	support
0	0.81	0.82	0.82	7400
1	0.69	0.66	0.67	4296
accuracy			0.76	11696
macro avg	0.75	0.74	0.74	11696
weighted avg	0.76	0.76	0.76	11696

Caso smokers

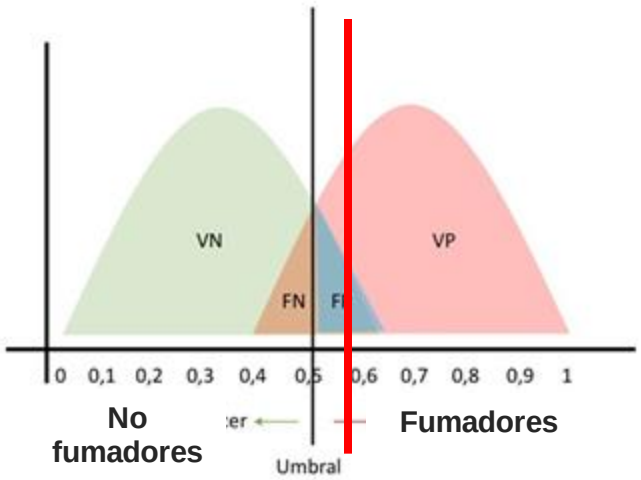
- Modifiquemos el punto de corte, para incluir una persona en la campaña decido que debe tener una propensión del 75% a ser fumador.
- Se me escaparan fumadores, pero tendré mayor certeza de que la campaña se aplica a un fumador real.

	Propension
801	1.000000
2570	0.618393
3143	0.576104
3998	0.516019
4193	0.499422
4456	0.478928
5964	0.352670
6699	0.280639
9177	0.037404
11625	0.000000

```
y_pred = (model.predict_proba(X_test)[:,1]>0.6)
print(classification_report(y_test,y_pred))
```

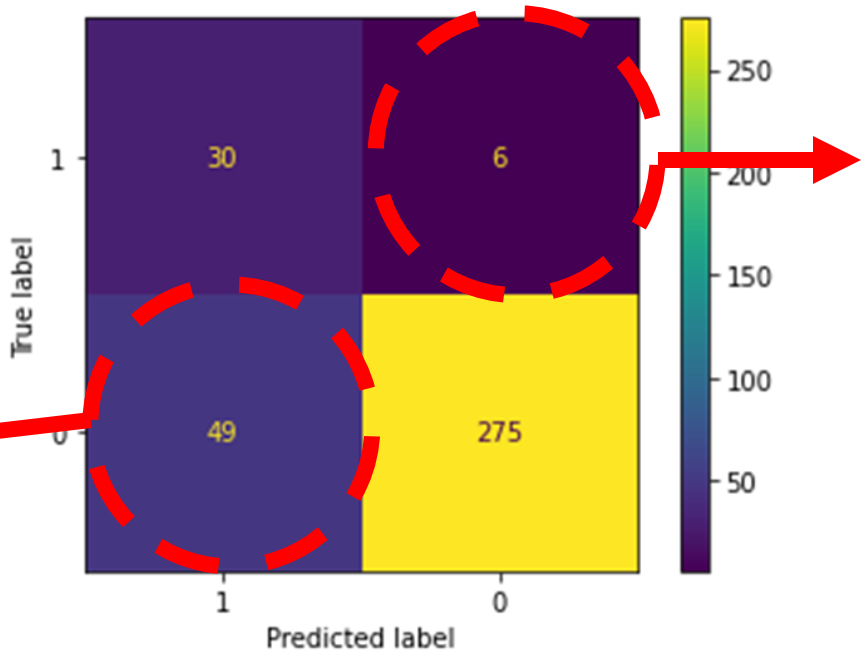
✓ 0.7s

	precision	recall	f1-score	support
0	0.76	0.90	0.83	7400
1	0.75	0.51	0.61	4296
accuracy			0.76	11696
macro avg	0.76	0.70	0.72	11696
weighted avg	0.76	0.76	0.74	11696



Punto de corte y costo

Si predigo que no va a llover (No reviso sumideros) y llueve tengo que asumir el costo del anegamiento.



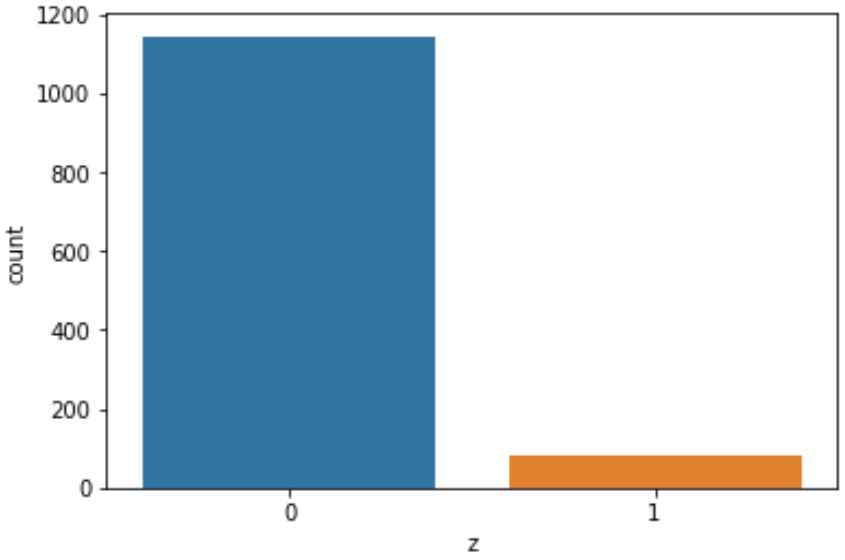
Si predigo que va a llover y no llueve tengo que pagar el costo del envío cuadrillas a revisar sumideros innecesariamente

El problema de las clases raras

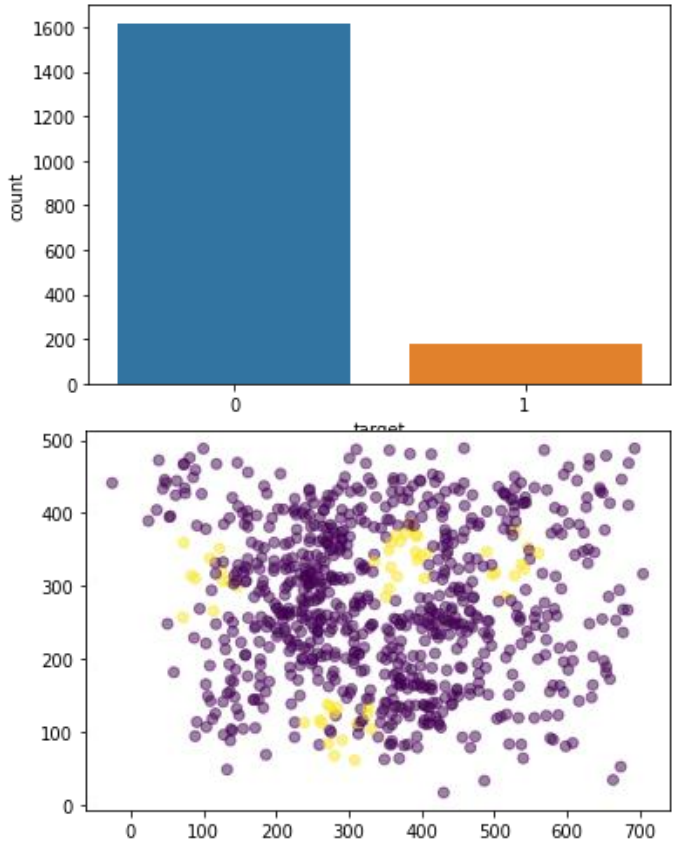
03_SmokersRareClassesUnderSample.ipynb

El problema de las clases raras

- El problema ocurre cuando una clase es preponderante
- Miles o millones contra solo unos pocos registros de la clase de interés
  - Transacciones de tarjetas de crédito
  - Fraude en seguros
- Un clasificador, en términos de exactitud, haría un excelente trabajo indicando todo como la clase más probable.
- ¿Cuándo la diferencia merece ser revisada?
  - No hay reglas absolutas, pero por debajo de un 25% merece atención.



Estratificación

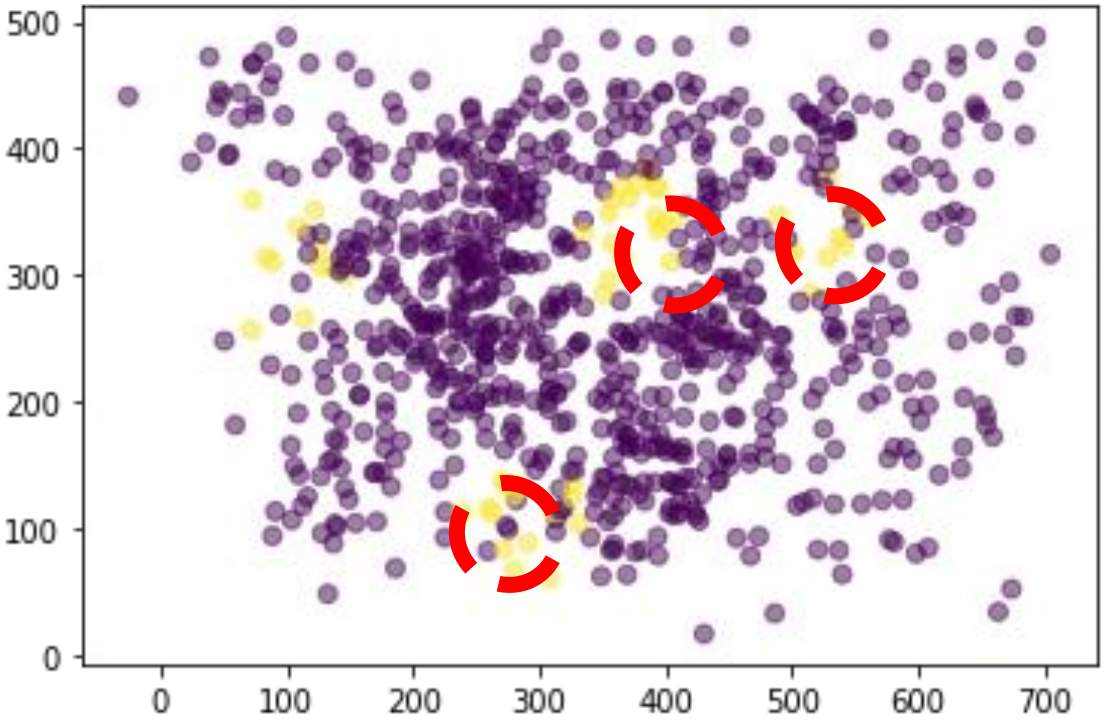


- De mínima se debe tener presente la estratificación de la muestra.
  - Si bien se puede estratificar por cualquier campo normalmente se espera que en validación y entrenamiento exista la misma proporción de clases
- Se busca evitar que un sesgo en las muestras de entrenamiento y validación produzcan un mal modelo o puntaciones irreales.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=42)
```

El problema de las clases raras

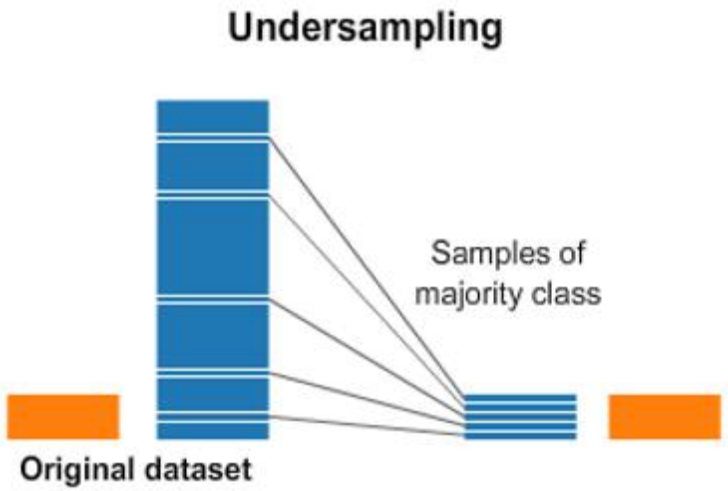
- La medida de exactitud no será adecuada para evaluar el clasificador
  - Si tengo que clasificar reclamos de seguros fraudulentos, tengo 1000 reclamos no fraudulentos contra información de 10 reclamos fraudulento un clasificador que marque todo como no fraudulento tendría una exactitud del 99%.
- Los clasificadores tendrán problemas para ponderar a las clases raras
  - KNN en las zonas bordes por simple cantidad de muestras tendrá tendencia a clasificar los registros como los de la clase prevalente



El problema de clases raras

- Undersampling
  - Random
  - Basado en lógica/Heurísticas (“near duplicates”)
- Se presupone registros redundantes
  - ¿Es realmente así? ¿Cuántos datos son suficientes?
  - Curva de aprendizaje

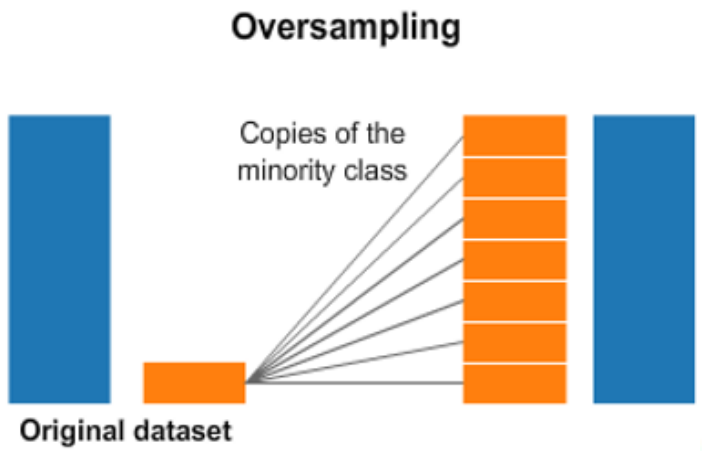
VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.".



El problema de clases raras

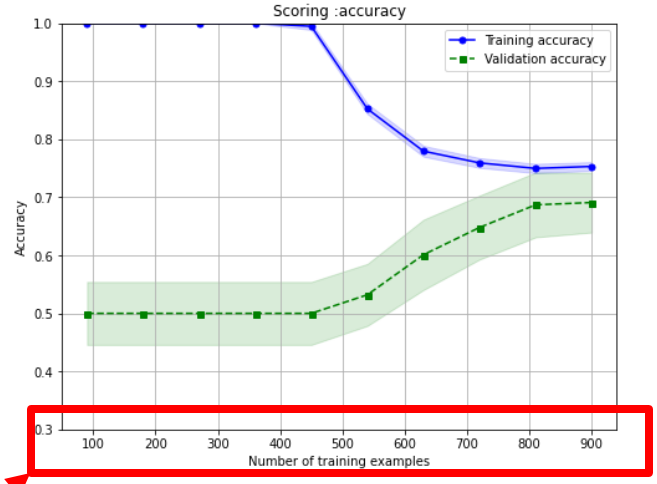
- Oversampling
  - Random (con reposición)
  - SMOTE
- Ponderación
- Ejemplo:
  - 02_SmokersLearningCurve.ipynb
  - 07_RareClassesOverSampling.ipynb
  - Al crear copias exactas de la clase minoritaria, el sobre sampleo aleatorio puede incrementar la probabilidad de tener un sobre ajuste.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. "O'Reilly Media, Inc.".

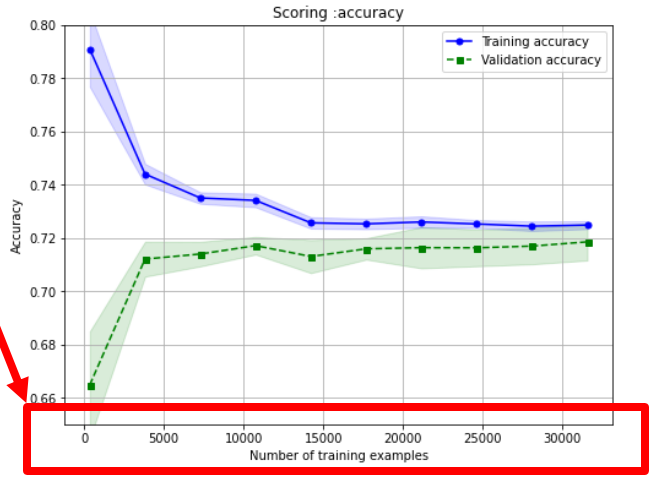


Curvas de aprendizaje

- Algoritmos más complejos requieren más datos.
- Algoritmos más complejos pueden clasificar fenómenos también más complejos
- A más dimensiones más registros (Maldición de la dimensionalidad)
- En mi problema puntual
  - ¿Tengo suficientes datos para hacer undersampling?



Curvas de aprendizaje, dataset: "Smokers", clasificador:KNN



Curvas de aprendizaje, dataset: "Smokers", clasificador: Árboles

Naive Bayes

Clasificación

Clasificación Bayesiana

- Clasificador para variables categóricas
- Es un conjunto de técnicas que resuelve problemas de clasificación aplicando un marco de trabajo probabilístico.
- Conceptos importantes para tener en cuenta para este tipo de clasificadores.
  - Probabilidad condicional
  - Teorema de Bayes

$$P\big(Y = \mathrm{i} \mid X_1, X_2, \ldots X_p\big) = \frac{P(Y = i)P(X_1, .., X_p \mid Y = 1)}{P(Y = 0)P\big(X_1, .., X_p \big| Y = 0\big) + P(Y = 1)P(X_1, .., X_p \mid Y = 1)}$$

Clasificación Bayesiana Exacta

	Ambiente	Temperatura	Humedad	Viento	Juega Tenis
1	Soleado	Alta	Alta	Leve	No
2	Soleado	Alta	Alta	Fuerte	No
3	Nublado	Alta	Alta	Leve	Si
4	Lluvioso	Media	Alta	Leve	No
5	Lluvioso	Baja	Normal	Fuerte	No
6	Lluvioso	Baja	Normal	Fuerte	No
7	Nublado	Baja	Normal	Leve	Si
8	Soleado	Media	Alta	Leve	Si
9	Soleado	Baja	Normal	Leve	Si
10	Lluvioso	Media	Normal	Leve	No
11	Soleado	Media	Normal	Fuerte	Si
12	Nublado	Media	Alta	Fuerte	Si
13	Nublado	Alta	Normal	Leve	Si
14	Lluvioso	Media	Alta	Fuerte	No

Queremos saber si se jugará al tenis bajo las siguientes condiciones:

Ambiente	Temperatura	Humedad	Viento	Juega Tenis
Soleado	Baja	Alta	Fuerte	?

$$P(Y = \mathrm{i} \mid X_1, X_2, \ldots X_p) = \frac{P(Y = i)P(X_1, .., X_p \mid Y = 1)}{P(Y = 0)P(X_1, .., X_p \mid Y = 0) + P(Y = 1)P(X_1, .., X_p \mid Y = 1)}$$

Clasificación Bayesiana Exacta

- Clasificación bayesiana exacta
  - Seleccionar los registros que contentan los mismos valores predictores de los registros que queremos clasificar.
  - Determinemos a que clase pertenecen y cuál es el prevalente
  - Asignemos esa clase al nuevo registro
- El problema
  - Si el número de variables excede a unas cuantas, rápidamente nos quedaremos sin registros con coincidencias exactas.
  - ¿Cuál es la probabilidad de que se juegue tenis si esta "soleado,baja,alta, fuerte"?

	Ambiente	Temperatura	Humedad	Viento
0	Soleado	Baja	Alta	Fuerte

	Ambiente	Temperatura	Humedad	Viento	JugarTenis
0	Soleado	Alta	Alta	Leve	No
1	Soleado	Alta	Alta	Fuerte	No
2	Nublado	Alta	Alta	Leve	Sí
3	Lluvioso	Media	Alta	Leve	No
4	Lluvioso	Baja	Normal	Fuerte	No
5	Lluvioso	Baja	Normal	Fuerte	No
6	Nublado	Baja	Normal	Leve	Sí
7	Soleado	Media	Alta	Leve	Sí
8	Soleado	Baja	Normal	Leve	Sí
9	Lluvioso	Media	Normal	Leve	No
10	Soleado	Media	Normal	Fuerte	Sí
11	Nublado	Media	Alta	Fuerte	Sí
12	Nublado	Alta	Normal	Leve	Sí
13	Lluvioso	Media	Alta	Fuerte	No

La solución ingenua

- No nos limitamos al cálculo de probabilidad de los registros que coinciden completamente.
- Se estiman las probabilidades condicionales individuales
  - Dada una respuesta binaria $(Y = i, \text{donde } i = 0 \; o \; 1)$
  - Estimamos probabilidades condicionales individualmente para cada predictora $\mathrm{P}(X_j|Y = \mathrm{i})$
  - Está última sería la probabilidad de que el valor de la predictora se encuentre en el registro cuando $Y = i$, para nuestro ejemplo la probabilidad de que llueva cuando se juega tenis
  - La probabilidad se estima como la proporción de valores X_j entre los valores $Y = i$. La proporción de días de lluvia entre los días que se juega tenis.

La solución ingenua

$$P(Y = \mathrm{i} \mid X_1, X_2, \ldots X_p) = \frac{P(Y = i)P(X_1, .., X_p \mid Y = 1)}{P(Y = 0)P(X_1, .., X_p \mid Y = 0) + P(Y = 1)P(X_1, .., X_p \mid Y = 1)}$$

- ¿Cuál es la probabilidad de jugar tenis si esta soleado?
- ¿Cuál es la probabilidad de jugar tenis si el viento es fuerte?
- Realizo la misma operación para cada campo.
- Asumiendo "ingenuamente" independencia de variables multiplico cada una de las probabilidades condiciones.
- Quedando la siguiente formula

$$P(Y = \mathrm{i} \mid X_1, X_2, \ldots X_p) = \frac{P(Y = i)P(X_1 \mid Y = i) \ldots P(X_p \mid Y = i)}{P(Y = 0)P(X_1 \mid Y = 0) \ldots P(X_p \mid Y = 0) + P(Y = 1)P(X_1 \mid Y = 1) \ldots P(X_p \mid Y = 1)}$$

Clasificación Bayesiana Ingenua

	Ambiente	Temperatura	Humedad	Viento	Juega Tenis
1	Soleado	Alta	Alta	Leve	No
2	Soleado	Alta	Alta	Fuerte	No
3	Nublado	Alta	Alta	Leve	Si
4	Lluvioso	Media	Alta	Leve	No
5	Lluvioso	Baja	Normal	Fuerte	No
6	Lluvioso	Baja	Normal	Fuerte	No
7	Nublado	Baja	Normal	Leve	Si
8	Soleado	Media	Alta	Leve	Si
9	Soleado	Baja	Normal	Leve	Si
10	Lluvioso	Media	Normal	Leve	No
11	Soleado	Media	Normal	Fuerte	Si
12	Nublado	Media	Alta	Fuerte	Si
13	Nublado	Alta	Normal	Leve	Si
14	Lluvioso	Media	Alta	Fuerte	No

Ambiente	Temperatura	Humedad	Viento	Juega Tenis
Soleado	Baja	Alta	Fuerte	?

$$P(Y = \mathrm{i} \mid X_1, X_2, \ldots X_p) = \frac{P(Y=i)P(X_1|Y=i) \ldots P(X_p|Y=i)}{P(Y=0)P(X_1|Y=0) \ldots P(X_p|Y=0) + P(Y=1)P(X_1|Y=1) \ldots P(X_p|Y=1)}$$

Bayes ingenuo

- Aprendizaje paramétrico
  - Determinar probabilidades a priori de cada clase y las probabilidades condicionales.

	Valores que toma	Cantidad de Casos	% casos totales
Ambiente	Soleado	5	35,7%
	Nublado	4	28,6%
	Lluvioso	5	35,7%
Temperatura	Alta	4	28,6%
	Media	6	42,8%
	Baja	4	28,6%
Humedad	Alta	7	50%
	Normal	7	50%
Viento	Leve	8	57,2%
	Fuerte	6	42,8%

$$\text{Casos}_{\text{Juega Tenis = Si}} = 7$$

$$\text{Casos}_{\text{Juega Tenis = No}} = 7$$

$$P(\text{Juega Tenis = Si}) = 0{,}5 = 50\%$$

$$P(\text{Juega Tenis = No}) = 0{,}5 = 50\%$$

Bayes ingenuo

Cantidad de Casos

	Valores que toma	Juega Tenis Si	Juega Tenis No
Ambiente	Soleado	3	2
	Nublado	4	0
	Lluvioso	0	5
Temperatura	Alta	2	2
	Media	3	3
	Baja	2	2
Humedad	Alta	3	4
	Normal	4	3
Viento	Leve	5	3
	Fuerte	2	4

Desglose de los 14 casos según si juegan o no al tenis

Probabilidades

	Valores que toma	Juega Tenis Si	Juega Tenis No
Ambiente	Soleado	3/7 = 42,8%	2/7 = 28,6%
	Nublado	4/7 = 57,2%	0
	Lluvioso	0	5/7 = 71,4%
Temperatura	Alta	2/7 = 28,6%	2/7 = 28,6%
	Media	3/7 = 42,8%	3/7 = 42,8%
	Baja	2/7 = 28,6%	2/7 = 28,6%
Humedad	Alta	3/7 = 42,8%	4/7 = 57,2%
	Normal	4/7 = 57,2%	3/7 = 42,8%
Viento	Leve	5/7 = 71,4%	3/7 = 42,8%
	Fuerte	2/7 = 28,6%	4/7 = 57,2%

Probabilidades condicionales

Ejemplo Bayes Ingenuo

- Predicción a realizar

Ambiente	Temperatura	Humedad	Viento	Juega Tenis
Soleado	Baja	Alta	Fuerte	?

P(Juega Tenis = Si) = 0,5
P(Juega Tenis = No) = 0,5

- P(Juega Tenis = Si) = 0,428 x 0,286 x 0,428 x 0,286 x 0,5 = 0,0075

- P(Juega Tenis = No) = 0,286 x 0,286 x 0,572 x 0,572 x 0,5 = 0,0133

Bayes Ingenuo

- Normalizando
  - P(Juega Tenis = Si) = 0,0075 / (0,0075 + 0,0133) = 36%
  - P(Juega Tenis = No) = 0,0133 / (0,0075 + 0,0133) = **64%**
- El clasificador bayesiano ingenuo va a predecir que no se juega al tenis con una probabilidad del 64%.
- ¿Y si los atributos son continuos?
  - Discretizar en rangos y luego aplicar la estrategia para atributos categóricos.
  - Usar la versión Gaussiana del clasificador. Normalmente denominada Gaussian Naïve Bayes

Bayes Ingenuo

- Problemas
  - Si no existen casos en el conjunto de entrenamiento para todas las combinaciones de atributos A_i y la clase, la probabilidad condicional de ese atributo es **0**, anulando toda la expresión.

Refund	Marital Status	Taxable Income	Evade
Yes	Divorced	90K	?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

  - P(Evade = Yes) = 0 x 1/3 x 1/3 x 3/10 = **0**
  - P(Evade = No) = 3/7 x 1/7 x 0 x 7/10 = **0**

Bayes Ingenuo

- La solución es corregir la estimación de la probabilidad
  - Ampliar el conjunto de datos de entrenamiento que contemple todas las combinaciones entre los atributos y la clase.
- Aplicar otras medidas para el cálculo de la probabilidad que permitan trabajar con probabilidades nulas.
  - Suavizador de Laplace
  - MultinomialNB
    - alpha (default=1.0)
    - force_alpha (Default=False, default 1e-10)
      - force_alpha = True*
- **This may cause numerical errors if alpha is too close to 0*

Bayes ingenuo

- Ventajas
  - Robusto a outliers.
  - Maneja valores faltantes ignorando la instancia cuando se estiman las probabilidades.
  - Robusto a atributos irrelevantes.
  - Funciona sorpresivamente bien.
- Desventajas
  - La independencia de los atributos no siempre es cierta.
  - Tampoco es cierta la distribución normal de los mismos.
  - Susceptible al envenenamiento bayesiano.

Bayes ingenuo y sklearn

- from sklearn.naive_bayes import MultinomialNB
- Los datos tienen que ser numéricos categóricos
  - OrdinalEncoder
  - Dummies
- Misma forma de operar de la API
  - fit
  - transform
  - Notebook: _00_PlayTennis.ipynb

Naive Bayes

Titanic: _02_Titanic.ipynb

Spam-No Spam: _03_Spam.ipynb

News Groups: _04_newsGroups.ipynb

Naive Bayes

- A pesar de no cumplirse la independencia de variables naive bayes suele funcionar muy bien
- En conjunto de datos pequeños incluso puede superar a herramientas más complejas
- Robusto y fácil de implementar
  - Clasificación de correos como spam
  - Tratamientos médicos
  - Clasificación de secuencias ADN

Naive bayes

- En terminos de spam - no spam
- Si tenemos 100 documentos para entrenar y recibimos un mensaje "hello world"
- Se puede observar que la asunción de que las variables son independientes no se cumplirá, en un documento que dice automóvil es más probable que diga Fiat o Ford que un documento que no se hable de vehículos.
- Con todo Naive Bayes obtiene buenos resultados

Naive bayes

- Probabilidad a priori
- ¿Cuál es la probabilidad de tener un spam?

$$P(\text{ham}) = 1 - P(\text{spam}).$$

- Esto último se calcula del dataset (Si fuese neutro en este sentido)
- O puedo ser aportado por un experto

Fin