

2023



ACT.3_2:
COMPARATIVA
CLASIFICADORES
NAIVE BAYES

Carolina María Montesdeoca Álvarez
SNS

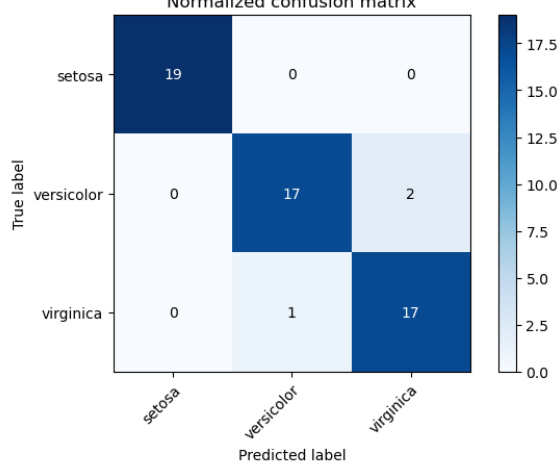
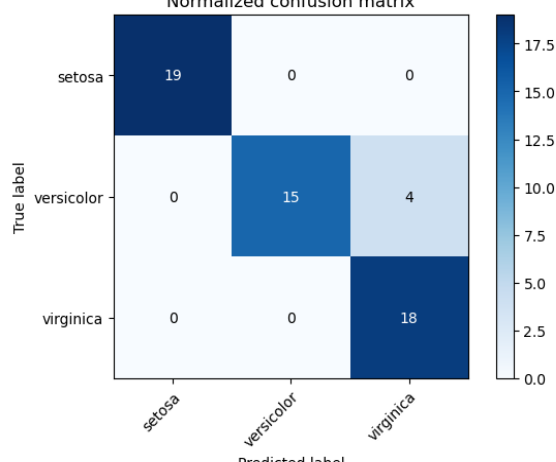
Actividad 3.2 – Comparativa clasificadores NaiveBayes

Utilizando como referencia el cuaderno Ejemplo_3_2_Iris_NaiveBayes - GaussianNB.ipynb, realizar otros tantos cuadernos, o desarrollar la solución como consideres oportuno, con los diferentes clasificadores NaiveBayes, de forma que para un mismo problema podamos comparar las precisiones obtenidas.

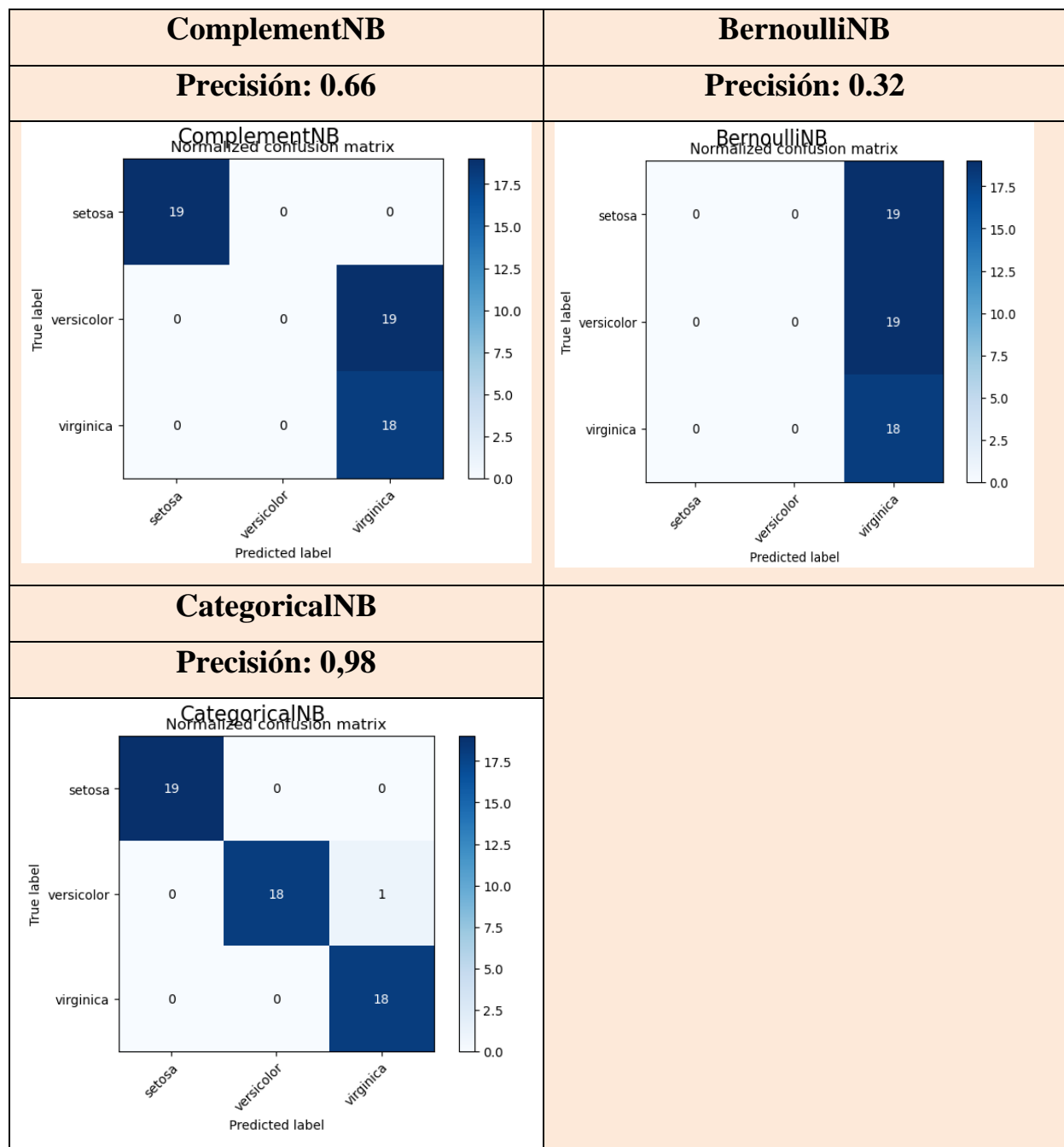
Realizar esta comparativa para el Dataset Iris y el Dataset Penguin (por separado)

A continuación, representamos en las siguientes tablas los datos obtenidos en los diferentes notebooks de Jupiter, tanto para el dataset de Iris como para el dataset de los Pingüinos.

DATASET IRIS

GaussianNB		MultinomialNB																																	
Precisión: 0.94		Precisión: 0.93																																	
<p>GaussianNB</p> <p>Normalized confusion matrix</p>  <table><tr><th></th><th>setosa</th><th>versicolor</th><th>virginica</th></tr><tr><th>setosa</th><td>19</td><td>0</td><td>0</td></tr><tr><th>versicolor</th><td>0</td><td>17</td><td>2</td></tr><tr><th>virginica</th><td>0</td><td>1</td><td>17</td></tr></table> <p>Predicted label</p>			setosa	versicolor	virginica	setosa	19	0	0	versicolor	0	17	2	virginica	0	1	17	<p>MultinomialNB</p> <p>Normalized confusion matrix</p>  <table><tr><th></th><th>setosa</th><th>versicolor</th><th>virginica</th></tr><tr><th>setosa</th><td>19</td><td>0</td><td>0</td></tr><tr><th>versicolor</th><td>0</td><td>15</td><td>4</td></tr><tr><th>virginica</th><td>0</td><td>0</td><td>18</td></tr></table> <p>Predicted label</p>			setosa	versicolor	virginica	setosa	19	0	0	versicolor	0	15	4	virginica	0	0	18
	setosa	versicolor	virginica																																
setosa	19	0	0																																
versicolor	0	17	2																																
virginica	0	1	17																																
	setosa	versicolor	virginica																																
setosa	19	0	0																																
versicolor	0	15	4																																
virginica	0	0	18																																

Para el *dataset de Iris* observamos que los modelos de clasificadores Naive Bayes con menor precisión son el modelo de BernulliNB con 0,32 y el modelo de ComplementNB con 0,66.



La precisión de un clasificador en el contexto del aprendizaje automático es una medida de cuán bien el modelo puede predecir correctamente la clase o categoría de nuevas observaciones. Se calcula como la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de predicciones realizadas. En este caso, estamos trabajando con diferentes variantes de clasificadores Naive Bayes y cada uno muestra una precisión diferente para el mismo conjunto de datos. Esto puede interpretarse de la siguiente manera:

GaussianNB (Precisión: 0.94): Este clasificador asume que los datos de cada característica siguen una distribución gaussiana (normal). Una precisión del 94% indica que el modelo es muy eficaz para este conjunto de datos, sugiriendo que la suposición de normalidad se ajusta bien a las características del conjunto de datos.

MultinomialNB (Precisión: 0.93): Este clasificador se utiliza generalmente para datos que pueden representarse como conteos de frecuencia o tasas de ocurrencia, como en el procesamiento de texto (recuento de palabras). Una precisión del 93% es también muy alta, lo que indica que este modelo es casi tan eficaz como GaussianNB para este conjunto de datos.

ComplementNB (Precisión: 0.66): Es una adaptación del MultinomialNB que funciona mejor con conjuntos de datos desequilibrados.

Una precisión del 66% es moderadamente baja comparada con los dos primeros, lo que podría sugerir que este clasificador no es tan adecuado para nuestro conjunto de datos o que los datos no son tan desequilibrados como para beneficiarse de este enfoque.

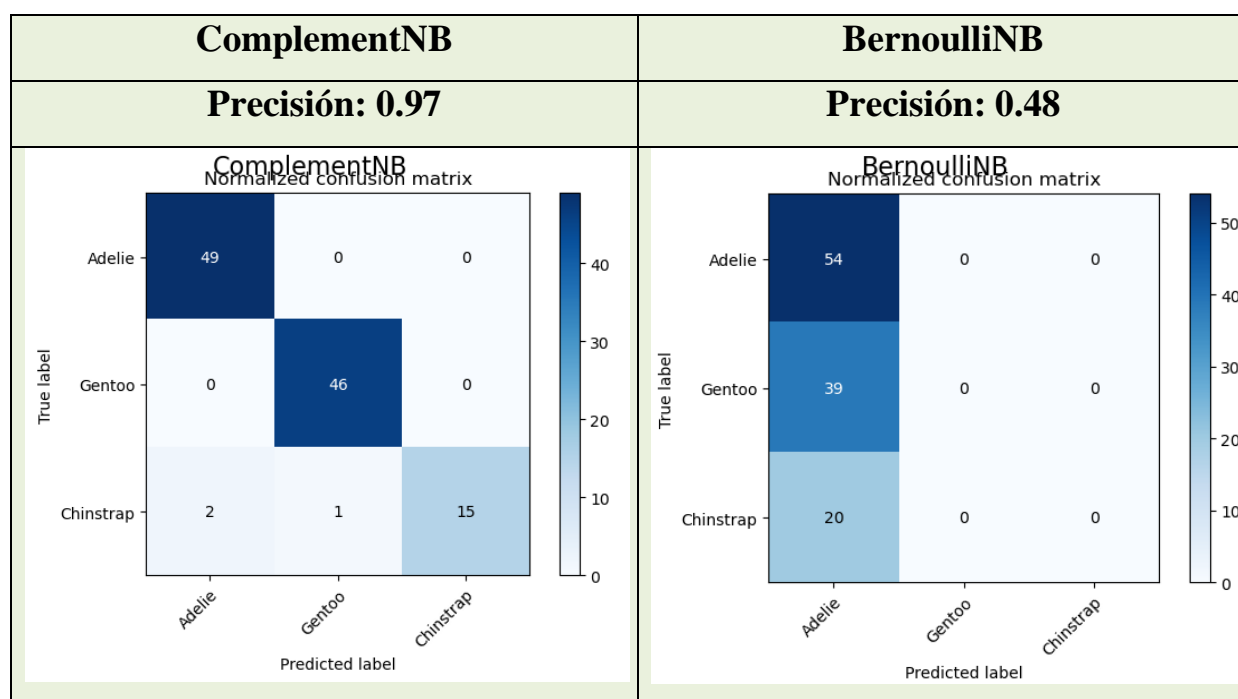
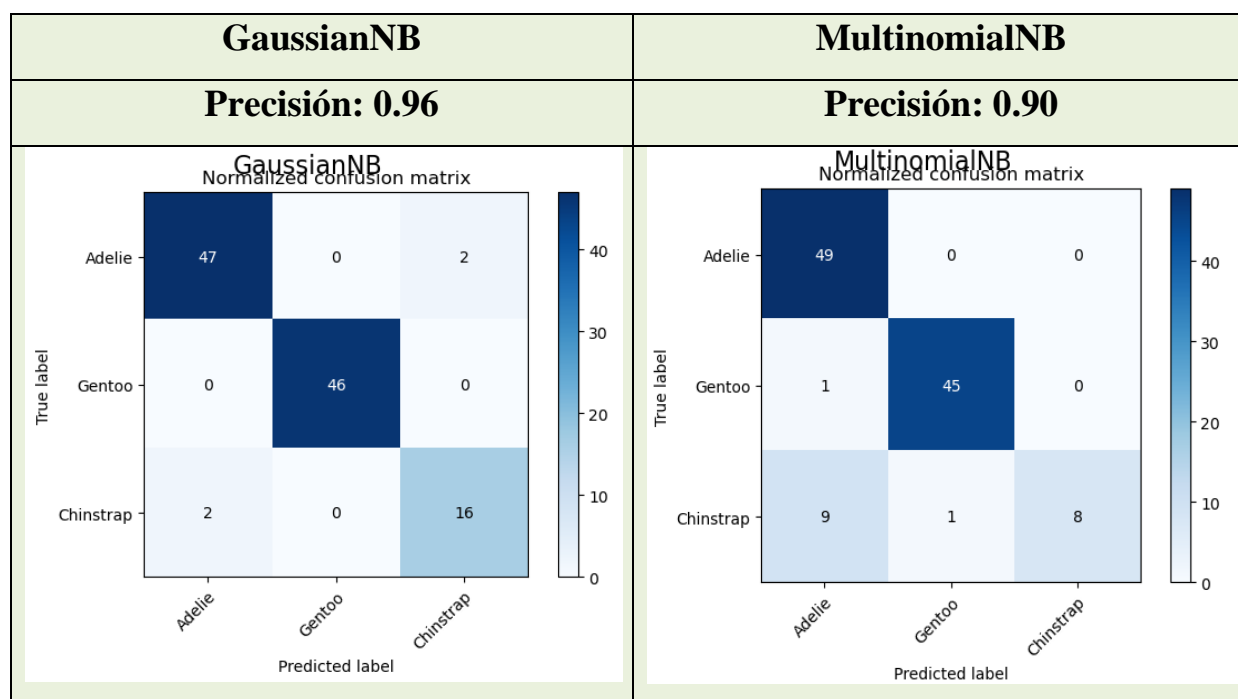
CategoricalNB (Precisión: 0.98): Este clasificador se utiliza para datos con características categóricas. Una precisión del 98% es extremadamente alta, lo que sugiere que este conjunto de datos Iris tiene características categóricas que son muy predictivas de la variable objetivo.

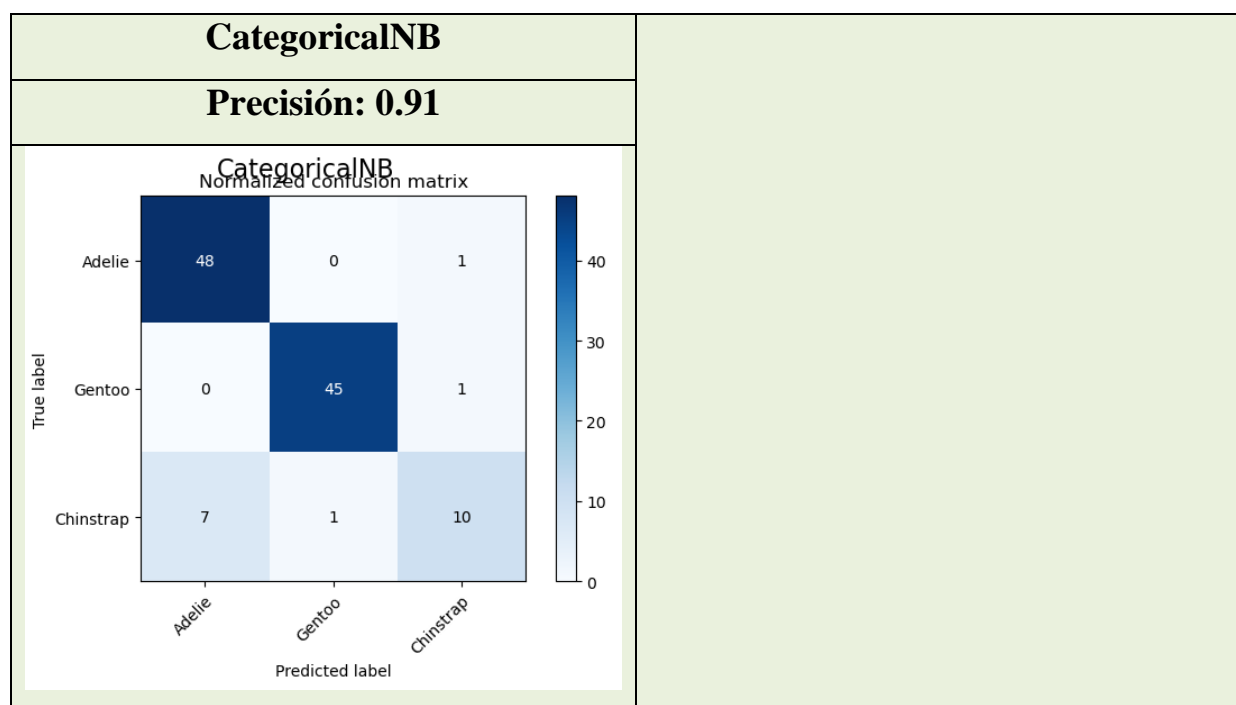
BernoulliNB (Precisión: 0.32): Este clasificador es adecuado para datos binarios/dicotómicos (como 'sí/no' o 'verdadero/falso'). Una precisión del 32% es bastante baja, lo que indica que este modelo no se ajusta bien a este conjunto de datos, posiblemente debido a que las características no son adecuadas para una modelización binaria.

En resumen, estos resultados indican que las suposiciones subyacentes de GaussianNB y MultinomialNB se alinean bien con la naturaleza de este conjunto de datos, mientras que las de ComplementNB y BernoulliNB no son tan adecuadas. Esto subraya la importancia de comprender las características de los datos y elegir un modelo que se ajuste bien a estas características.

Por otra parte, Para el *dataset de Penguins* observamos que los modelos de clasificadores Naive Bayes con menor precisión es el modelo de BernulliNB con 0,48. También hay que destacar que en el dataset de los pingüinos se prescindieron de tres columnas: Isla, Sexo, Masa corporal y se eliminaron los datos NaN.

DATASET PINGÜINOS





El análisis de la precisión de diferentes clasificadores Naive Bayes en un conjunto de datos de pingüinos, del cual se han eliminado tres columnas de datos y los valores NaN, ofrece información valiosa sobre cómo estos modelos se ajustan a los datos. Aquí está lo que cada precisión podría significar en este contexto:

GaussianNB (Precisión: 0.96): Con una precisión del 96%, GaussianNB, que asume una distribución normal de los datos, parece ser muy efectivo para este conjunto de datos. Esto sugiere que las características restantes de los pingüinos (después de eliminar las columnas y los NaN) siguen una distribución que se asemeja a la normal.

MultinomialNB (Precisión: 0.90): Este modelo es adecuado para características que representan frecuencias o tasas de ocurrencia. Una precisión del 90% es alta, indicando que, aunque no es tan efectivo como GaussianNB, este modelo aún se ajusta bastante bien a los datos.

ComplementNB (Precisión: 0.97): Este es una variante del MultinomialNB diseñado para manejar mejor los conjuntos de datos desequilibrados. La precisión más alta de 97% sugiere que este modelo es particularmente adecuado para este conjunto de datos,

posiblemente debido a que maneja bien cualquier desequilibrio presente en los datos después de la eliminación de columnas y valores NaN.

CategoricalNB (Precisión: 0.91): Este modelo está diseñado para variables categóricas y ofrece una precisión del 91%. Esto implica que el conjunto de datos tiene características categóricas que este modelo puede explotar eficazmente.

BernoulliNB (Precisión: 0.48): Adecuado para características dicotómicas, una precisión del 48% sugiere que este modelo no es adecuado para este conjunto de datos. Esto podría ser porque las características del conjunto de datos no son binarias o porque la eliminación de columnas y valores NaN ha alterado la estructura de los datos de manera que no se ajusta bien a un modelo Bernoulli.

En general, estos resultados indican que para el dataset específico de pingüinos (post-limpieza), los modelos ComplementNB y GaussianNB son los más adecuados, mientras que el modelo BernoulliNB es el menos adecuado. Esto demuestra la importancia de elegir un modelo de clasificación que se alinee bien con las características y la estructura de los datos.

URL: https://github.com/carolProg/SNS_23_24/tree/main