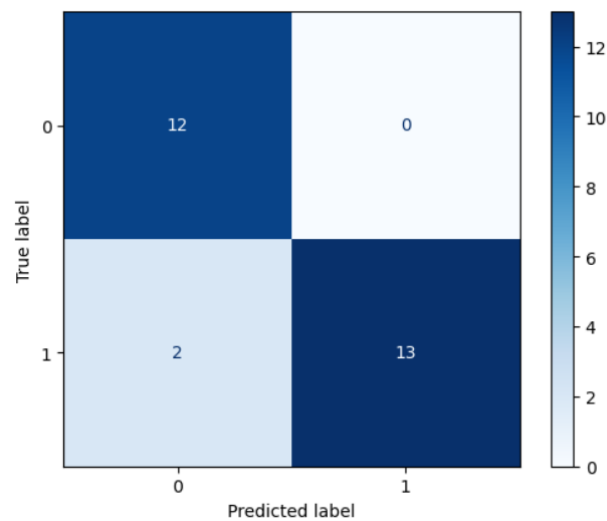


BLOQUE 1: NAIVE BAYES Y REGRESIÓN

Naive Bayes

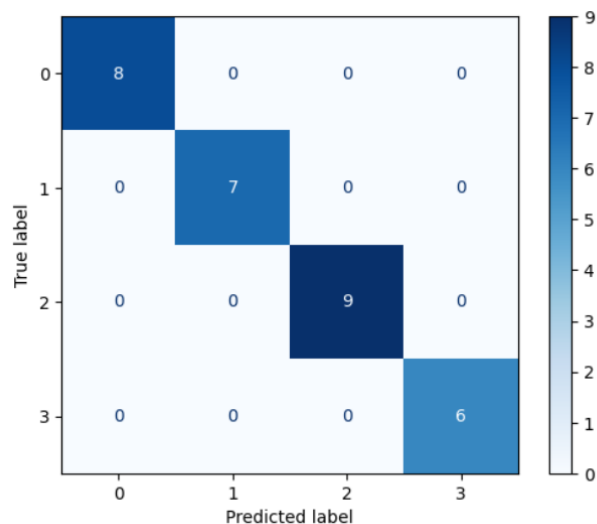
Bernoulli: Ejercicio 5

El modelo tuvo 25 predicciones correctas de 27 posibles. En el caso del 0 tuvo 12 aciertos y 0 fallos, mientras que en el caso del 1 tuvo 13 aciertos y 2 fallos. Esto quiere decir que es muy bueno para detectar clase 0 y bastante bueno para detectar clase 1.



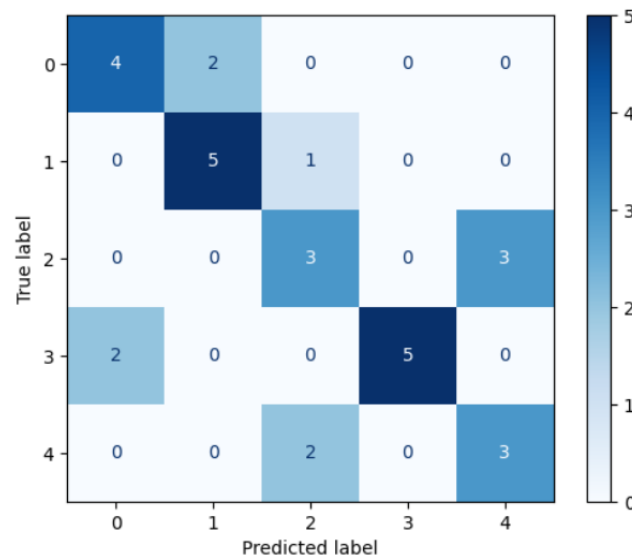
Multinomial: Ejercicio 5

El modelo tuvo 30 predicciones correctas de 30 posibles, lo cual indica que con los datos de entrenamiento tuvo un 100 % de precisión en el conjunto de prueba. Esto implica que lo más probable es que el modelo memorizase los datos del entrenamiento y que pueda no ser tan preciso con nuevos datos, sobreajuste.



Gaussiana: Ejercicio 4

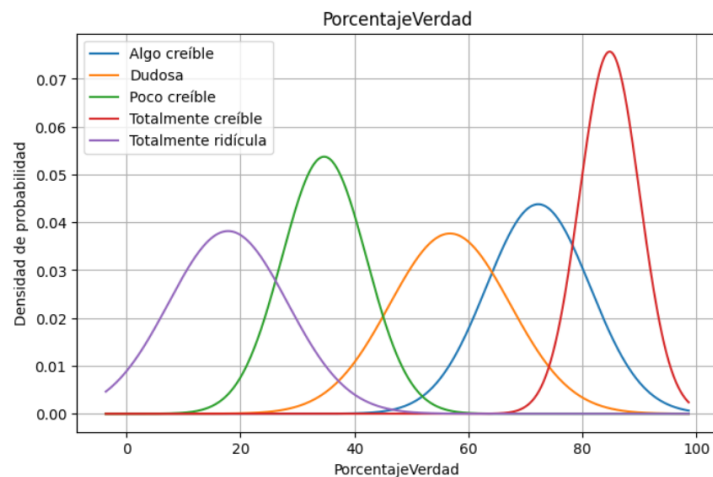
El modelo tuvo 20 predicciones correctas de 33 posibles. Se puede observar que el modelo no clasificó con una precisión perfecta, pero si acertó la gran mayoría. En las clases 1 y 3 tienen un buen rendimiento, mientras que en las otras muestra mayor confusión con clases cercanas. El modelo parece confundir clases cercanas en credibilidad por pequeños matices, pero se entiende que entiende la tendencia general.



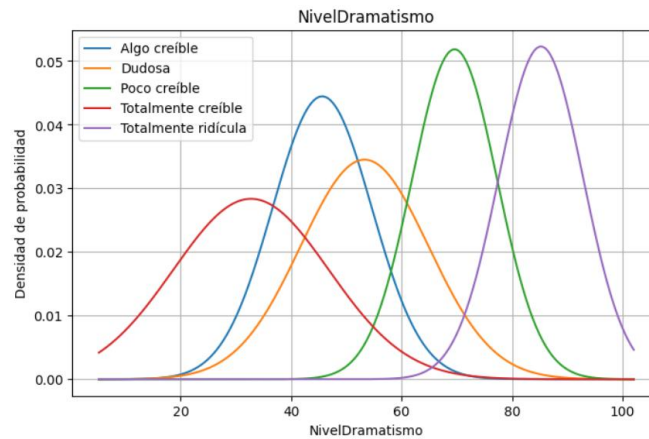
Gaussiana: Ejercicio 5

En los gráficos se pueden ver representadas las distribuciones de probabilidad de las características para cada clase.

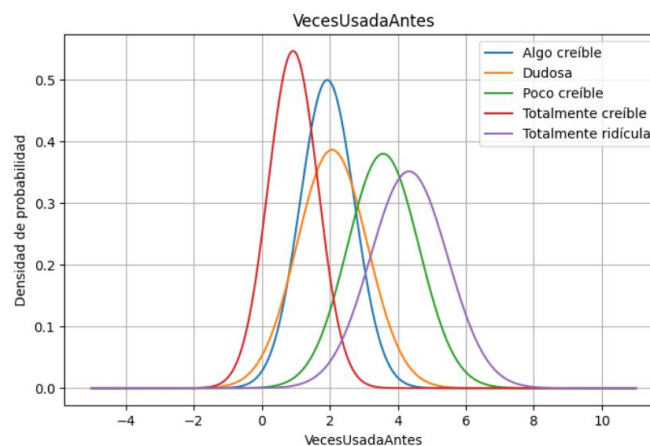
En PorcentajeVerdad las categorías extremas se discriminan muy bien, haciendo también un buen trabajo con las clases “poco creíble” y “algo creíble”, aunque se sitúan en el rango medio pudiendo confundirse con “dudosa”.



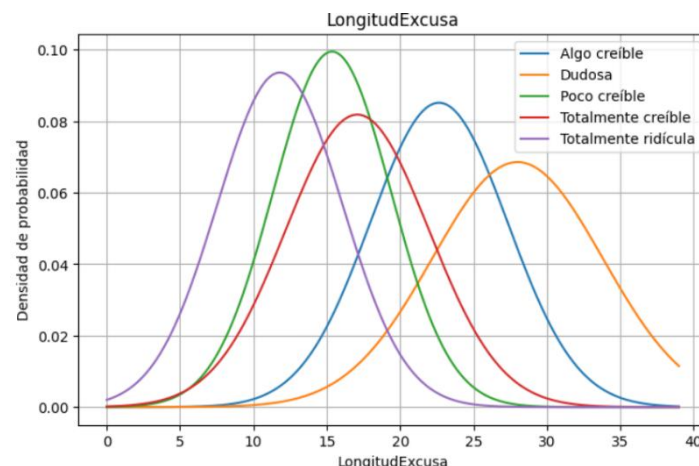
En NivelDramatismo “Totalmente ridícula” y “Poco creíble” tienen altos niveles de dramatismo, ya que suele asociarse con baja credibilidad, pero podría haber confusión cuando niveles de dramatismo altos y niveles más bajos ya que se concentran las clases más. Las excusas más dramatizadas serán percibidas como menos creíbles.



En VecesUsadaAntes se nota gran concentración, especialmente “algo creíble” y “dudosa” tienen valores más concentrados alrededor de 0-3 usos, creando confusión. Mientras que “Totalmente ridícula” presenta una distribución más amplia y desplazada. Las excusas repetidas más veces serán percibidas como menos creíbles.



En LongitudExcusa que se diferencia muy bien “Totalmente ridícula” y “Dudosa”, pero la alta concentración entre 10 y 20 genera que pueda haber confusión entre “Totalmente ridícula”, “Poco creíble” y “Totalmente creíble”. Las excusas más largas generan duda pero pueden llegar a creerse, mientras que si es demasiado corta no, en el término medio es donde más confusión genera, si es creíble o no.



Regresión Lineal

Cálculo de métricas

El cálculo de métricas nos ha dado los siguientes resultados:

$MAE = 1.0658141036401503e-13$.

Este es el error absoluto medio, consiste en cuanto se equivoca de media en las predicciones. Cuanto más cerca de 0 mejor.

$MSE = 1.1777630206077289e-26$

El error cuadrático medio penalizará más los errores grandes. Cuanto más cerca de 0 mejor.

$R^2 = 1$

El coeficiente de determinación medirá que tan bien explica el modelo los datos. Si es 1, el modelo predice perfectamente todos los valores.

En teoría, según las métricas el modelo predice perfectamente los datos del conjunto de pruebas, ya que ($MAE \approx 0$, $MSE \approx 0$ y $R^2 = 1$), esto puede apuntar a un data leakage o si el modelo está sobre ajustado.

Learning Rate

Nuestro código entrena un modelo de regresión lineal mediante descenso del gradiente desde 0, es decir, se ajustan los parámetros paso a paso para reducir el error.

- Utilizando de Learning Rate = 0.0001 y de épocas = 1000 podemos observar el coste esta disminuyendo de forma estable, comenzando en 121 623 y terminando en 101 586. Probaremos con Learning Rates más grandes ya que el aprendizaje va muy lento, el coste aún no se ha estabilizado tras 1000 épocas.

Los coeficientes obtenidos [47.43, -0.59, 0.14, 1.59, -0.19, -0.15, 1.50] nos muestran cómo afectan las diferentes características al precio del seguro, por ejemplo, a mayor edad se bajará ligeramente (-0.59) pero cuanto más viejo el coche más sube (1.50).

- Al aumentar Learning Rate a 0.0005, manteniendo el número de épocas, se observa un coste más bajo, comenzando en 99 574 y terminando en 40 472, bajando mucho más rápido.
En este caso los coeficientes son más grandes en magnitud, esto puede indicar que está convergiendo hacia otro mínimo del coste, o que todavía no está estabilizado.
- Por último, hemos aumentado también el número de épocas a 3000, haciendo esto se observa un coste mucho más bajo al comenzar, de 36 620 y terminando en 2 013, reduciendo bastante el error cuadrático medio, estando muchísimo más ajustado. El Learning Rate actual también es el óptimo.