

Redes neuronales secuenciales: Sommelier

Martín Quijano - Martina Coletto

Variables

- fixed acidity – Acidez fija
- volatile acidity – Acidez volátil
- citric acid – Ácido cítrico
- residual sugar – Azúcar residual
- chlorides – Cloruros
- free sulfur dioxide – Dióxido de azufre libre
- total sulfur dioxide – Dióxido de azufre total
- density – Densidad
- pH – Nivel de pH
- sulphates – Sulfatos
- alcohol – Contenido de alcohol
- quality – Calificación
- sensorial del vino (target)

Modelos

Redes secuenciales

Redes funcionales

Objetivo

Realizar un análisis exhaustivo de variables (gráfico y analítico) para determinar el tipo y calidad del vino para encontrar errores en embotellamiento.

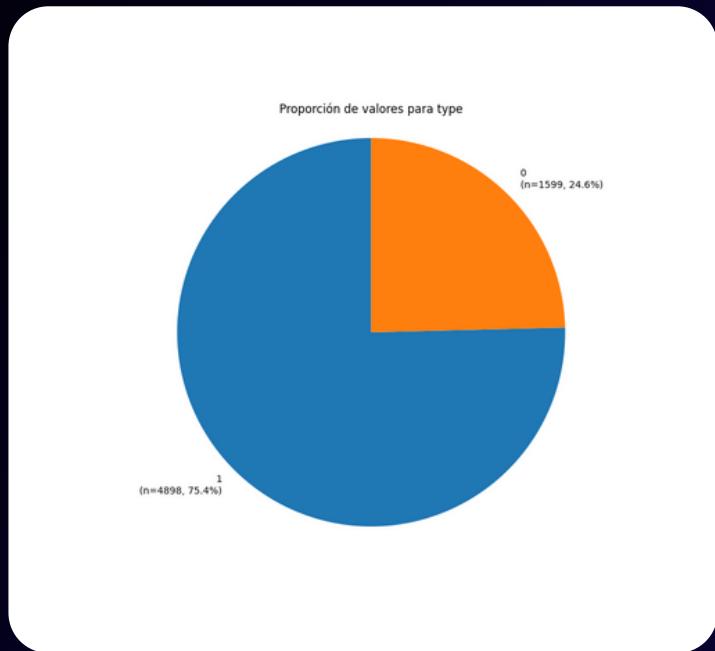


Análisis exploratorio

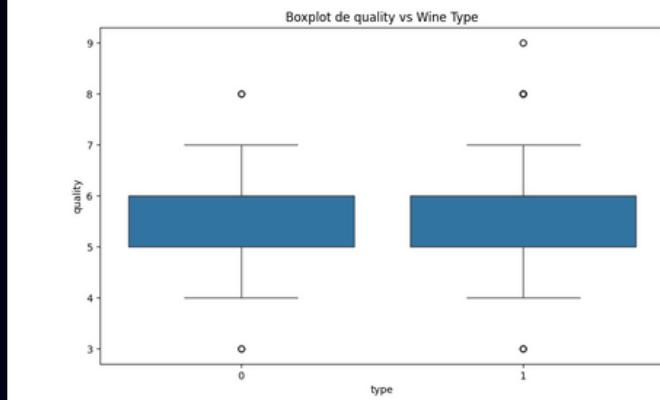
Lo importante a notar es que el dataset proviene de un productor de vinos de Portugal.

La medición de tipo de vino es algo objetivo, mientras que la calidad es algo subjetivo. Por lo que leímos, para medir la calidad se tomaron a al menos tres sommeliers por vino, pidieron que den su rating, y se tomó la media de los valores de los sommeliers.

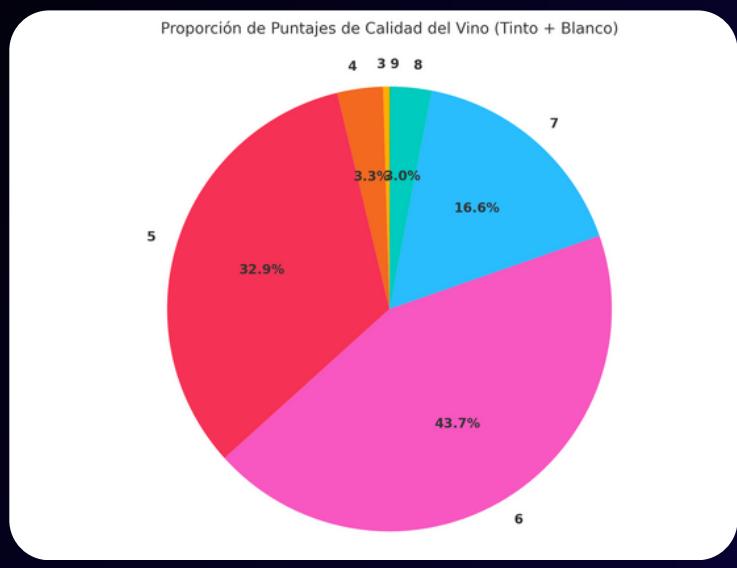
Es importante notar esto, ya que nosotros pensamos estos modelos como parte del control de calidad dentro del proceso de embotellamiento



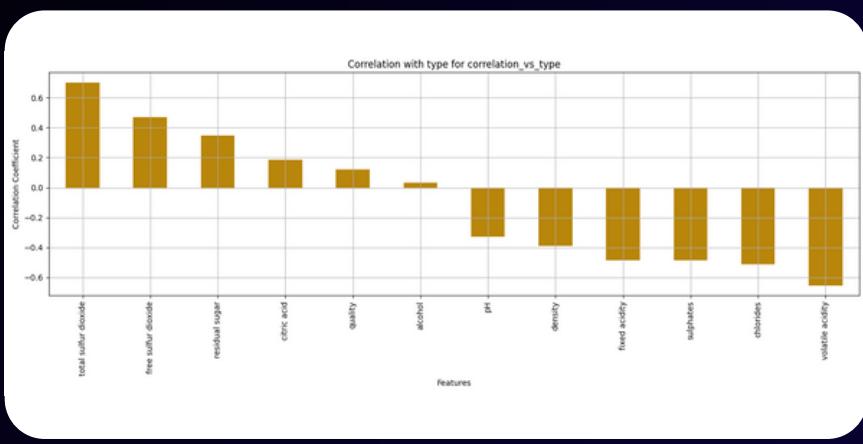
La muestra esta desbalanceada, hay muchos mas vinos blancos



Aca vemos la proporcion de calidad segun tipo de vino. Vemos que los vinos blancos estan mas presentes en calidades mas altas, pero no hay mucha diferencia en la calidad segun el tipo de vino.



Vemos que la calidad esta bastante desproporcionada, siendo 6 lo que mas tiene. Por lo tanto, como minimo a los modelos le vamos a exigir un accuracy de predecir solo 6, es decir, un accuracy de 43,7%

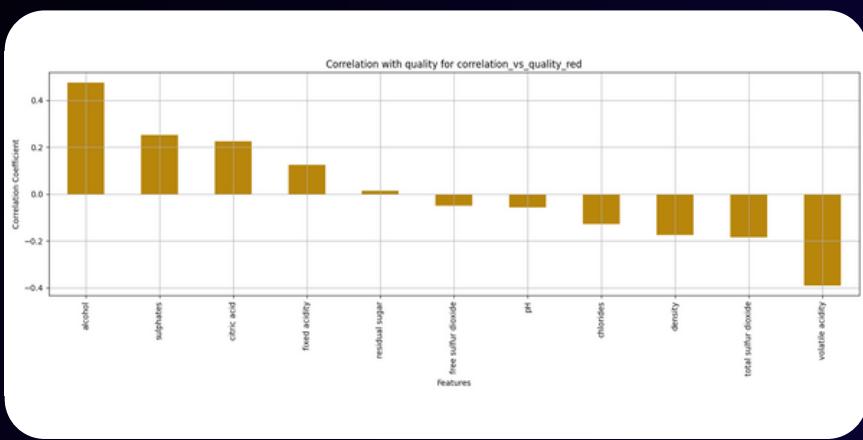


Esto usamos para hacer feature importance y elegir las variables de los modelos simples e intermedios

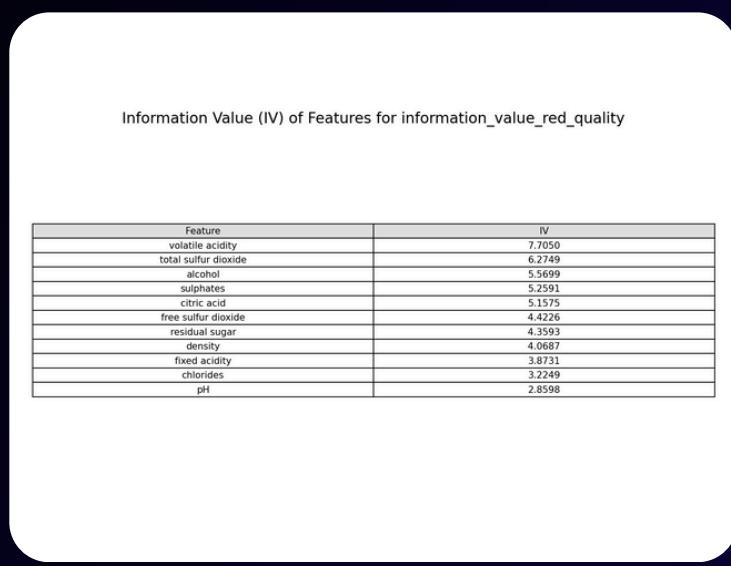
Information Value (IV) of Features for information_value_combined_type

Feature	IV
total sulfur dioxide	5.8053
chlorides	5.0505
volatile acidity	2.9440
residual sugar	2.7259
free sulfur dioxide	2.0103
sulphates	1.9836
density	1.6751
fixed acidity	1.3370
citric acid	1.0961
pH	0.6840
alcohol	0.1988

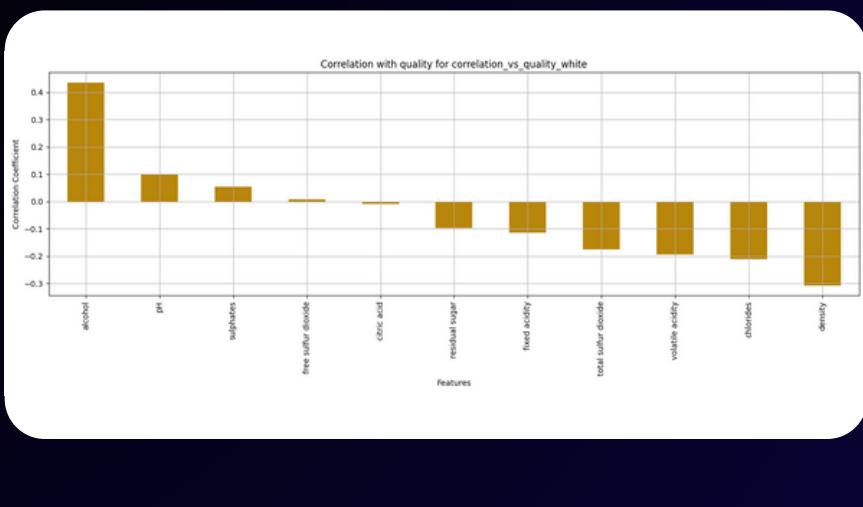
Esto usamos para hacer feature importance y elegir las variables de los modelos simples e intermedios



Esto usamos para hacer feature importance y elegir las variables de los modelos simples e intermedios



Esto usamos para hacer feature importance y elegir las variables de los modelos simples e intermedios

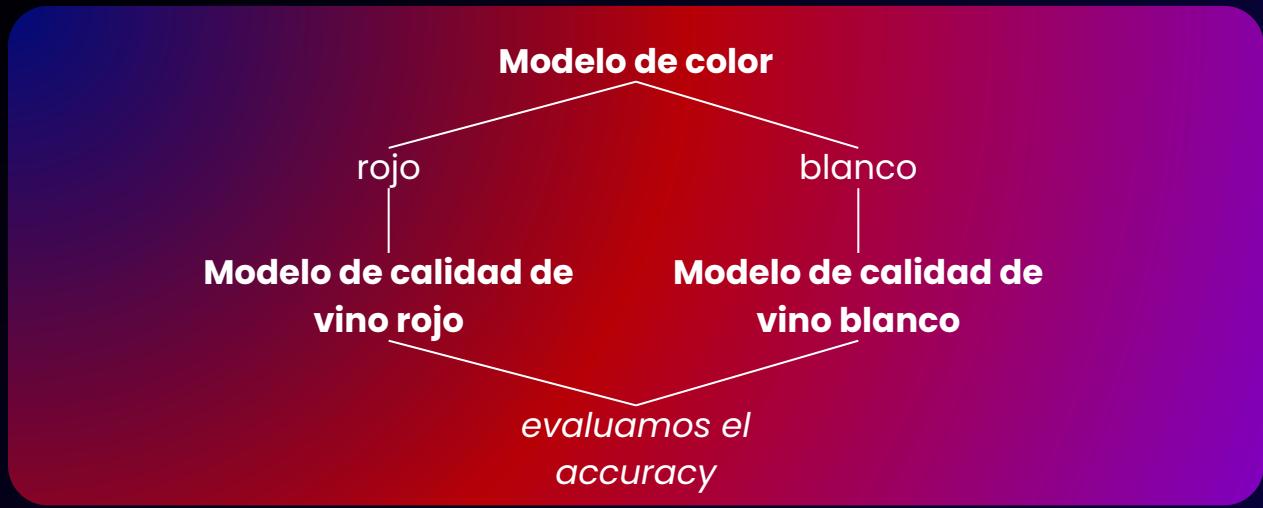


Esto usamos para hacer feature importance y elegir las variables de los modelos simples e intermedios

Information Value (IV) of Features for information_value_white_quality

Feature	IV
alcohol	7.6872
density	7.5182
chlorides	7.1089
volatile acidity	6.3086
residual sugar	6.1237
citric acid	6.1137
pH	6.1079
total sulfur dioxide	6.0780
free sulfur dioxide	6.0748
fixed acidity	5.1690
sulphates	4.8222

Esto usamos para hacer feature importance y elegir las variables de los modelos simples e intermedios



Lo que hicimos fue hacer un modelo de color, en base a la elección usar un modelo de calidad específico de cada tipo de vino, y en base a eso evaluamos el accuracy



**Redes
Secuenciales**



**Modelo
Color**

explicar que hicimos modelo simple, intermedio y completo.

el mejor en relacion a accuracy/tiempo de computo es el intermedio, usando las variables tal tal y tal

Para el dataset intermedio, hicimos un modelo con una capa oculta y un modelo con dos capas ocultas.

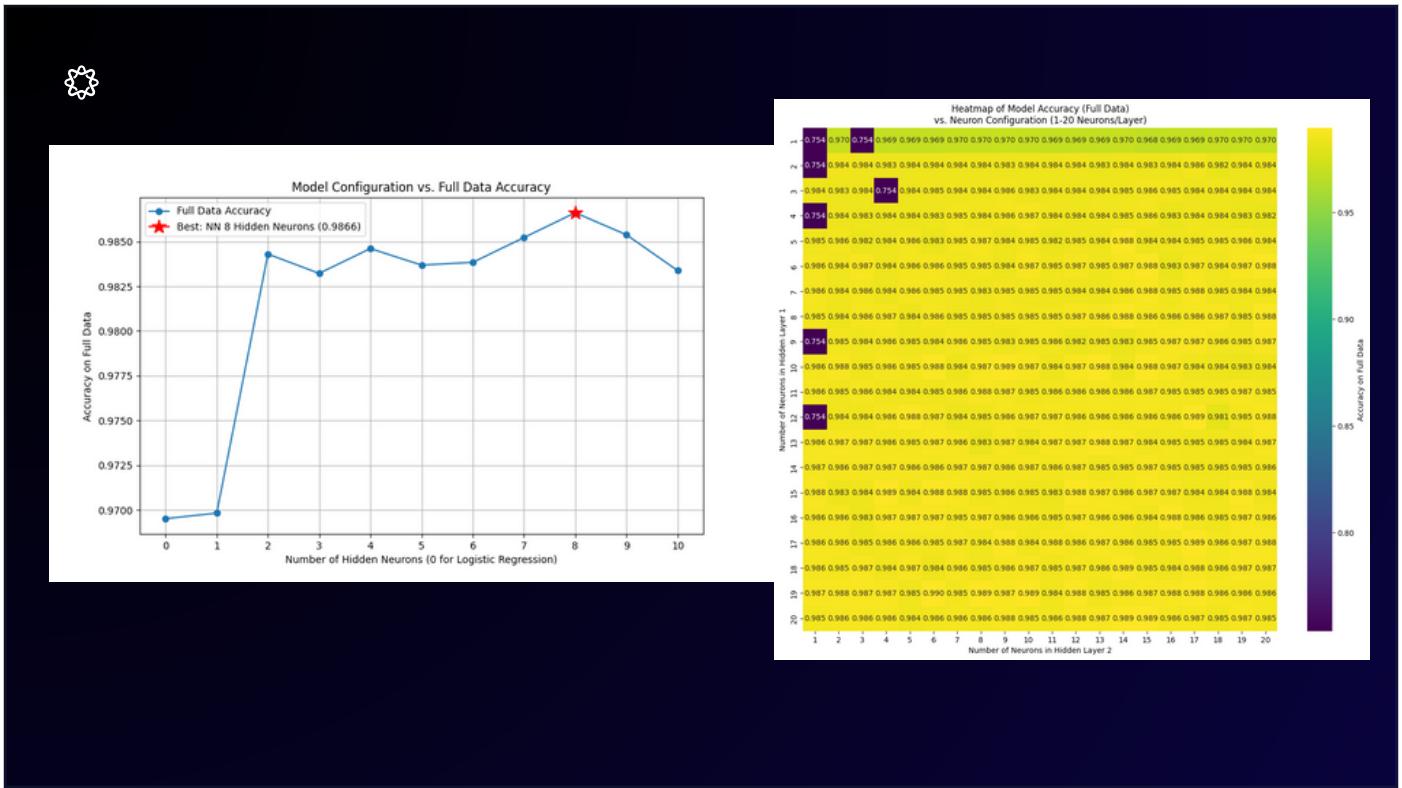
El modelo intermedio usa las variables ['total sulfur dioxide', 'chlorides', 'volatile acidity', 'residual sugar']



Parámetros de entrenamiento

- Capa de salida: Sigmoid
- Capas ocultas: Relu
- Función de pérdida: binary_crossentropy
- Optimizador: Adam
- Épocas: 100
- Tamaño del batch: 32
- Validación: split 0.8/0/2

al querer predecir tipo de vino, usamos sigmoid, ya que la variable de respuesta tipo la definimos como 0 vino rojo y 1 vino blanco.

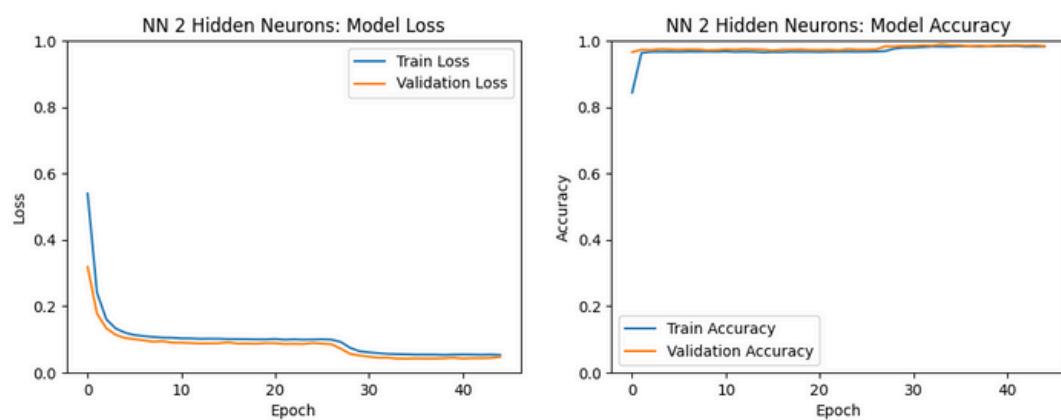


Para el dataset intermedio, hicimos un modelo con una capa oculta y un modelo con dos capas ocultas.

El mejor modelo de una capa oculta es con 8 neuronas, aunque ya con 2 neuronas tenemos un accuracy muy bueno.

Para modelos con dos capas ocultas, el mejor modelo tiene en la primer capa 19 neuronas y 6 en la segunda con un accuracy de 0.9895.

En relacion al accuracy/costo computacional, nos quedamos con el modelo intermedio de una capa oculta y 2 neuronas

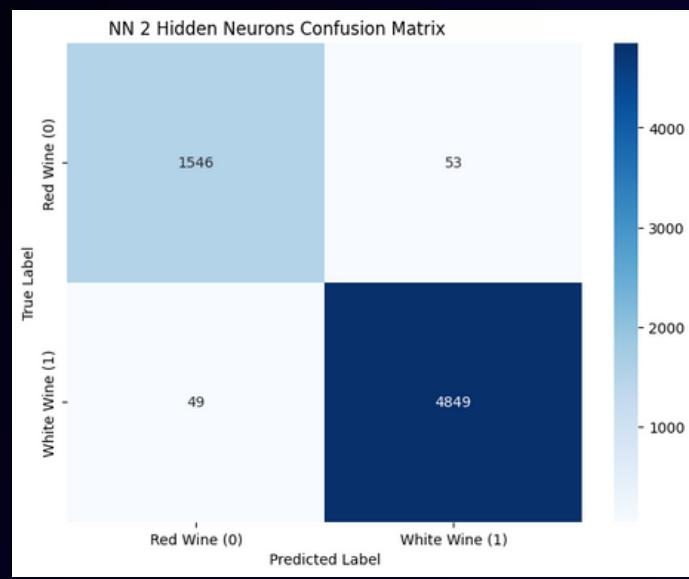


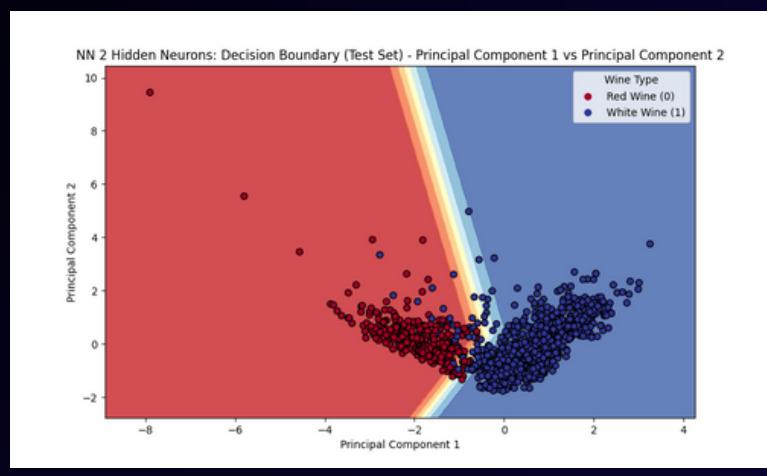
Aca vemos como el modelo intermedio de una capa con dos neuronas converge rápidamente



NN 2 Hidden Neurons Classification Metrics

	precision	recall	f1-score	support	accuracy
Red Wine (0)	0.9693	0.9669	0.9681	1599	--
White Wine (1)	0.9892	0.9900	0.9896	4898	--
macro avg	0.9792	0.9784	0.9788	6497	--
weighted avg	0.9843	0.9843	0.9843	6497	--
accuracy	--	--	--	--	0.9843







Modelos de Calidad

para medir calidad, vamos a hacer modelos que sea general (que sirva para ambos tipos de vino), otro modelo pensando en que solo se va a aplicar para vino rojo, y otro modelo que se va a aplicar para vino blanco. Para esos 3 modelos, hacemos modelos simples de 2 variables, modelos intermedios con 4 o 5 variables y modelos con todas las variables.

Lo que cambia para cada modelo es las variables que elegimos (en simple e intermedio), pero tambien el dataset que usamos para entrenar a esos modelos. Para elegir las variables usamos `iv_combined_quality`, `iv_red_quality`, y `iv_white_quality` que mostramos antes.



Modelo calidad General

Como dije, para el modelo de calidad general tenemos la versión simple, intermedia y completa.

Estos modelos fueron entrenados con ambos datasets.

Por la proporción entre accuracy y complejidad computacional, decidimos usar los modelos intermedios.

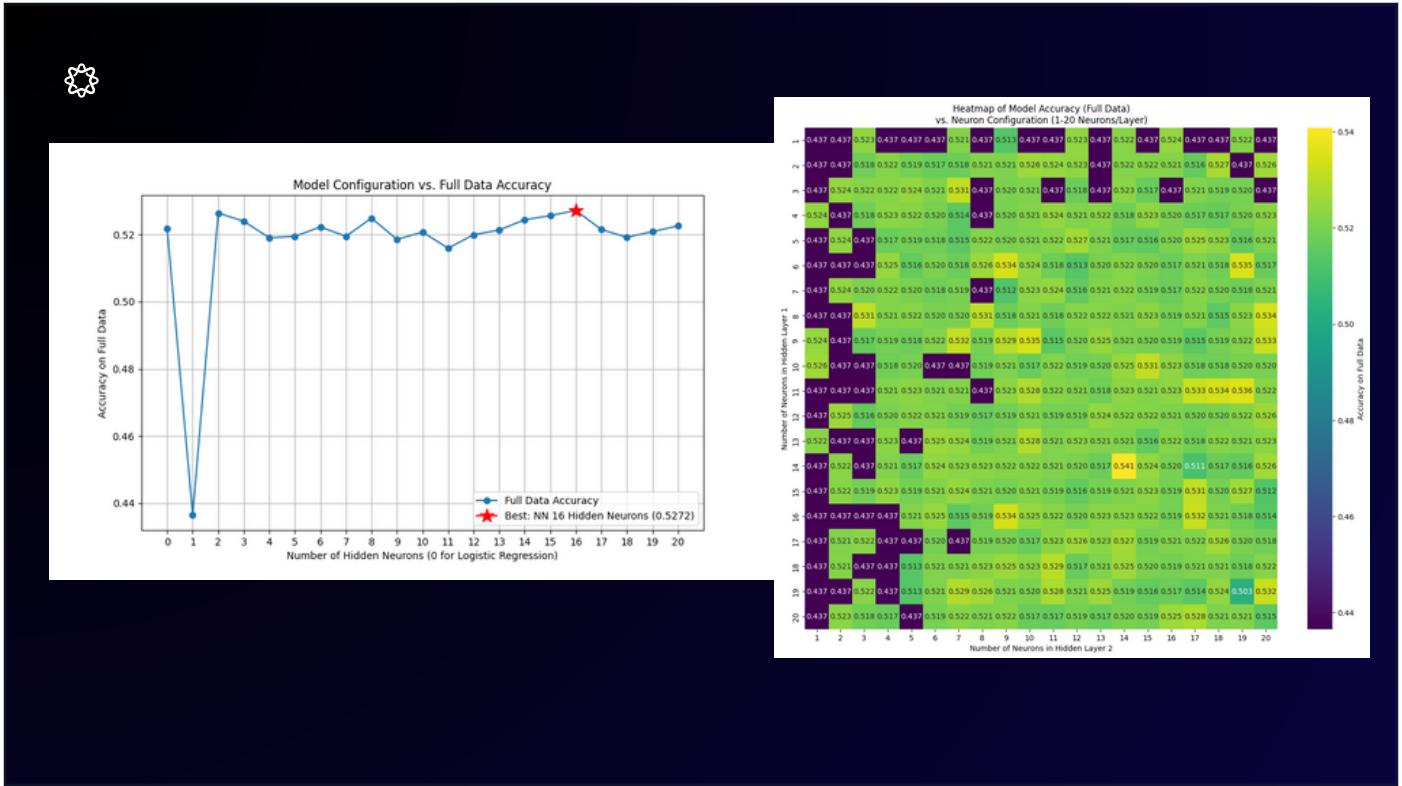
El modelo intermedio para calidad general usa las variables ['alcohol', 'density', 'chlorides', 'pH', 'volatile acidity']



Parámetros de entrenamiento

- Capa de salida: Softmax
- Capas ocultas: Relu
- Función de pérdida: categorical_crossentropy
- Optimizador: Adam
- Épocas: 100
- Tamaño del batch: 32
- Validación: split 0.8/0/2

usamos categorical crossentropy, porque la variable de respuesta es categorica. Hay 7 valores posibles para calidad

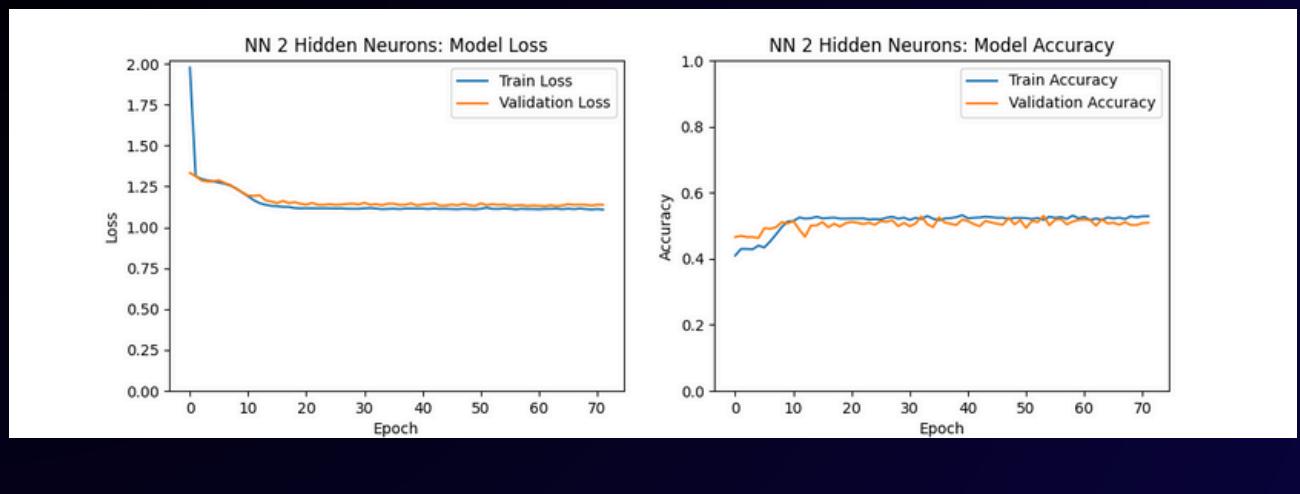


Para el dataset intermedio, hicimos un modelo con una capa oculta y un modelo con dos capas ocultas.

El mejor modelo de una capa oculta es con 16 neuronas, aunque ya con 2 neuronas tenemos un accuracy muy bueno de 0.5264.

Para modelos con dos capas ocultas, el mejor modelo tiene en la primer capa 14 neuronas y 14 en la segunda con un accuracy de 0.5409.

En relacion al accuracy/costo computacional, nos quedamos con el modelo intermedio de una capa oculta.



Aca vemos como el modelo intermedio de una capa con dos neuronas converge rápidamente

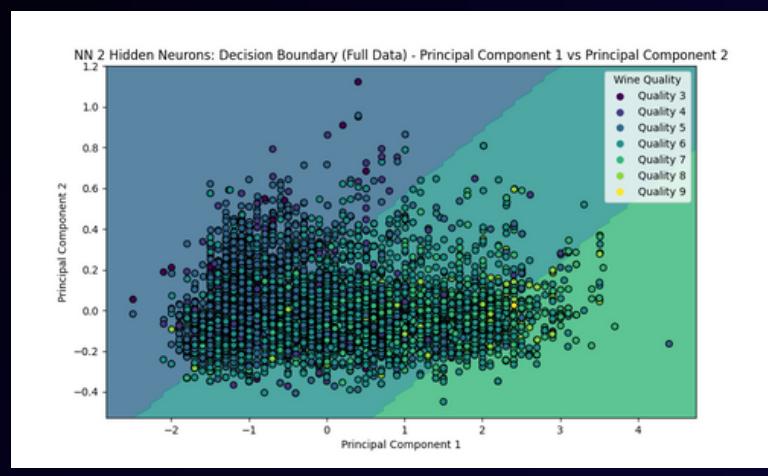


NN 2 Hidden Neurons Classification Metrics

	precision	recall	f1-score	support	accuracy
Quality 3	0.0000	0.0000	0.0000	30	--
Quality 4	0.0000	0.0000	0.0000	216	--
Quality 5	0.5844	0.5617	0.5729	2138	--
Quality 6	0.5033	0.7264	0.5946	2836	--
Quality 7	0.4556	0.1474	0.2227	1079	--
Quality 8	0.0000	0.0000	0.0000	193	--
Quality 9	0.0000	0.0000	0.0000	5	--
macro avg	0.2205	0.2051	0.1986	6497	--
weighted avg	0.4877	0.5264	0.4850	6497	--
accuracy	--	--	--	--	0.5264



		NN 2 Hidden Neurons Confusion Matrix (7d)						
		Predicted						
True Label	Q3	Q3	Q4	Q5	Q6	Q7	Q8	Q9
		0	0	14	16	0	0	0
Q4	Q4	0	0	110	106	0	0	0
Q5	Q5	0	0	1201	930	7	0	0
Q6	Q6	0	0	634	2060	142	0	0
Q7	Q7	0	0	78	842	159	0	0
Q8	Q8	0	0	18	136	39	0	0
Q9	Q9	0	0	0	3	2	0	0





**Modelo
calidad
Rojo**

Aca tambien tenemos modelos de calidad version simple, intermedia y completa.

Este modelo fue entrenado solo con el dataset de vino rojo

Por la proporcion entre accuracy y complejidad computacional, decidimos usar los modelos intermedios.

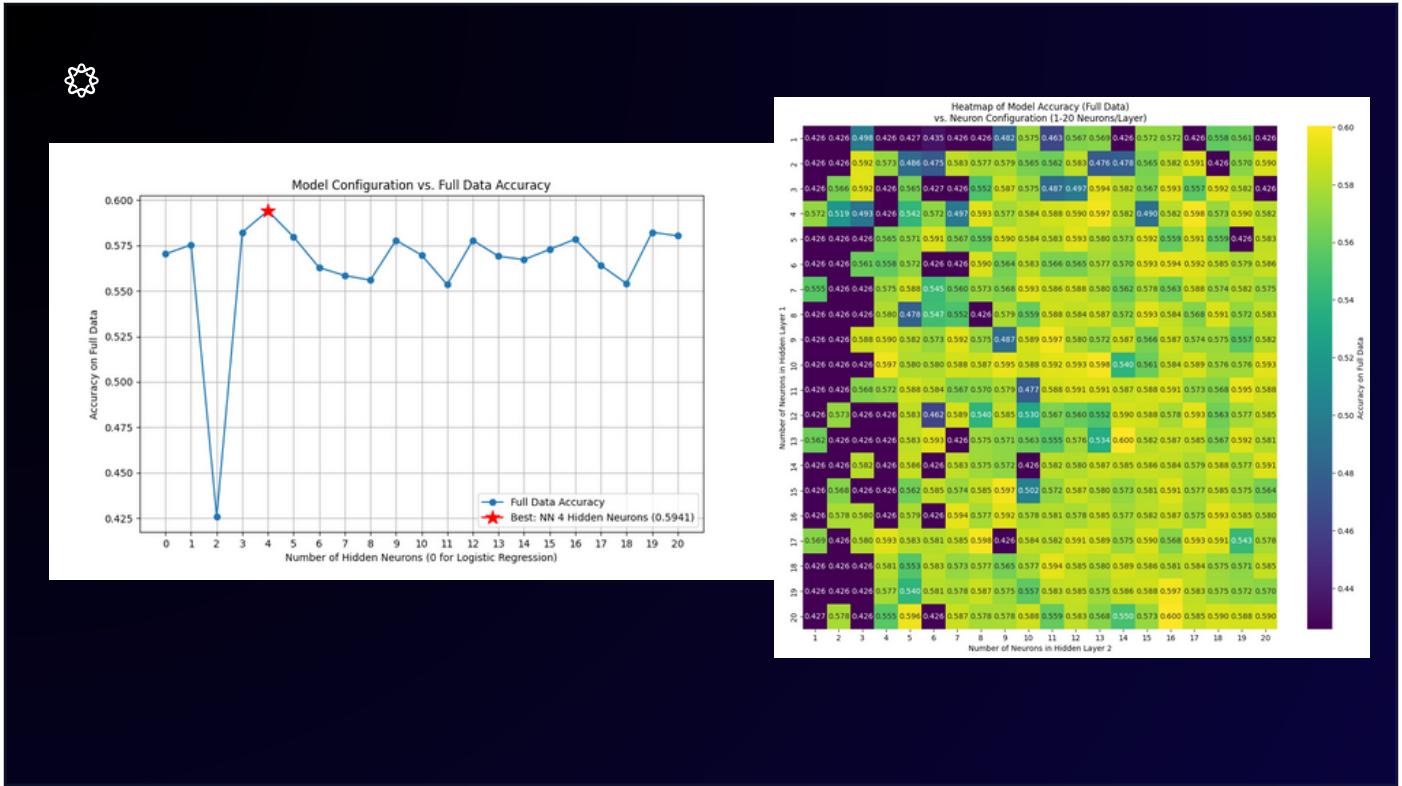
El modelo intermedio para calidad rojo usa las variables ['volatile acidity', 'total sulfur dioxide', 'alcohol', 'sulphates', 'citric acid']



Parámetros de entrenamiento

- Capa de salida: Softmax
- Capas ocultas: Relu
- Función de pérdida: categorical_crossentropy
- Optimizador: Adam
- Épocas: 100
- Tamaño del batch: 32
- Validación: split 0.8/0/2

usamos categorical crossentropy, porque la variable de respuesta es categorica. Hay 7 valores posibles para calidad

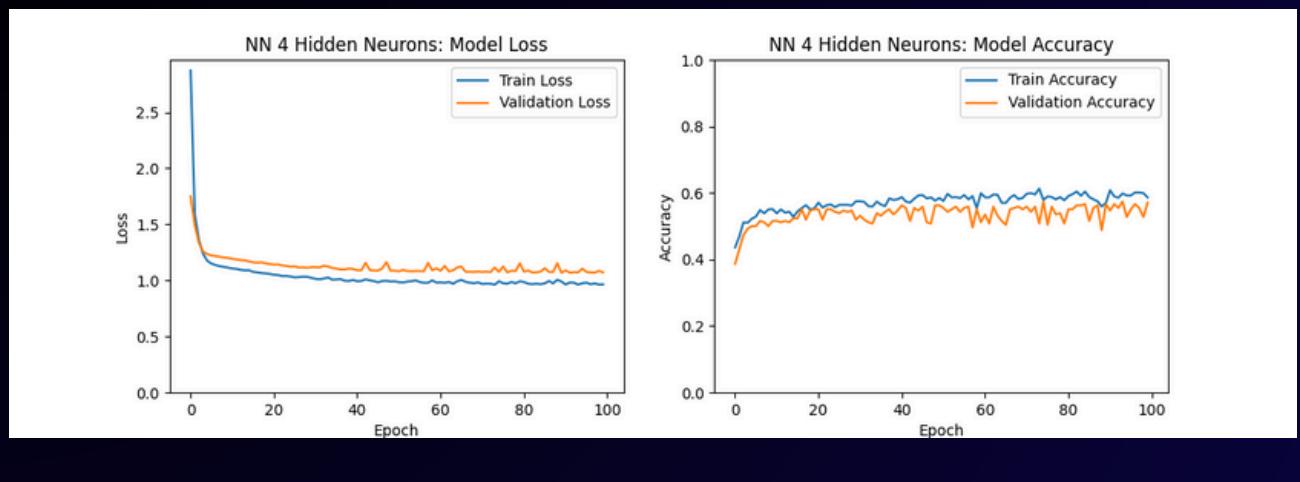


Para el dataset intermedio, hicimos un modelo con una capa oculta y un modelo con dos capas ocultas.

El mejor modelo de una capa oculta es con 4 neuronas con un accuracy de 0.5941.

Para modelos con dos capas ocultas, el mejor modelo tiene en la primer capa 13 neuronas y 14 en la segunda con un accuracy de 0.6004

En relacion al accuracy/costo computacional, nos quedamos con el modelo intermedio de una capa oculta.

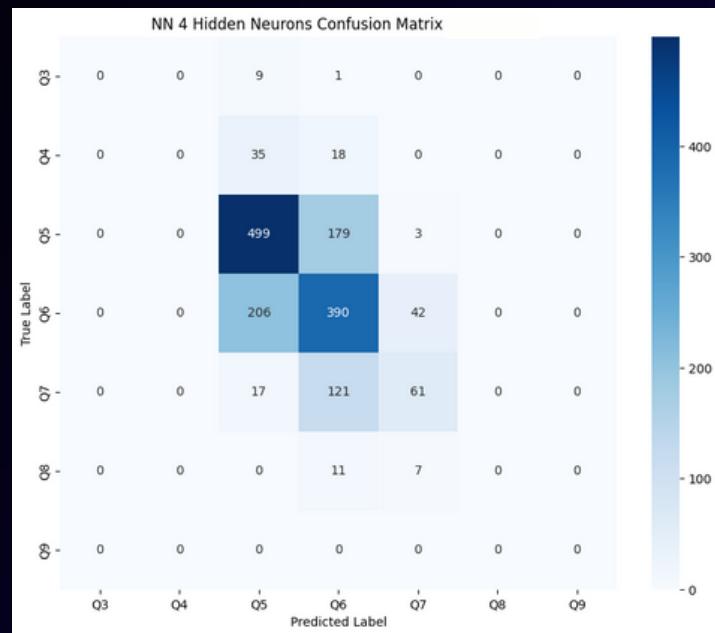


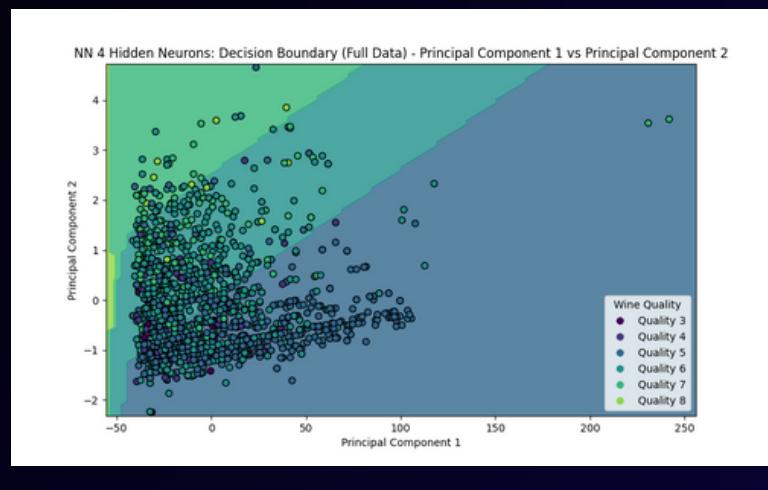
Aca vemos como el modelo intermedio de una capa con dos neuronas converge rápidamente



NN 4 Hidden Neurons Classification Metrics

	precision	recall	f1-score	support	accuracy
Quality 3	0.0000	0.0000	0.0000	10	--
Quality 4	0.0000	0.0000	0.0000	53	--
Quality 5	0.6514	0.7327	0.6897	681	--
Quality 6	0.5417	0.6113	0.5744	638	--
Quality 7	0.5398	0.3065	0.3910	199	--
Quality 8	0.0000	0.0000	0.0000	18	--
Quality 9	0.0000	0.0000	0.0000	0	--
macro avg	0.2476	0.2358	0.2364	1599	--
weighted avg	0.5607	0.5941	0.5716	1599	--
accuracy	--	--	--	--	0.5941







**Modelo
calidad
Blanco**

Aca tambien tenemos modelos de calidad version simple, intermedia y completa.

Este modelo fue entrenado solo con el dataset de vino blanco

Por la proporcion entre accuracy y complejidad computacional, decidimos usar los modelos intermedios.

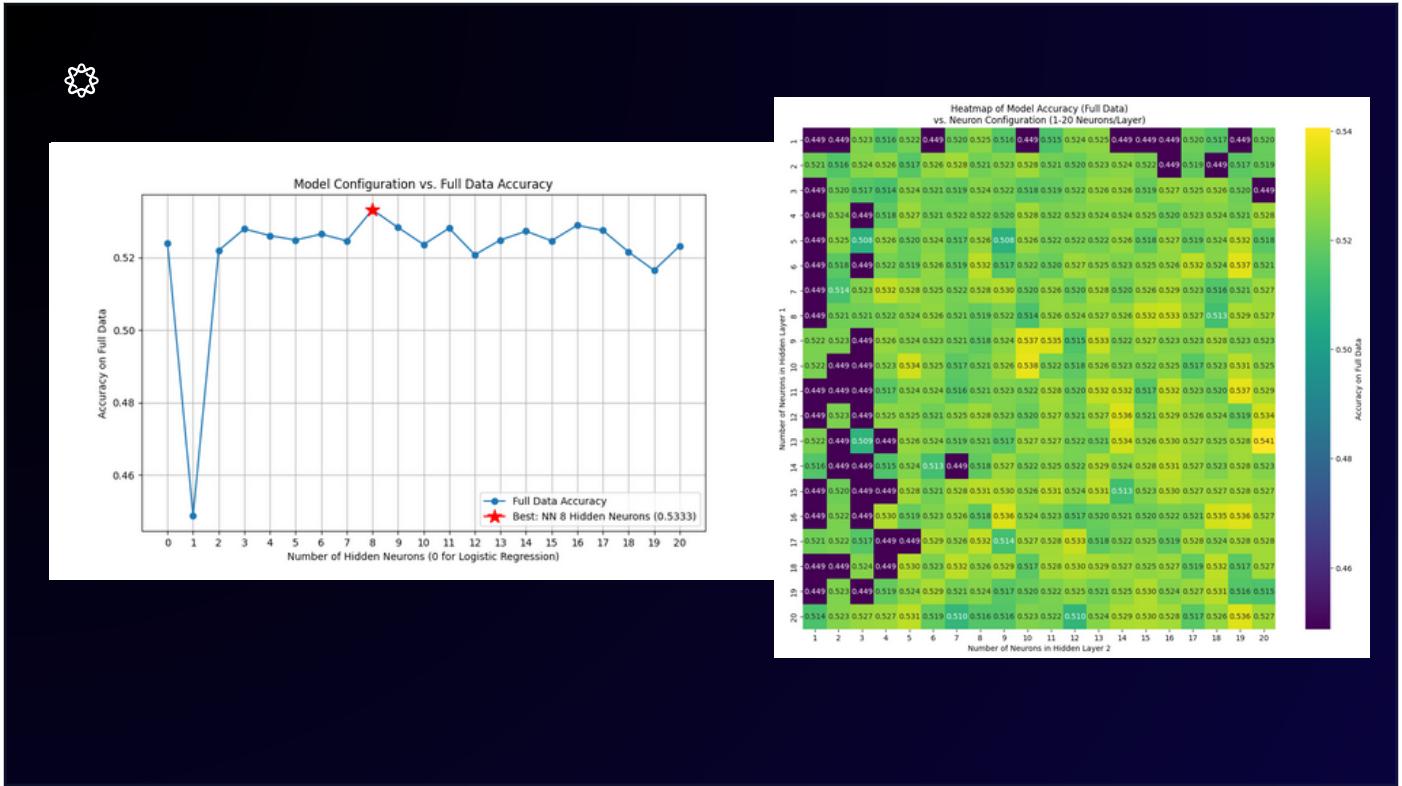
El modelo intermedio para calidad blanco usa las variables ['alcohol', 'density', 'chlorides', 'volatile acidity', 'residual sugar']



Parámetros de entrenamiento

- Capa de salida: Softmax
- Capas ocultas: Relu
- Función de pérdida: categorical_crossentropy
- Optimizador: Adam
- Épocas: 100
- Tamaño del batch: 32
- Validación: split 0.8/0/2

usamos categorical crossentropy, porque la variable de respuesta es categorica. Hay 7 valores posibles para calidad

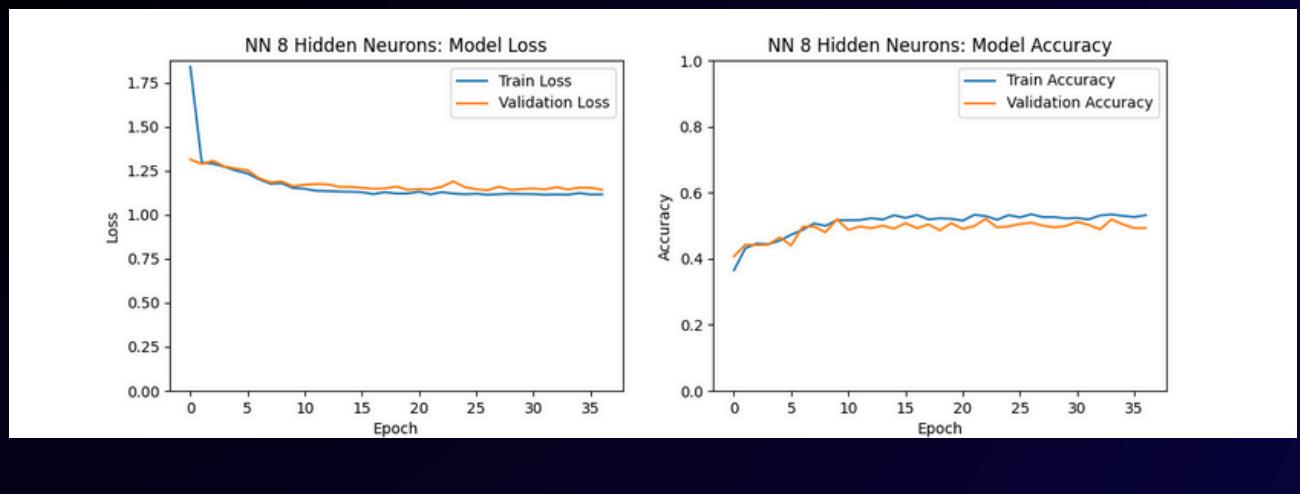


Para el dataset intermedio, hicimos un modelo con una capa oculta y un modelo con dos capas ocultas.

El mejor modelo de una capa oculta es con 8 neuronas y da un accuracy de 0.5333

Para modelos con dos capas ocultas, el mejor modelo tiene en la primer capa 13 neuronas y 20 en la segunda con un accuracy de 0.5406

En relacion al accuracy/costo computacional, nos quedamos con el modelo intermedio de una capa oculta.

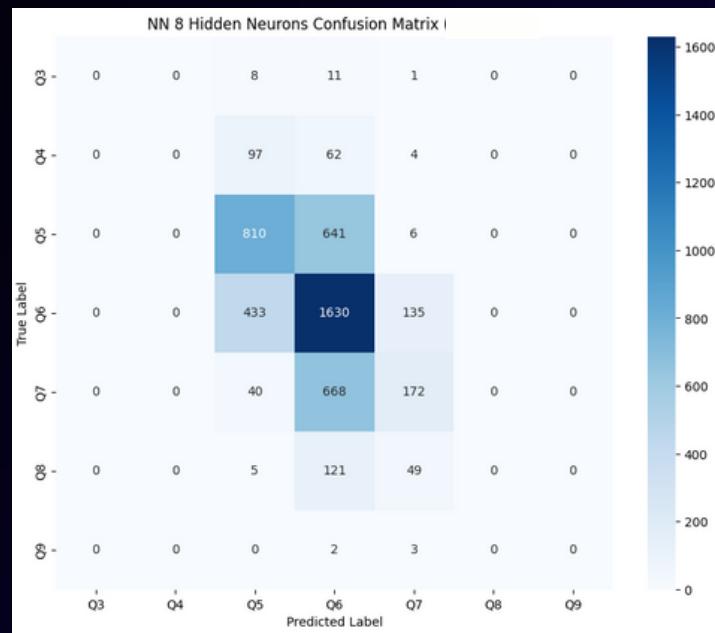


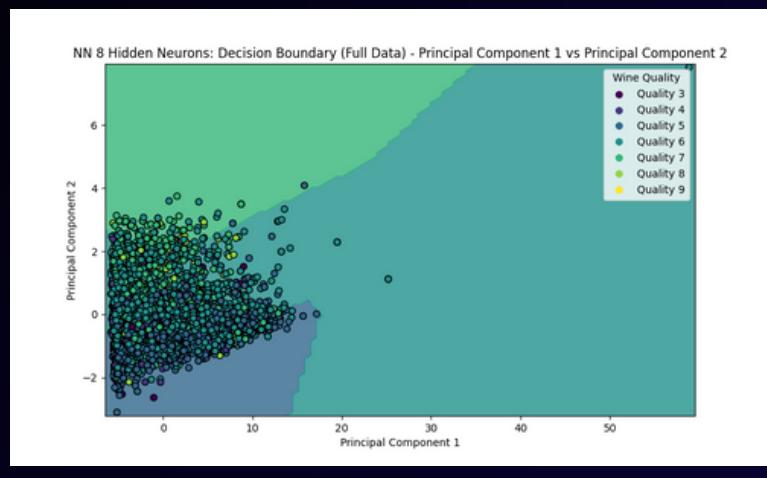
Aca vemos como el modelo intermedio de una capa con dos neuronas converge rápidamente



NN 8 Hidden Neurons Classification Metrics

	precision	recall	f1-score	support	accuracy
Quality 3	0.0000	0.0000	0.0000	20	--
Quality 4	0.0000	0.0000	0.0000	163	--
Quality 5	0.5815	0.5559	0.5684	1457	--
Quality 6	0.5199	0.7416	0.6113	2198	--
Quality 7	0.4649	0.1955	0.2752	880	--
Quality 8	0.0000	0.0000	0.0000	175	--
Quality 9	0.0000	0.0000	0.0000	5	--
macro avg	0.2238	0.2133	0.2078	4898	--
weighted avg	0.4898	0.5333	0.4928	4898	--
accuracy	--	--	--	--	0.5333







Conclusiones

Si promediamos la proporcion de cada tipo de vino con su accuracy, sacamos una accuracy ponderada que da 0.5482.

Es decir, el metodo de predecir el color y en base a eso usar el modelo correcto dio un accuracy de 0.5482. Esto implica una mejor sobre el modelo general.

Entonces, fue mejor calcular primero el tipo y en base a eso usar un modelo especifico de ese tipo de vino para predecir su calidad que usar un modelo generico.

Eso puede ser debido a que el modelo de color tiene un accuracy demasiado alto.



Modelos Funcionales

Al ser un problema con dos variables a predecir, podemos usar un modelo funcional.

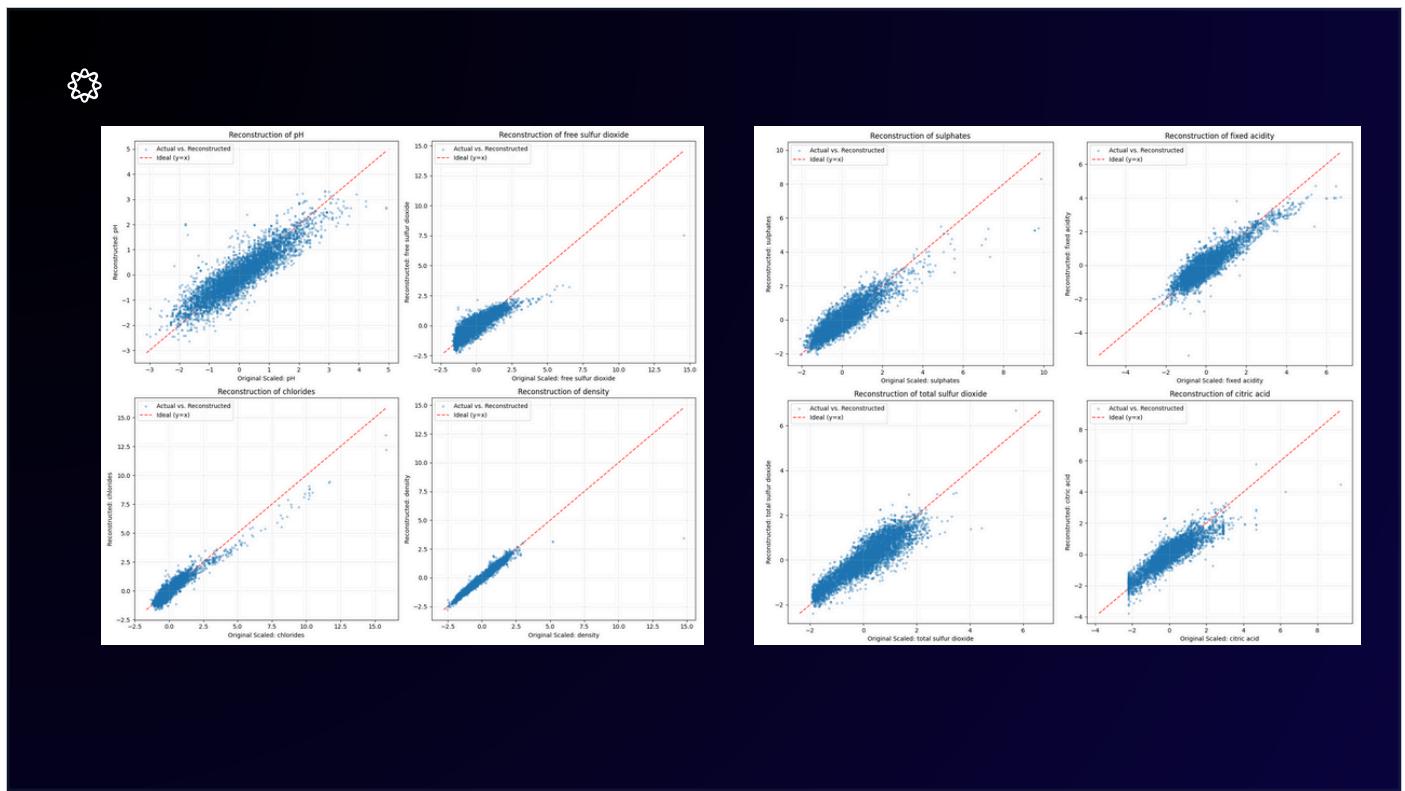
Y en este caso, en vez de hacer modelo completo y modelo intermedio, vamos a usar autoencoder para reducir la cantidad de variables a usar.



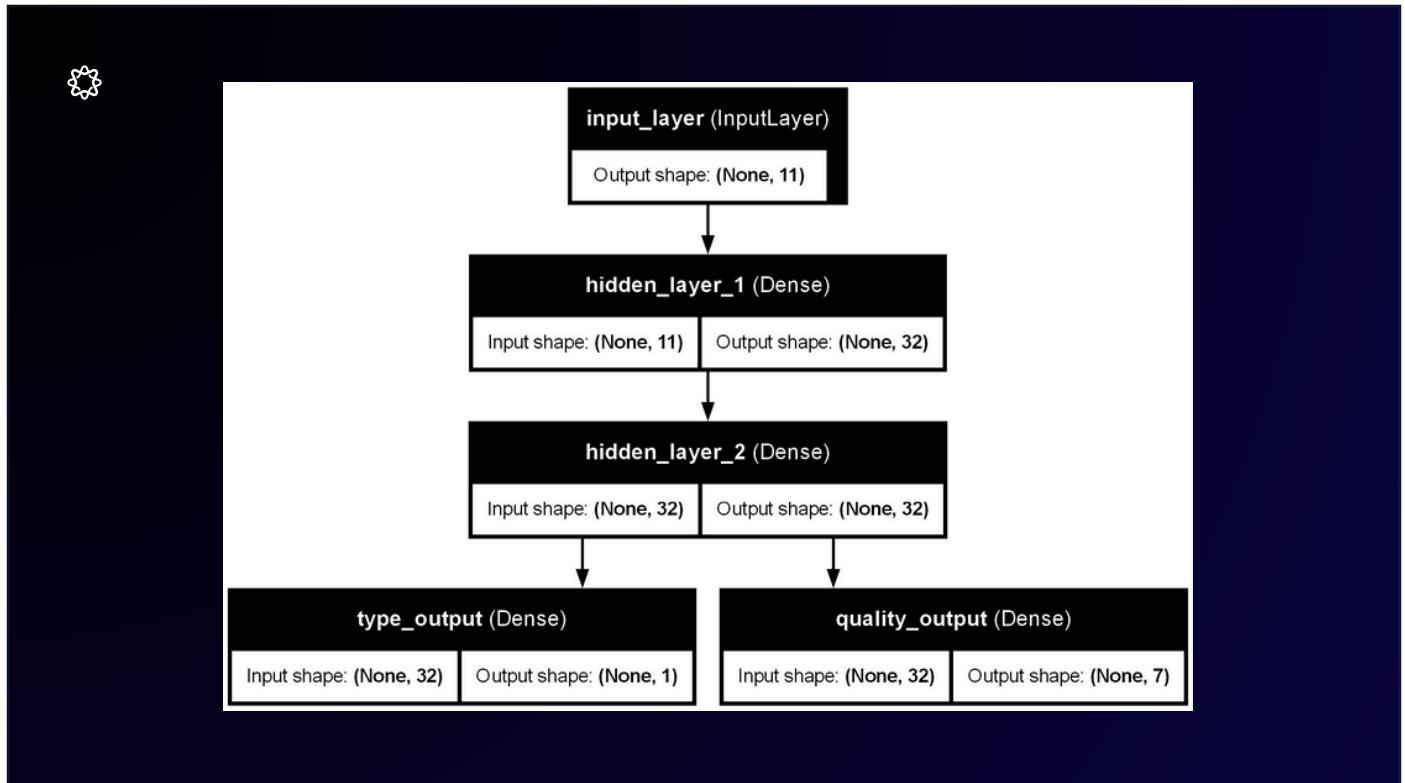
Feature Variance Details: Original vs. Reconstructed

Feature	Original Scaled Variance	Reconstructed Variance	Variance Retained (%)
fixed acidity	1.0000	0.8042	80.4200%
volatile acidity	1.0000	0.6901	69.0100%
citric acid	1.0000	0.7727	77.2700%
residual sugar	1.0000	0.7500	75.0000%
chlorides	1.0000	0.8540	85.4000%
free sulfur dioxide	1.0000	0.7394	73.9400%
total sulfur dioxide	1.0000	0.8034	80.3400%
density	1.0000	0.9411	94.1100%
pH	1.0000	0.7763	77.6300%
sulphates	1.0000	0.8050	80.5000%
alcohol	1.0000	0.7071	70.7100%

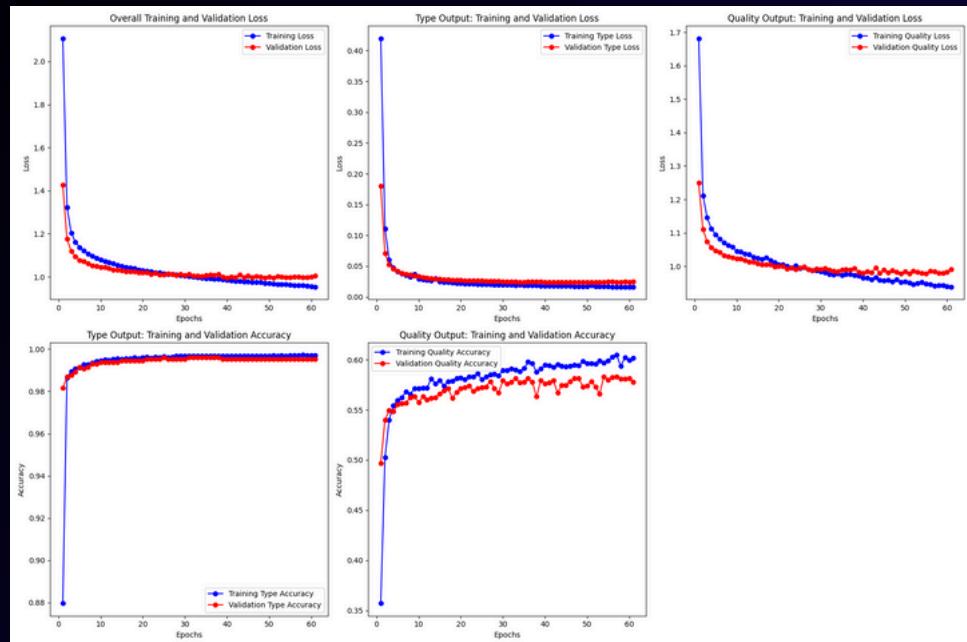
aca vemos que el autoencoding no nos saco tanta varianza



aca vemos que el autoencoding no nos saco tanta varianza. Lo ideal seria que los puntos sigan la linea roja



El modelo que usamos fue de 32 neuronas en cada capa oculta. Nos hubiese gustado hacer un análisis de sensibilidad como en el resto de modelos.



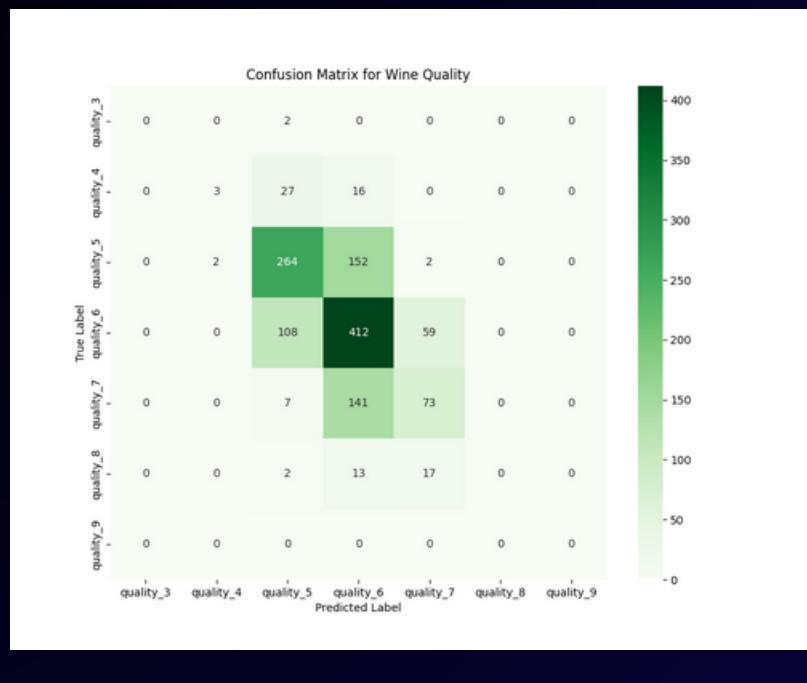


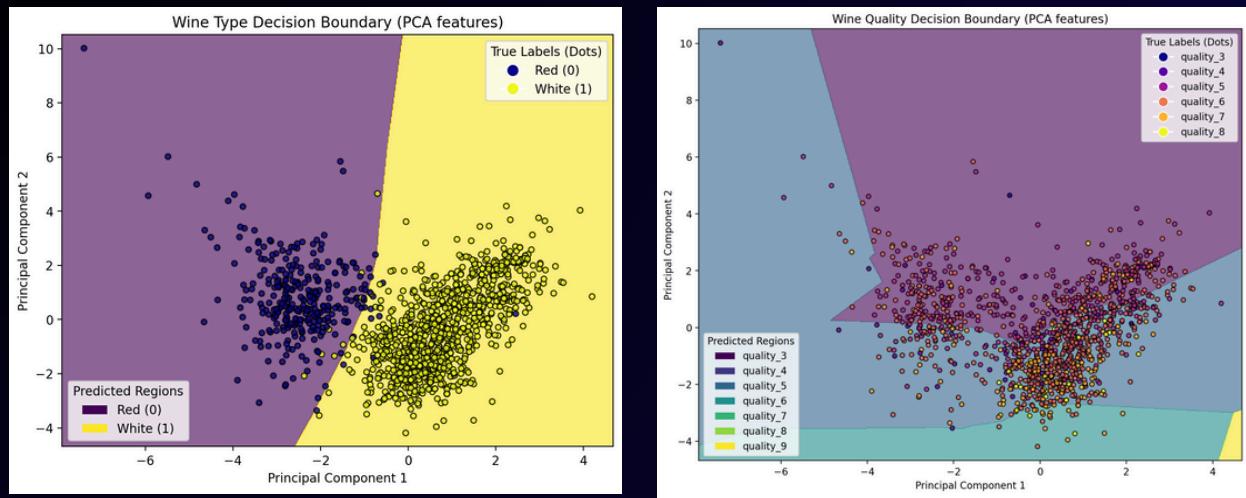
Classification Report for Wine Quality

	precision	recall	f1-score	support
quality_3	0.0	0.0	0.0	2.0
quality_4	0.6	0.065	0.118	46.0
quality_5	0.644	0.629	0.636	420.0
quality_6	0.561	0.712	0.628	579.0
quality_7	0.483	0.33	0.392	221.0
quality_8	0.0	0.0	0.0	32.0
quality_9	0.0	0.0	0.0	0.0
accuracy	nan	nan	0.578	1300.0
macro avg	0.327	0.248	0.253	1300.0
weighted avg	0.561	0.578	0.556	1300.0

Classification Report for Wine Type

	precision	recall	f1-score	support
Red (0)	0.994	0.987	0.99	314.0
White (1)	0.996	0.998	0.997	986.0
accuracy	nan	nan	0.995	1300.0
macro avg	0.995	0.993	0.994	1300.0
weighted avg	0.995	0.995	0.995	1300.0







Conclusiones

el mejor modelo es el que mas variables tiene



Tipo de modelo	Accuracy
Color	0.9843
Calidad General	0.5406
Calidad vino tinto	0.5941
Calidad vino blanco	0.5333
Concatenacion color y calidad especifica	0.5482
Secuencial calidad	0.5780
Secuencial color	0.9950



Conclusión

La concatenación del modelo de color y calidad específica mejora el accuracy contra el modelo general, aunque no de manera significativa. Aun así, usamos eso porque es lo mas realista.

Nuestra forma de concatenación de modelo es un intento rudimentario de redes funcionales.

El modelo funcional no solo es mas eficiente en términos de complejidad computacional, sino que obtiene una mejor accuracy

Como conclusión entonces, vemos como el modelo funcional es mejor que hacer todo lo que hicimos.

Lo que podemos rescatar es que la concatenación de modelos que hicimos manualmente fue mejor que el modelo general de calidad, aunque la mejora no es significativa. Aun así, usamos eso porque es lo mas realista.

Nuestro intento de hacer primero color y en base a eso hacer el modelo de calidad específico es una forma muy rudimentaria de hacer un modelo funcional.

Gracias

Martín Quijano - Martina Coletto