Proyecto 2 – Redes Neuronales

Juan Esteban Urquijo Huerfano

Santiago Zúñiga Martínez

Iván Alejandro Martínez Gracia

juan-urquijoh@javeriana.edu.co

s.zunigam@javeriana.edu.co

iamartinezg@javeriana.edu.co

Introducción a la inteligencia artificial

Pontificia Universidad Javeriana

Bogotá D.C, Colombia

2022

Comprensión del dataset

- ¿Qué información presenta el dataset?

Se escogió el dataset de los vinos tintos y tiene como información todas las características de la composición química que puedan llegar a tener estos.

- ***Acidez Fija:*** Conjunto de ácidos naturales del vino (tartárico, málico, cítrico, succínico y láctico). Gracias a estos se preservan las cualidades del vino.
- ***Acidez Volátil:*** La cantidad de ácido acético (vinagre) que tiene el vino, altamente indeseable y se realizan esfuerzos para que sea mínimo.
- ***Ácido Cítrico:*** Presente en las frutas ácidas, pero de escasez en la uva, aporta frescura y toque amargo.
- ***Azúcar residual:*** La cantidad total de azúcar que queda en el vino y que no ha podido poder ser fermentado por las levaduras.
- ***Cloruros:*** Principales componentes de sales en el vino realzan los sabores
- ***Dióxido de azufre libre***: Como agente antibacteriano y que evita la modificación del color del vino.
- ***Dióxido de azufre total***: Usado como agente antioxidante y conservante que evita la formación de desperfectos. Sin embargo, está regulado y no debe exceder 210 mg/L por regulación
- ***Densidad:*** Se busca que la densidad del vino está muy cerca a la del agua 1g/ml.
- ***pH:*** Mide la acidez del vino, por lo general el vino tinto se ubica entre 3.3 y 3.6.
- ***Sulfatos:*** Conservantes, antioxidantes, agentes antimicrobianos y antioxidásicos
- ***Alcohol:*** El vino es una bebida fermentada poseyendo una graduación alcohólica entre 3.5 y 15 grados.
- ***Calidad:*** Es un valor de 0 a 10 que se le asigna al vino de acuerdos a las demás características presentadas previamente.

- Elección de la variable

La variable sobre la que se va a realizar la estimación es la calidad del vino en base a las otras variables. Con los datos de las demás variables, se realizará la predicción de la calidad del vino de acuerdo con tres modelos diferentes: perceptrón, una red neuronal con una capa oculta con un numero de neuronas igual al número de entradas y otra red neuronal con dos capas oculta con dos neuronas en cada capa oculta.

- Análisis Gráfico

- Correlación entre variables

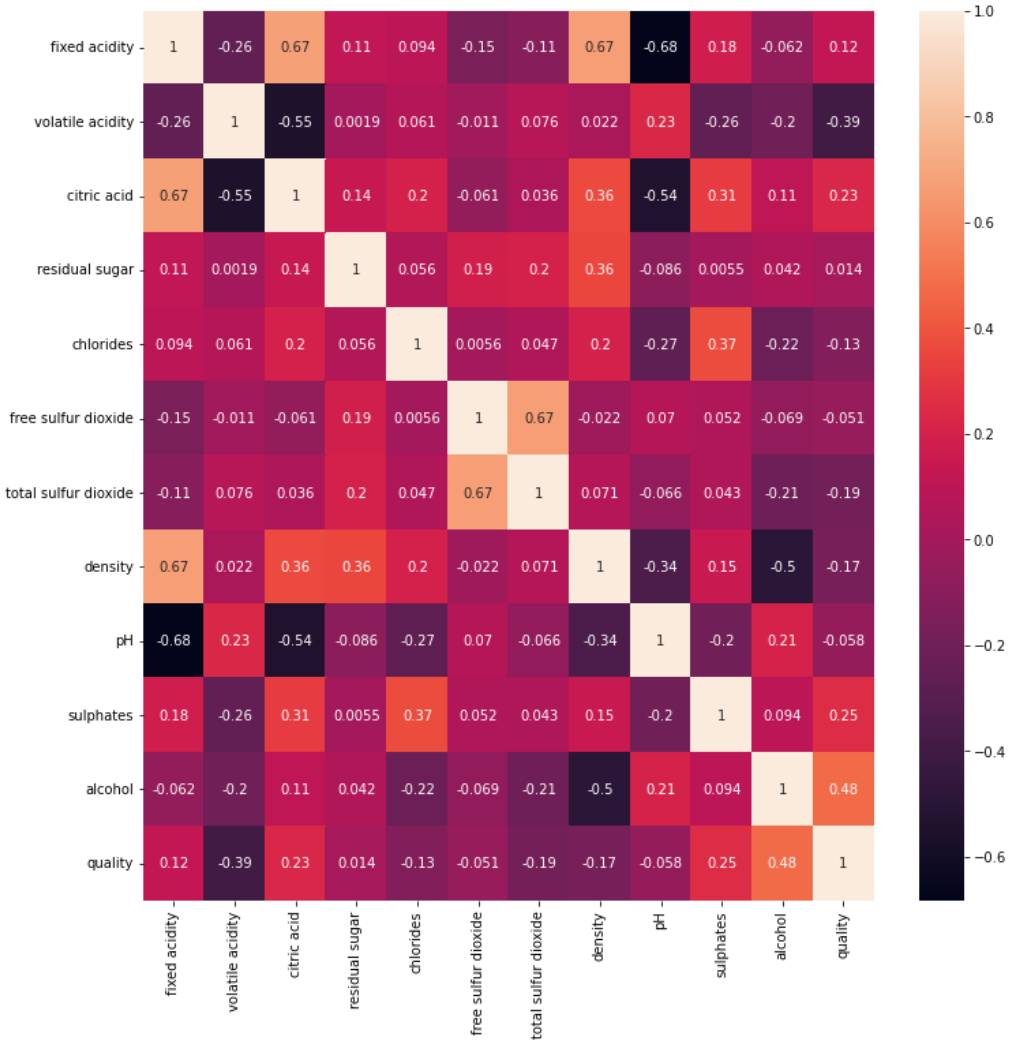

Figura 1. Matriz de correlación entre variables

La matriz de correlación es bastante util para ver la relación entre las variables del dataset. Esta muestra los valores de correlación que miden el grado de relación lineal entre cada par de variables. Los valores de correlación se pueden ubicar entre -1 y +1. Si las dos variables tienden a aumentar o disminuir al mismo tiempo, el valor de correlación es positivo.

En este caso, como escogimos la variable de calidad para el vino tinto, es interesante conocer aquellas variables están altamente relacionadas con esta y así entender como es el comportamiento de los datos. Como se puede ver en la matriz, la calidad y el alcohol están considerablemente correlacionadas con un valor de 0.48, de igual manera la calidad y los sulfatos poseen una correlación de 0.25 y la calidad y la acidez fija tienen una correlación de 0.12. Estos valores no son bastante altos porque no se acercan al 1 pero dan muestra de que hay correlación débil entre cada par de variables, pero son las que más se correlacionan frente a las demás variables. De igual

manera en el caso de la correlación de calidad y acido volátil es de -0.39 lo que indica que entre más acidez volátil puesta en el vino asocia una peor calidad.

- Distribución de las variables

Para realizar un análisis de como se distribuyen los valores de cada variable, se realizo un histograma y se analizó que valores se repetían con frecuencia para poder entender como afecta esto en la calidad del vino que es la variable que se escogió previamente. Como se puede evidenciar, los datos están bastante agrupados para cada variable a excepción de del acido cítrico o el alcohol.

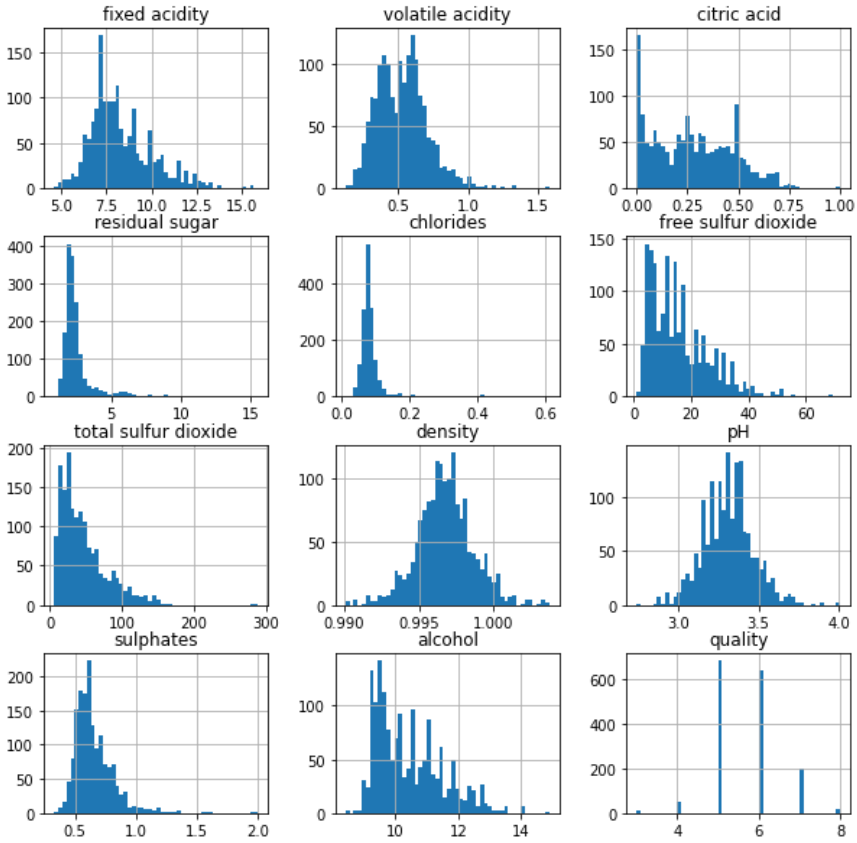

Figura 2. Histograma de las variables del dataset

- Diagrama de Pares (Pairplot)

Finalmente, para dar un resumen e investigar de manera rápida la relación cruzada entre todas las variables presentes en el dataset usamos es un gráfico de pares o parejas, el cual muestra la relación entre los pares de todas las variables y como es su distribución gráfica. Esto nos permitió analizar que tan agrupados estaban los datos, las fuertes relaciones entre variables y de igual manera su dispersión.

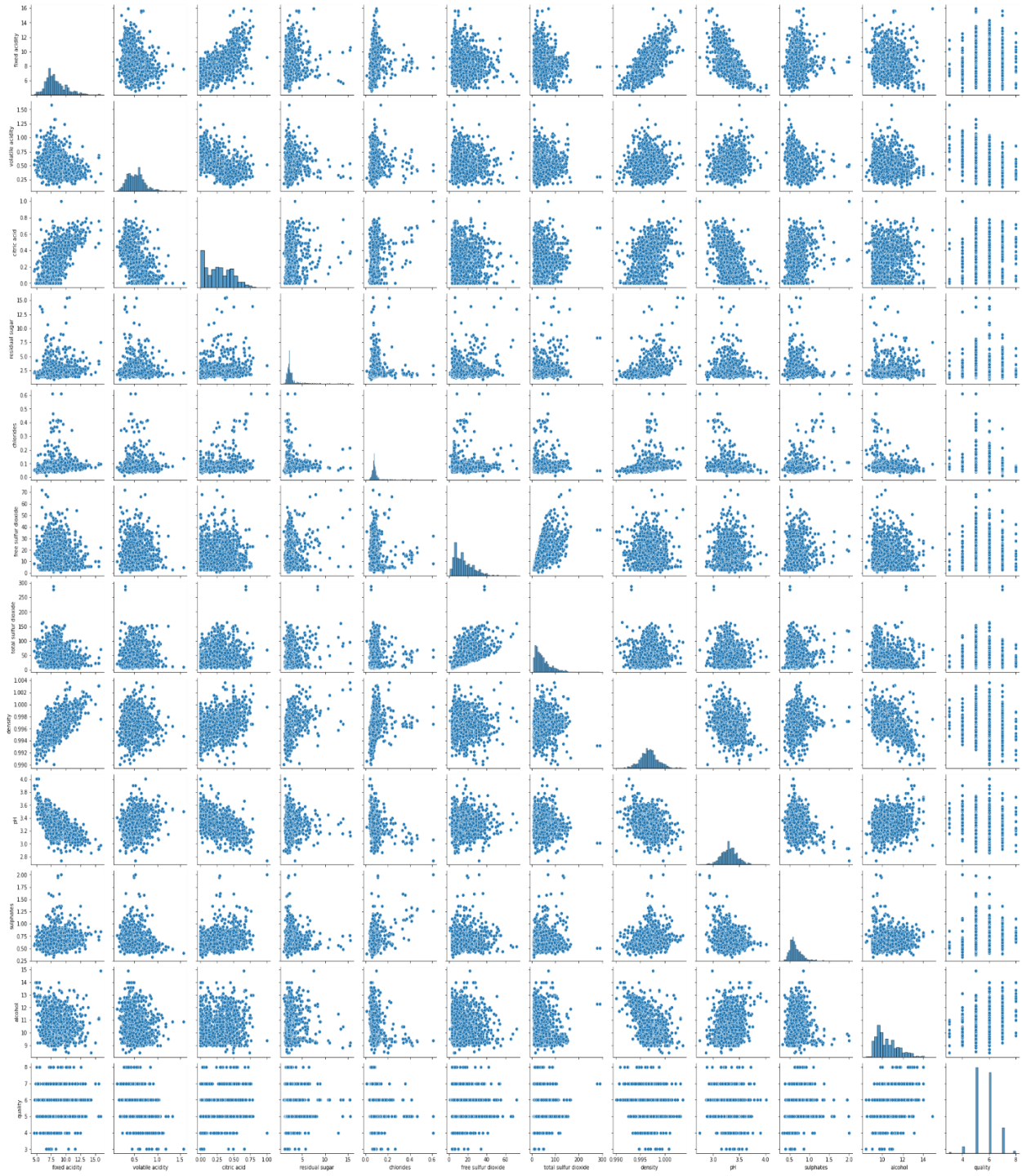

Figura 3. Diagrama de pares del dataset

Construcción del dataset

- ¿Qué proporción del conjunto de entrenamiento y del conjunto de pruebas?

El conjunto de entrenamiento y prueba tendrá una razón de 80/20

- ¿Cambiarían los resultados si se modifica la proporción?

En teoría, al ser un dataset relativamente pequeño (1600) el cambio en la muestra de entrenamiento debería afectar considerablemente el modelo, sin embargo, si se observa detalladamente la columna calidad, los valores de esta tienen una dispersión mínima, por lo que compensa altamente esto haciendo que, si se afecte el modelo, pero a escalas mínimas haciendo que el resultado en práctica siga siendo relativamente bueno.

Elaboración del modelo

Se diseñaron diferentes soluciones empleando diferentes arquitecturas para el dataset de vino tinto.

• Perceptrón

• Red neuronal con una capa oculta con un numero de neuronas igual al número de entradas con función de activación sigmoide

• Red neuronal con dos capas oculta con dos neuronas en cada capa oculta función de activación sigmoide

Luego de construir los modelos se busco realizar una predicción de la calidad de vino teniendo en cuenta los datos para entrenamiento y los datos de prueba seleccionados previamente, así como el uso de las demás variables presentes en el dataset. En el siguiente apartado se muestran los resultados y el análisis para cada uno de los modelos.

Análisis de resultados

Para evaluar la calidad de los modelos hechos previamente, se usó una matriz de confusión y un reporte de clasificación que contiene unas métricas comunes que sirven para entender que tan bueno fue el modelo que se construyó. Algunas de estas métricas son acuraccy, precisión, recall y f1 score.

- Modelo Perceptrón

```
array([[ 0,  0,  2,  0,  0,  0],
       [ 0,  0,  8,  3,  0,  0],
       [ 0,  2, 60, 72,  1,  0],
       [ 0,  4, 55, 65, 17,  1],
       [ 0,  0, 13,  4, 10,  0],
       [ 0,  0,  3,  0,  0,  0]])
```

Figura 4. Matriz de confusión en perceptrón

	precision	recall	f1-score	support
3	0.00	0.00	0.00	0
4	0.00	0.00	0.00	6
5	0.44	0.43	0.43	141
6	0.46	0.45	0.45	144
7	0.37	0.36	0.36	28
8	0.00	0.00	0.00	1
accuracy			0.42	320
macro avg	0.21	0.21	0.21	320
weighted avg	0.43	0.42	0.43	320

Figura 5. Reporte de clasificación en perceptrón

Para el modelo perceptrón se obtuvieron los resultados que se ven en la imagen, la métrica *accuracy* indica que tan acertado fue el modelo frente a las predicciones hechas, *precision* indica el porcentaje de predicciones positivas correctas respecto al total de predicciones positivas, el *recall* indica porcentaje de predicciones positivas correctas en relación con el total de positivos reales y el *f1-score* es una media armónica ponderada de precisión y recuperación. Cuanto más cerca de 1, mejor es el modelo.

Para este modelo se obtuvo una precisión del 42% lo que indica pues que el modelo no es tan bueno realizando las predicciones de vino, se obtuvo un *recall* del 43% que indica que de todas las prtedicciones hechas solo predijo correctamente para dicho porcentaje de esos datos y finalmente el *f1-score* nos dio un valor de 0.45 para los valores de 6 en la calidad de vino y 0.43 para valores de 5 lo que indica que el modelo no es tan bueno porque no se acerca a 1.

- Red Neuronal 1

```
array([[  0,   0,   2,   0,   0,   0],
       [  0,   1,  10,   0,   0,   0],
       [  0,   2, 123,  10,   0,   0],
       [  0,   2,  94,  46,   0,   0],
       [  0,   0,   9,  18,   0,   0],
       [  0,   0,   0,   3,   0,   0]])
```

Figura 6. Matriz de confusión en la red neuronal 1

	precision	recall	f1-score	support
3.0	0.00	0.00	0.00	0
4.0	0.09	0.20	0.13	5
5.0	0.91	0.52	0.66	238
6.0	0.32	0.60	0.42	77
7.0	0.00	0.00	0.00	0
8.0	0.00	0.00	0.00	0
accuracy			0.53	320
macro avg	0.22	0.22	0.20	320
weighted avg	0.76	0.53	0.59	320

Figura 7. Reporte de clasificación en la red neuronal 1

Para este modelo se obtuvo una precisión del 91% para valores de 5 en la calidad del vino lo que indica pues que el modelo es bueno realizando las predicciones de vino, se obtuvo un *recall* del 52% que indica que de todas las predicciones hechas solo predijo correctamente para dicho porcentaje de esos datos y finalmente el *f1-score* nos dio un valor de 0.66 para los valores de 5 en la calidad de vino y 0.42 para valores de 6 lo que indica que el modelo es considerablemente bueno realizando las predicciones.

- Red Neuronal 2

```
array([[  0,   0,   2,   0,   0,   0],
       [  0,   1,   9,   1,   0,   0],
       [  0,   0, 114,  21,   0,   0],
       [  0,   1,  99,  42,   0,   0],
       [  0,   0,   8,  19,   0,   0],
       [  0,   0,   0,   3,   0,   0]])
```

Figura 8. Matriz de confusión en la red neuronal 2

	precision	recall	f1-score	support
3.0	0.00	0.00	0.00	0
4.0	0.09	0.50	0.15	2
5.0	0.84	0.49	0.62	232
6.0	0.30	0.49	0.37	86
7.0	0.00	0.00	0.00	0
8.0	0.00	0.00	0.00	0
accuracy			0.49	320
macro avg	0.21	0.25	0.19	320
weighted avg	0.69	0.49	0.55	320

Figura 9. Reporte de clasificación en la red neuronal 2

Para este modelo se obtuvo una precisión del 49% lo que indica pues que el modelo esta en un punto medio, no es bueno ni es tan malo realizando las predicciones de vino, se obtuvo un *recall* del 49% que indica que de todas las predicciones hechas solo predijo correctamente para dicho porcentaje de esos datos y finalmente el *f1-score* nos dio un valor de 0.62 para los valores de 5 en la calidad de vino y 0.37 para valores de 6 lo que indica que el modelo es medianamente bueno realizando las predicciones, pero en general podría mejorarse.

Análisis Comparativo y Conclusión

Analizando los resultados obtenidos para los tres modelos se evidenció que se el modelo que obtiene mejores resultados es el de la red neuronal con una capa oculta que tiene el mismo numero de neuronas que de entradas, seguido del modelo perceptrón y por último el modelo de la red

neuronal con dos capas ocultas y dos neuronas cada una. Para el primer modelo, como se contaba con una capa oculta con el mismo numero de neuronas igual al número de entradas permite que los cálculos para detecten características o tendencias en los datos de entrada de manera más eficiente ya que cada neurona se puede encargar de un dato de entrada.

Para el modelo de perceptrón como es conocido al ser un modelo bastante simple en donde solo se cuenta con una capa de entrada y una de salida usando una función escalón de activación, posee ciertas limitaciones frente a la entrada de datos que en este caso era de 11 variables para predecir la calidad del vino. Además, en este caso tuvo buenos resultados debido al asignarse un 80% de datos de entrenamiento frente al 20% de prueba y que el dataset era medianamente pequeño. En caso de que hubiese un mayor flujo de datos posiblemente perceptrón arrojaría peores resultados.

Finalmente, el modelo con dos capas ocultas y con dos neuronas arrojaron los peores resultados, ya que como las entradas eran mayores que el número de neuronas disponibles en esa primera capa oculta los cálculos en las predicciones no serían tan acertados y nos damos cuenta de que para construir el modelo no depende tampoco del número de capas ocultas sino de factores como la función de activación y como están distribuidos los datos.