Máster	Análisis y Visualización de Datos Masivos
Tema	**Técnicas de Inteligencia Artificial - Actividad 2**
Alumna	**Shirley Claudette Martínez Cerrato**

APARTADO DE REGRESIÓN

Descripción de Dataset

Houses: Este dataset está basado en el censo de casas de California en el año de 1990, su objetivo es predecir el valor medio de una casa en dólares, en base a los diferentes parámetros de entrada que serán descritos posteriormente, esta variable representa el atributo de salida y es de tipo numérico. El dataset se compone de 20,640 instancias, 9 atributos, 8 de entrada y 1 atributo de salida, 0 valores nulos, a continuación, se describen los atributos de entrada:

1-median_income: Determina el ingreso promedio de las personas que viven en la casa, tipo numérico. **2-housing_median_age:** La edad promedio o aproximada de la casa, tipo numérico. **3-total_rooms:** Total de secciones que posee la vivienda, tipo numérico. **4-total_bedrooms:** Total de habitaciones para dormir que posee la casa, tipo numérico. **5-population:** Número de personas que viven en el mismo bloque donde se encuentra la casa, tipo numérico. **6-households:** Número de personas que viven en el mismo bloque donde se encuentra la casa, tipo numérico. **7-latitude:** Distancia entre el centroide del bloque, tipo numérico. **8-longitude:** Distancia entre el centroide del bloque, tipo numérico

A continuación, se muestra y explica la distribución de los datos:

Diagrama de Frecuencias Atributos de Dataset Houses

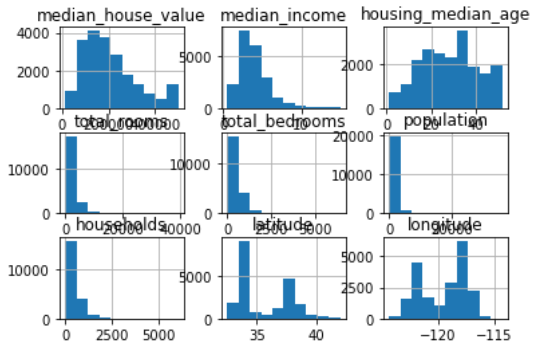
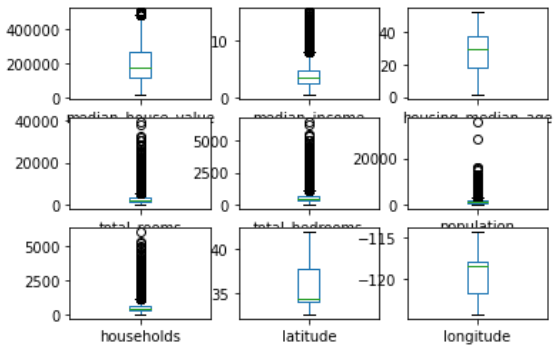


Diagrama de Cajas Atributos de Dataset Houses



Como se puede observar ninguno de los atributos, tanto de entrada como de salida poseen una distribución normal, en el caso del atributo de salida, la distribución del precio de las casas tiene mayor aglomeración en casas de menor valor.

Definición de Los Modelos Predictivos

Para los modelos de predicción se ha utilizado una red neuronal y la regresión lineal. A continuación, se exponen los principales parámetros de la red neuronal:

-Layers: Toman una entrada y devuelven una salida, dependiendo de las diferentes capas que utilicemos, la computación de los datos es diferente. **-Metrics:** Determina en base a que cálculo estadístico se medirá la precisión de la red neuronal. -**Epoch:** Es el procesamiento completo de todos los datos de entrenamiento por el modelo predictivo, entre mayor sea el número de épocas, mayor será la precisión del modelo.

A continuación, se exponen los principales parámetros de la Regresión Lineal:

- normalize: Este parámetro indica si los datos deben ser normalizados o no, en este caso se usó la opción por defecto y se normalizaron los datos aparte. **- copy_X:** Este parámetro indica si al momento de pasar los datos de entrenamiento, estos serán copiados o podrán ser sobrescritos, se usó la opción por defecto. **- positive:** Indica si los resultados arrojados por el modelo serán forzados a ser siempre positivos o no, se utilizo la opción por defecto

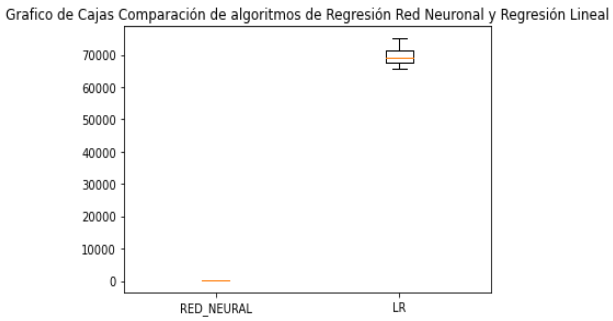
Arquitectura de la Red Neural

La arquitectura de la red neuronal se compone de una capa de entrada densa de 128 unidades, con función de activación “relu”, luego se añadieron dos capas intermedias densas de 64 y 32 unidades respectivamente, ambas con activación “relu”, se finalizó con una capa de salida densa de una unidad, para ofrecer un único valor de salida. Para entrenar el modelo se utilizaron 10 épocas.

Comparación de Resultados

Para la comparación de los modelos se utilizó la técnica de cross-validation y también se realizaron cálculos de forma individual. Obteniendo los siguientes resultados:

Validación Cruzada



Validación Individual

En ambos modelos se utilizó como métrica de evaluación el error cuadrático medio.

Red Neuronal:

```
Train Error: 64907.23157430566
Test Error: 66478.51085704441
```

Regresión Lineal:

```
Train Error: 69262.70269311876
Test Error: 70582.49506314681
```

Conclusiones

Para este dataset, que está conformado enteramente por atributos numéricos, y cuyo objetivo era predecir un valor numérico, la red neuronal presentó un mejor desempeño frente al modelo de regresión lineal. En este caso, la ventaja x10 que ofrece las épocas, sumado con la capacidad de aprender de la red neuronal, permitió que el modelo predictivo redujera considerablemente el error cuadrático medio en cada entrenamiento y, por ende, realizará mejores tareas de predicción que la regresión lineal. Para mejorar las predicciones para ambos modelos se podría aumentar la proporción de datos de entrenamiento, así como una mejor técnica de normalización.

APARTADO DE CLASIFICACIÓN

Descripción de Dataset

Artificial -Characters: Este dataset ha sido generado artificialmente y describe la estructura de las letras capitales A, C, D, E, F, G, H, L, P, R, indicadas por un numero de 1 a 10, en ese orden (A=1,C=2,...). estos números representan las clases a predecir del atributo de salida (Class). Cada estructura de una letra es descrita por un conjunto de segmentos (líneas), que representan los atributos de entrada. El dataset se compone de 10,218 instancias, con 8 atributos, 1 de salida, y siete atributos de entrada, 0 valores nulos. A continuación, se describen los atributos de entrada:

1-V1: El número de segmento, tipo numérico. **2-V2:** Coordenadas iniciales y finales en el plano cartesiano, tipo numérico. **3-V3:** Coordenadas iniciales y finales en el plano cartesiano, tipo numérico. **4-V4:** Coordenadas iniciales y finales en el plano cartesiano, tipo numérico. **5-V5:** Coordenadas iniciales y finales en el plano cartesiano, tipo numérico. **6-V6:** Largo de un segmento computado, usando distancias geométricas entre dos puntos, tipo numérico. **7-V7:** Largo de la diagonal del rectángulo más pequeño que incluye la imagen del carácter, tipo numérico.

A continuación, se muestra y explica la distribución de los datos:

Diagrama de Frecuencias Atributos Dataset Artificial-Characters

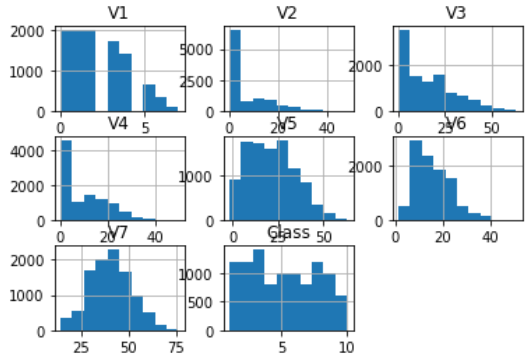
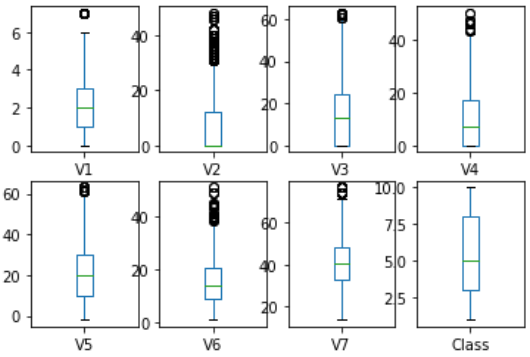


Diagrama de Cajas Atributos de Dataset Artificial-Characters



Para los atributos de entrada, la mayoría de los datos están desbalanceados, pero se puede observar en el atributo de salida, que la distribución es casi uniforme.

Definición de los Modelos Predictivos

Para realizar la clasificación, se ha utilizado una red neuronal y un árbol de clasificación CART. A continuación, se exponen los principales parámetros de la red neuronal:

- activation: Determina como será computado el peso de cada capa, los cuales serán pasados como entradas a las capas siguientes. **-Density Layer Units:** Indica la densidad del vector de salida de la capa. **-loss:** Parámetro que nos permite observar, durante la fase de entrenamiento, la calidad del comportamiento de nuestro modelo con cada optimización, idealmente con cada entrenamiento, la perdida debería ser menor.

A continuación, se exponen los principales parámetros del árbol de decisión:

-splitter: Parámetro que indica la estrategia utilizada para elegir la mejor partición del nodo, se utilizó el valor por defecto. **-max_depth:** Indica la profundidad del árbol, se utilizó el valor por defecto. **-criterion:** Parámetro que indica el criterio para seleccionar la mejor partición, se utilizó el valor por defecto que es “gini”.

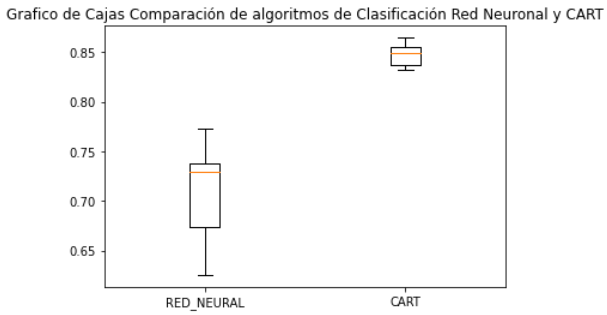
Arquitectura de la Red Neuronal

La arquitectura de la red neuronal se compone de una capa de entrada densa de 100 unidades, con función de activación “relu”, luego se añadieron dos capas intermedias densas de 50 y 25 unidades respectivamente, ambas con activación “relu”, se finalizó con una capa de salida densa de 10 unidades, debido al one-hot-encoding de las 10 clases a predecir, con función de activación “softmax” para convertir el vector de valores de las capas densas a un conjunto de distribución de probabilidades para cada clase. Para entrenar el modelo se utilizaron 10 épocas.

Comparación de Resultados

Para la comparación de los modelos se utilizó la técnica de cross-validation y también se realizaron cálculos de forma individual. Obteniendo los siguientes resultados:

Validación Cruzada



Validación Individual

En ambos modelos se utilizó como métrica de evaluación la precisión.

Red Neuronal:

```
Accuracy Score: 0.8065883887801696
```

Árbol de Decisión:

```
Accuracy Score: 0.8584474885844748
```

Conclusiones

A pesar de que, el modelo de red neuronal demostró una mejora progresiva, aparte de tener la ventaja X10 (número de épocas) en cada entrenamiento sobre el árbol de decisión, al visualizar las métricas finales, podemos ver que el árbol de decisión hace un mejor trabajo a la hora de realizar tareas de clasificación. Es seguro que la red neuronal mejorará con cada entrenamiento por su capacidad de aprender, pudiendo llegar a tener una precisión igual o mayor al árbol de decisión, pero si el entrenamiento siempre se realiza con los mismos datos, puede causar overfitting. Para mejorar las métricas de la red neuronal se puede aumentar la densidad de las capas o añadir más capas intermedias.