Roberto Arnoso y Adrià Navarrete

Entrega 5.1 - Preparación Dataset

1. Definición del problema

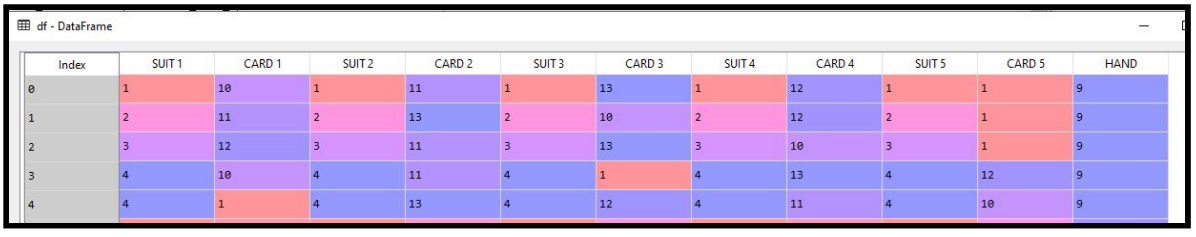
Nos encontramos con un dataset enorme con muchas muestras que nos iban a complicar a la hora de elegir el clasificador y entrenarlo. Exactamente el dataset que nos descargamos tiene un millón de muestras por cada categoría (lo que significa que tanto cada carta como cada palo tienen un millón de muestras. Además de su correspondiente jugada). En total el dataset estaba formado por 13 millones de muestras. Por ello nos hemos visto obligados a adaptar un dataset de entrenamiento más reducido, que permita a los clasificadores absorber el volumen de datos de entrenamiento y que nos permita realizar una predicción correcta si alterar los resultados de la misma.

2. Obtenció i preparació de les dades

Finalmente tenemos un dataset reducido a 25010 muestras de cada categoría además de su correspondiente jugada (pareja, doble pareja, etc...). En resumen un dataset de 325130 muestras, un 2,5% del total de las muestras que teníamos originalmente. Esta reducción se ha realizado de manera que los datos (cartas y jugadas) queden lo más balanceados posibles. Es decir que no se altere la probabilidad de obtener una mano u una carta, al haber realizado esta reducción

Junto a la entrega del documento adjuntamos los archivos y el script para la preparación del dataset, el cual vamos a comentar a continuación:

Primero de todo tenemos que recibir los datos de un .data a través de la librería pandas, este lo tenemos que leer como una tabla y añadir los separadores en las comas. Por lo que nos queda un dataframe con valores, el siguiente paso es añadirle las columnas correspondientes. Hemos decidido categorizar el dataframe de la siguiente manera:



df - DataFrame

Index	SUIT 1	CARD 1	SUIT 2	CARD 2	SUIT 3	CARD 3	SUIT 4	CARD 4	SUIT 5	CARD 5	HAND
0	1	10	1	11	1	13	1	12	1	1	9
1	2	11	2	13	2	10	2	12	2	1	9
2	3	12	3	11	3	13	3	10	3	1	9
3	4	10	4	11	4	1	4	13	4	12	9
4	4	1	4	13	4	12	4	11	4	10	9

En el poker nos encontramos con 4 suites, y de cada uno de estos nos encontramos 13 cartas (As,1,2,3,4,5,6,7,8,9,J,Q,K). Por lo que a la hora de indicar la carta que tenemos en nuestro mazo debemos indicar las siguientes características:

1. Suit (Palo) atribuido a la carta
2. Número (Valor) de la carta

En nuestra mano nos encontraremos con 5 cartas, cada una con su suite y número correspondiente, y ésta formará una jugada:

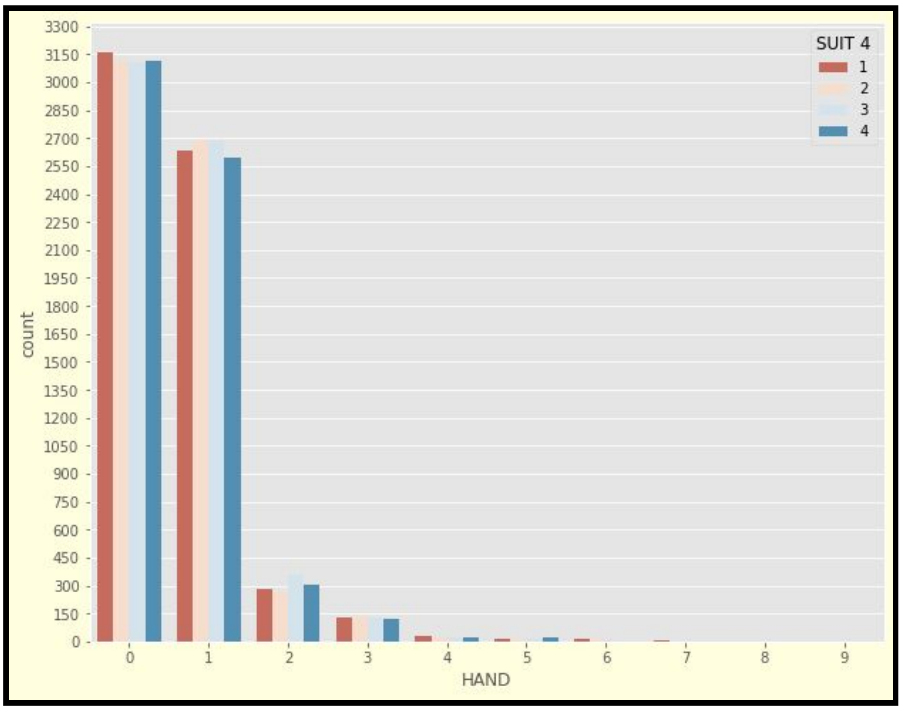
1. Nada
2. Una pareja de números
3. Dos parejas de números
4. Un trío de números
5. Escalera de números
6. Color, 5 cartas de la misma suite
7. Full, un trío de números + una pareja de números
8. Poker, 4 cartas con el mismo número
9. Escalera de color
10. Escalera real

Por lo que en nuestro dataset deberemos de indicar por cada una de las muestras las 5 cartas que tenemos (El número y la suite) más una última columna que nos indique el tipo de jugada que le correspondía. Un ejemplo de una muestra en nuestro data frame sería:

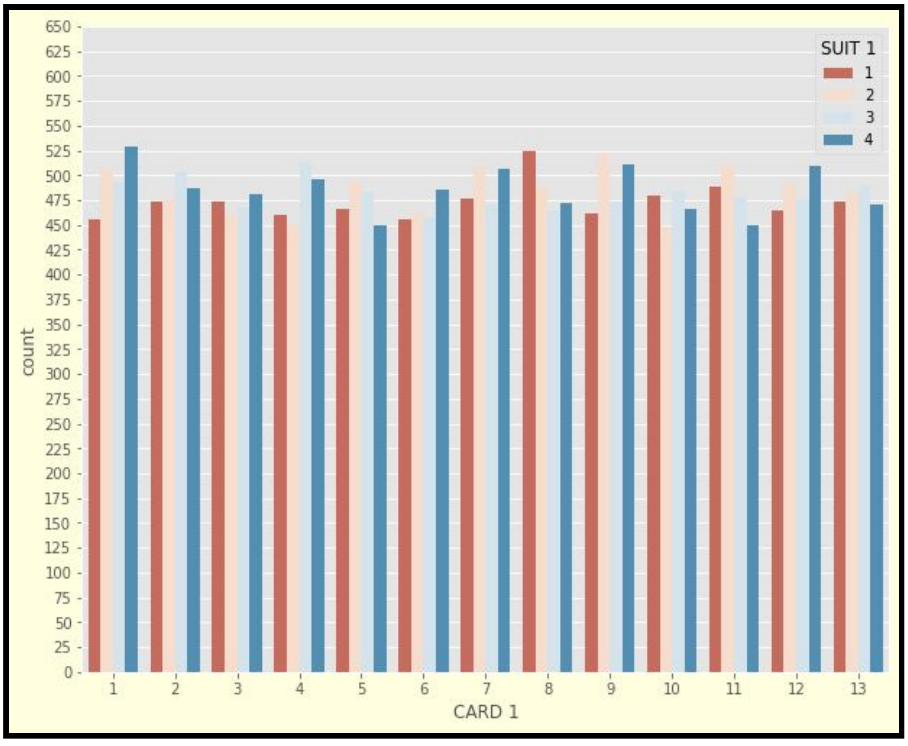
Index	Suite Carta 1	Número Carta 1	Suite Carta 2	Número Carta 2	Suite Carta 3	Número Carta 3	Suite Carta 4	Número Carta 4	Suite Carta 5	Número Carta 5	Jugada
0	1..4	1..13 (1-> As, 13 -> Rey)	1..4	1..13 (1-> As, 13 -> Rey)	1..4	1..13 (1-> As, 13 -> Rey)	1..4	1..13 (1-> As, 13 -> Rey)	1..4	1..13 (1-> As, 13 -> Rey)	0..9 (0 -> Nada, 9 -> Escalera real)

Una vez definido el diseño, analizamos todas la muestras que tenemos para el entrenamiento del clasificador.

3. Análisis Gráfico



Este gráfico (#Plot1) representa el número de jugadas (HANDS) obtenidas por palo. Como es lógico las jugadas más probables son las asociadas a los valores 0 y 1 (Mano vacía y pareja). Estas manos son las más probables de obtener como se aprecia en el gráfico. Esta suposición se adecua a la que podríamos formular en la realidad.



Este gráfico (#Plot2) representa la distribución equiprobable de cada una de las cartas dentro de los distintos palos de la baraja. Con este gráfico mostramos que la posibilidad de obtener una carta u otro es equiprobable e independiente del palo escogido. Lo que hace totalmente independientes ambas variables (color del palo, valor de la carta).