Proyecto Capstone

Informe final

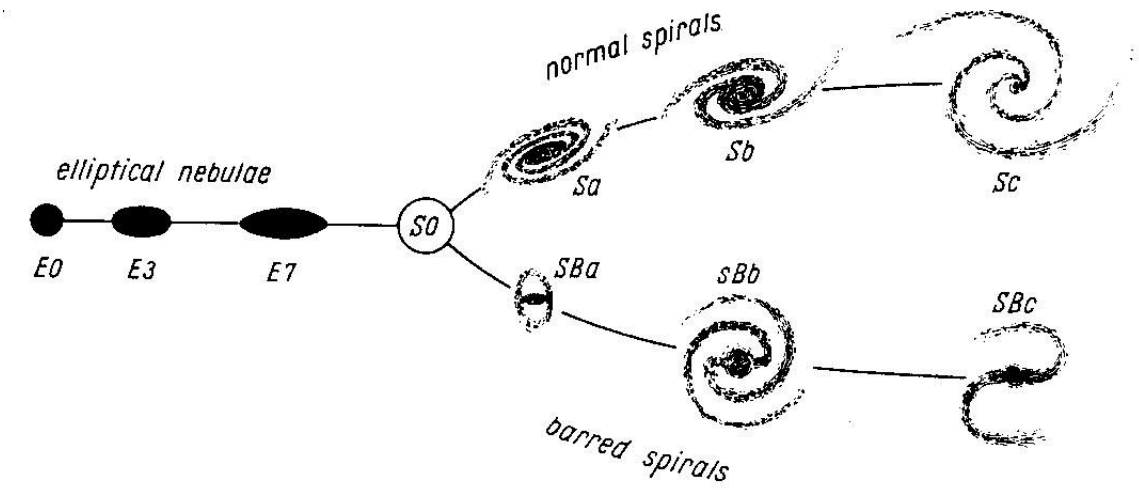
1. Objetivo del proyecto

El objetivo de este proyecto capstone es aplicar técnicas y herramientas para el big data.

2. Criterio de clasificación de las galaxias

2.1 Las galaxias se pueden clasificar por su forma en espiral o elíptica, pero existen muchos estados intermedios con formas irregulares, tamaño y brillo que dependen de su edad y posición, identificar cada una de las imágenes con un tipo de galaxia. En este caso la clasificación es binaria, o bien es elíptica o bien es espiral, y siempre tiene que haber una respuesta. No existe el concepto “no lo sé”.

Para distinguirlas se usan los siguientes criterios, En 1926 Edwin Hubble inventó un esquema de clasificación morfológica para las galaxias, la secuencia de Hubble, o como se conoce coloquialmente, diagrama diapasón, a consecuencia de la forma de su representación gráfica:



Por lo que podemos decir también que, brillo es una característica importante ya que cuando menor es la magnitud mas brillante un objeto y se le asignan unas letras que identifican los filtros y cada una corresponde a una longitud diferente por donde pasa la luz.

3. Descripción de los datos y las herramientas usadas

3.1 Para este proyecto tenemos a disposición los siguientes ficheros de parámetros:

Fichero	**Formato**	**Descripción**
SDSS_parametros.zip	ZIP	contiene los ficheros SDSS_PhotObj.csv y SDSS_SpecObj.csv, con parámetros de objetos celestes
SDSS_imagenes.zip	ZIP	contiene ficheros de imágenes de galaxias en formato JPG
Usuarios	csv	Usuarios que clasificaron las imágenes
Votos	Csv	Voto que realiza el usuario
T_F_DR14_ZooSpec_10000	csv	Features y target

3.2 Las herramientas informáticas utilizadas son:

Herramienta	**Descripción**	**Usada para**
Cloudera MV	Máquina virtual	Espacio de trabajo para correr aplicaciones para trabajar big data
HDFS	Es un sistema distribuido basado en Java se usa por comandos	Esta pieza hace posible almacenar data sets masivos con tipos de datos estructurados, semiestructurados y no estructurados como imágenes, vídeo, datos de sensores
HIVE y Beeline	gestionar enormes datasets almacenados bajo el HDFS, se usa por comandos	Crear datos, importar datos
SPARK	Analisis, visualizacion e interpretacion de resultados	su potencia de procesamiento agiliza la detección de patrones en los datos, la clasificación organizada de la información

Jupyter nootebooks	Interfaz usada para spark y python	Lee y ejecuta código de maquina

4. Exploración de los datos

La exploración de los datos con los parámetros de los objetos observados ha dato el siguiente resultado:

Fichero	**n. de elementos/ tamaño**	**Observaciones**
SDSS_PhotObj	25100/2.8 mb	No existen elementos repetidos, se encuentran enumeradas por el “objid” para identificarla, todos los datos son validos
SDSS_SpecObj	25100/1.2 mb	Existen algunos elementos repetidos, cuenta con una columna “objid” usada para identificar el objeto, contiene datos validos.
Voto	379149/16.1 mb	Cuenta con 4 columnas, id galaxya, id usuario, forma y tiempo, existen datos duplicados pero necesarios para exploracion.
usuario	14919/248 kb	Cuenta con 3 columnas, id_usuario, edad, país. Aun asi contiene usuarios para filtrarse.
T_F_DR14_ZooSpec_10000	10000/605 mb	Estos datos son validos aun asi necesitan filtrarse.

5. Modelización de los datos de voto

La normalización del modelo de datos se hace para el correcto manejo de los datos y posterior análisis que facilite su comprensión.

5.2 Modelo normalizado

Nombre de la tabla: USUARIOS

Descripción: contiene los datos de los usuarios que participaron en la clasificación.

Nombre	**Tipo**	**Descripción**
id_usuario	entero	Identificador del usuario
edad	entero	Edad del usuario

País	carácter	País del usuario

Nombre de la tabla: VOTOS

Descripción: contiene los datos de la votación durante la clasificación.

Nombre	Tipo	Descripción
id_galaxia	entero	identificador de la galaxia
id_usuario	entero	Identificador del usuario
forma	entero	Forma elegida (0=elíptica, 1=espiral)
tiempo	flotante	Tiempo de respuesta en segundos

Nombre de la tabla: sdss_photobj_csv

Descripcion

Nombre	Tipo	Descripción
objid	entero	Identificador del objeto a clasificar, llave primaria
fileid	entero	identificador usado para el fichero que contiene la imagen correspondiente del objeto
ra, dec	flotante	Coordenadas celestes del objeto
u, g, r, i, z	flotante	Brillo (magnitud) medido en los diferentes filtros de colores por cada objeto.
field	entero	Número que identifica el área de cielo en la que se encuentra el objeto

Nombre de la tabla: sdss_specobj_csv

Descripcion:

Nombre	Tipo	Descripción
objid	entero	Identificador único de la galaxia, llave primaria
redshift	flotante	Indicador de la distancia del objeto
plate, fiberid	entero	Identificadores de los instrumentos del telescopio que se ha utilizado para hacer la medida estimada de redshift
mjd	entero	identificador de la fecha en la que se ha tomado la medida

class	carácter	nombre de la clase del objeto, según las características de la luz emitida en el espectro electromagnético observado

6. Exploración de los datos de voto

6.1 De la exploración de los datos de voto podemos comprobar que existen columnas que debemos normalizar tambien hay datos que necesitan ser limpiados, ya que en algunos votos el usuario se tomó muy poco tiempo para la clasificación y algunos usuarios son menores de 10 y otros mayores de 100 años, en total tenemos 379149 votos y 14919 usuarios

6.2 Gráficos

Grafico de usuarios por edad

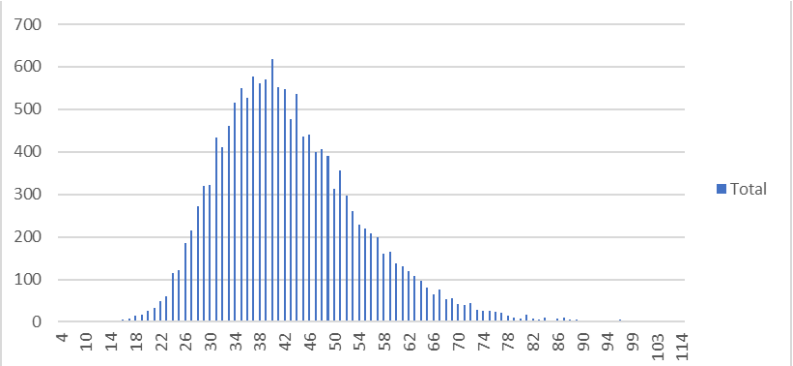


Gráfico de usuarios por país

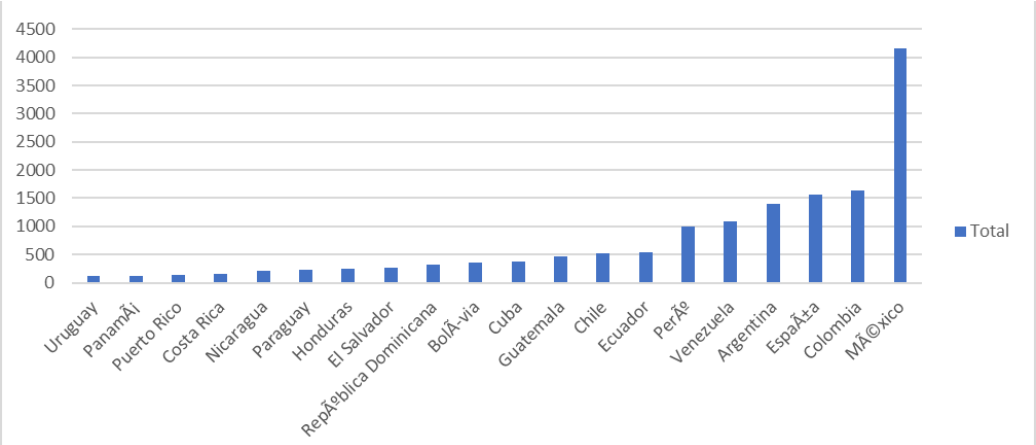
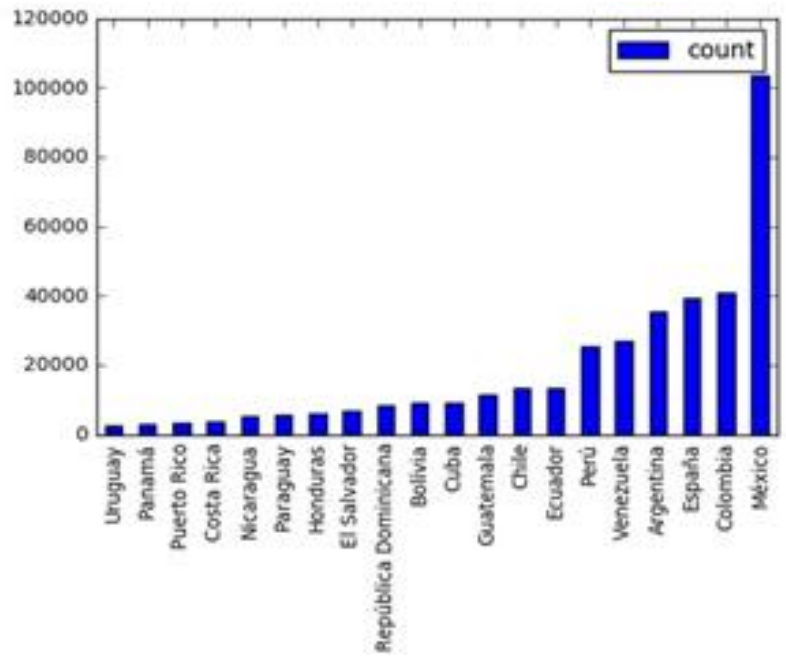


Gráfico votos por país



7.Creación del clasificador

7.1 Los datos usados para crear el clasificador son data set de imágenes galácticas (features); para cada una y cada columna contiene la información de un pixel, estandarizado, la clasificación (target) se obtienen de las clasificaciones que generamos a partir de los votos con los siguientes valores 0 incierto, 1 elíptica, 2 espiral, agrega una columna para identificar a cada objeto.

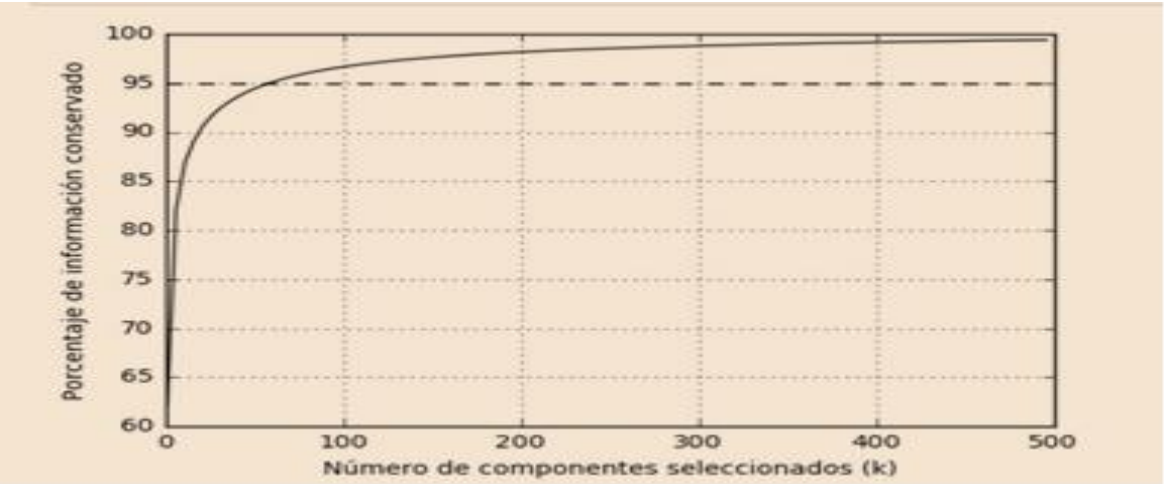
Se han reunidos en una misma tabla con las siguientes características

Nombre	Tipo	Descripción
dr7objid	Entero	identificador del objeto
target	entero	clase del objeto según la definición anterior
F0 a F4095	entero	correspondientes a la tupla de 4096 atributos anteriormente descrita

PCA

7.2 Antes de empezar, aplicamos el método PCA para reducir el número de atributos y conservar la mayor cantidad de información posible desde el análisis de los 64 componentes conservados se ve que 95,3% de la información se conserva

.7.3 Análisis de los componentes conservados



Regresión Logística

7.4 Para entrenar un algoritmo de clasificación hay que separar el conjunto de datos en Dividiremos nuestro set de datos en dos subsets más pequeños; df_train y df_test.

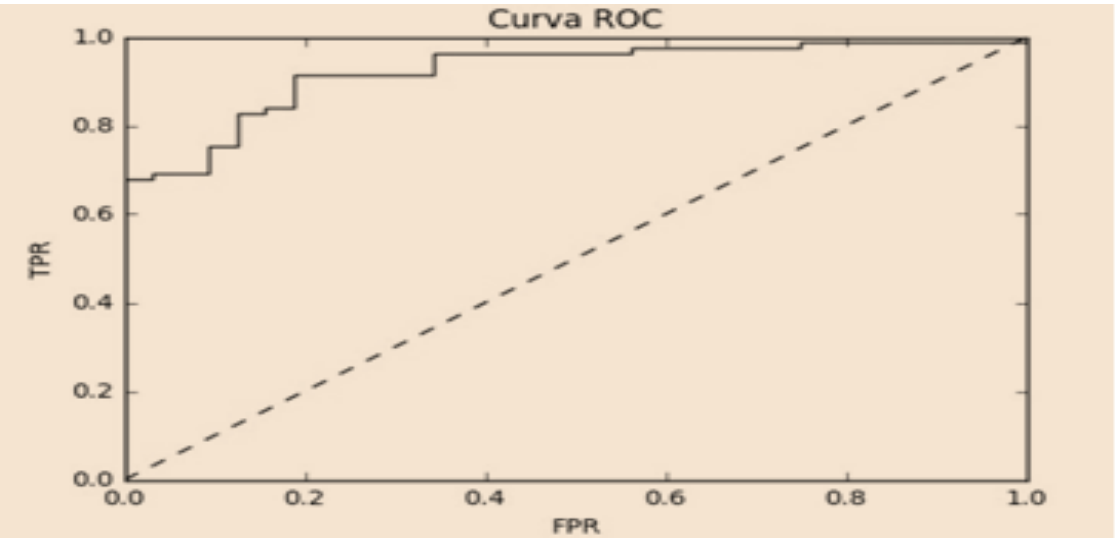
Cada uno con: df_train: para el entrenamiento del modelo (Train) con un 70% de los datos. df_test: para la validación (Test) con un 30% de los datos

7.5 El algoritmo de clasificación se ha preparado en el siguiente modo:

Entrenamiento: donde corremos el 70% de los datos.

Validación: donde corremos el 30% de los datos, aplicación el modelo de predicción y se compara con la clasificación real.

7.6 Se obtiene un porcentaje de acierto de destacar que el porcentaje de aciertos (Accuracy) del modelo es del 83%.



8. Redes neuronales

8.1 Las redes neuronales son unos de los algoritmos más utilizados para clasificación de imágenes y por este motivo lo hemos aplicado es conveniente que, aún y habiendo obtenido un porcentaje de aciertos razonablemente bueno mediante un modelo de regresión logística, valoremos la utilización de otros métodos. con el objetivo de obtener una fiabilidad más alta.

8.2 Las características de la red neuronal que vamos a utilizar son:

Capas: [64, 16, 8, 2] se utilizarán 4 capas con sus respectivos nodos 64, 16, 8y 2.

Resultado: Podemos observar un incremento el porcentaje de aciertos se ha incrementado con respecto al algoritmo de regresión logística. Ahora tenemos 92,41% de aciertos

9. Conclusiones

Las Redes neuronales es un tipo de clasificador bastante bueno para nuestro conjunto de datos ya que esta diseñado óptimamente para ello la cual nos da un número mayor de aciertos.

Si bien el modelo de regresión logística da unos resultados excelentes se hace necesario probar algunos de los otros modelos conocidos para retroalimentar la información brindada por los datos.