Clasificación de Datos estelares

Modelos para clasificar datos de tipo estelar

Beatriz González Merino 12-6-2025

Indice

- Objetivo del Modelo
- Análisis de datos
- Modelos
  - Logistic Regression
  - Support Vector Machines
  - Random Forest
- Conclusiones

Objetivo del model

- El objetivo del modelo es clasificar un conjunto de datos con objetos estelares en estrellas, galaxias o quasars (qso).
- Para ello vamos a entrenar varias modelos de clasificación y nos quedaremos con el que nos proporcione una mejor tasa de acierto.

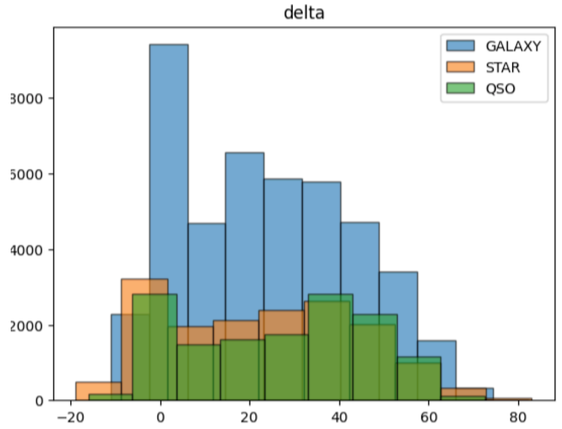
Análisis de datos

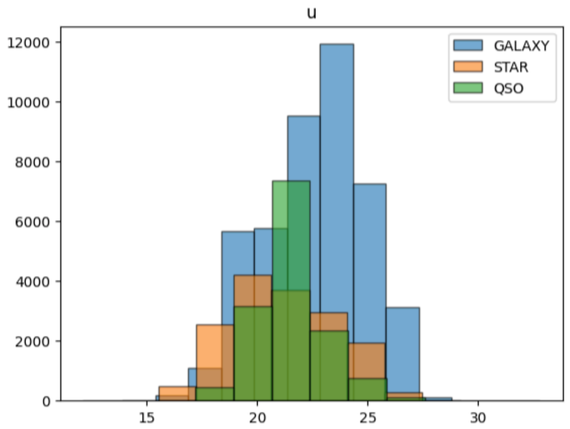
Limpieza y entendimiento de los datos

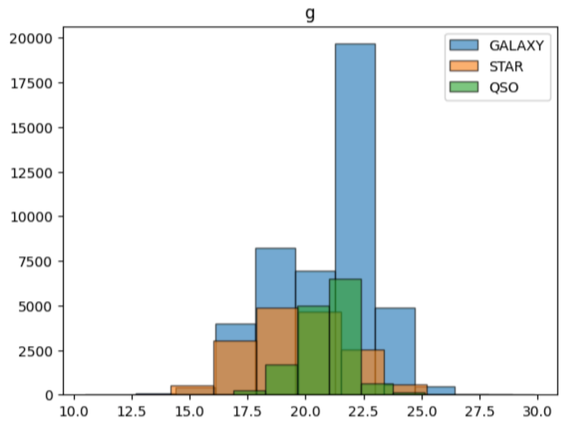
- Consideramos sólo las columnas con información sobre los objetos, no las identificativas
- Sustituimos los valores nulos, -9999, presentes en los datos por la media de la columna
- Comprobamos si hay outliers que necesiten ser eliminados y cuál es la distribución de las variables para cada una de las clases representándolas en forma de histogramas

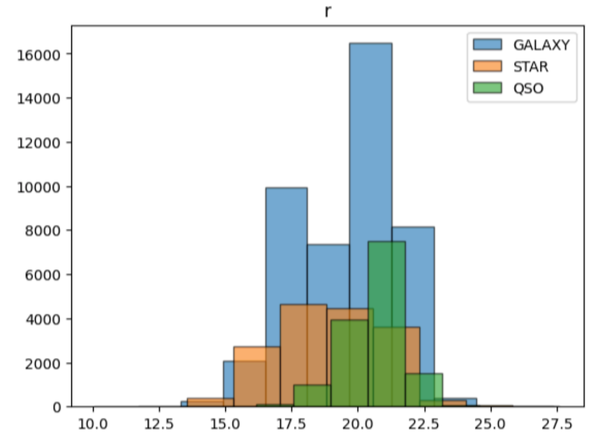
Análisis de datos

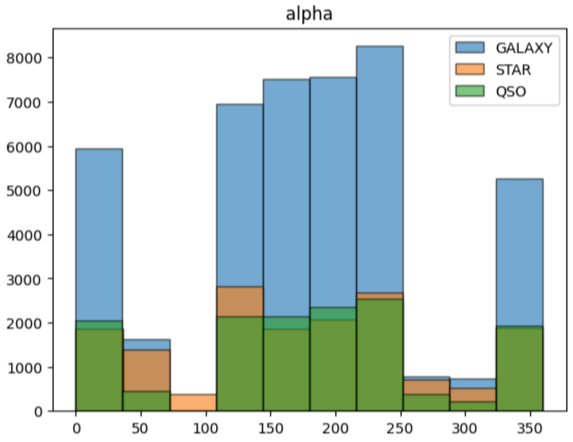
Representación de las variables por clases

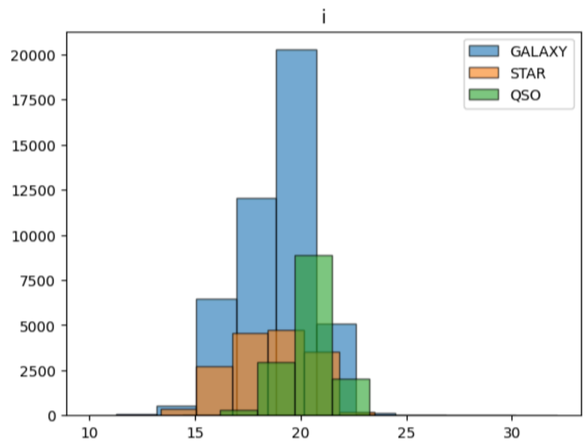


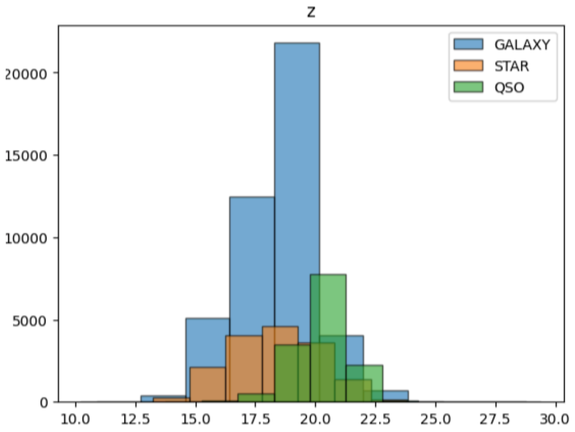


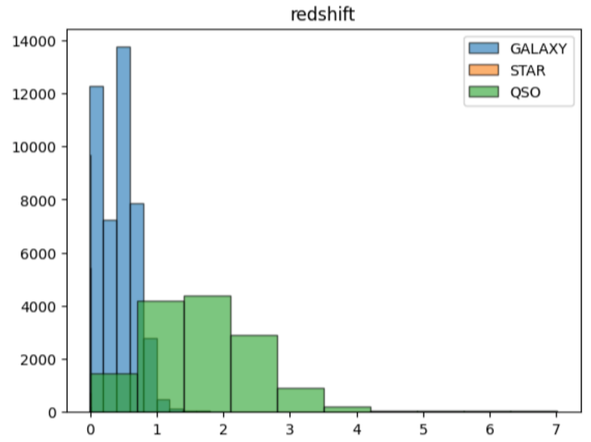












Análisis de datos

Representación de las variables

- Las galaxias son el objeto más común en el set de datos
- Todos los objetos observados se encuentran en el hemisferio norte (delta>0)
- Las emisiones en los diferentes filtros de observación dependen de la energía del objeto, y por eso presentan distribuciones diferentes para cada clase
- La variable donde mayor diferencia se aprecia es en el corrimiento al rojo, redshift. Las estrellas están más cerca a nosotros y por eso su redshift es menor, los quasars son los elementos más lejanos y su redshift es mucho mayor. Las galaxias cubren el rango intermedio

Análisis de datos

Distribución del número de observaciones por clase

- Como ya hemos mencionado, las galaxias son el elemento más presente en los datos seleccionados
- Esto indica que los datos no están balanceados. Para que los modelos puedan aprender las características de las diferentes clases por igual es necesario balancearlo

class	obj_ID
GALAXY	44621
QSO	14206
STAR	16173

Modelos

- Vamos a ajustar tres modelos diferentes a los datos para ver cuál de ellos funciona mejor:
  - Logistic Regression: auc: 0.986 y accuracy=0.933
  - Support Vector Machine (se han considerado dos posibilidades: ovo y ovr, pero los resultados son muy parecidos): auc=0.990 y accuracy=0.953
  - Random Forest (se ha elegido la mejor opción usando hyper parameter tunning con un grid search): auc=0.995 y accuracy=0.970
- Random Forest es el modelo que mejores resultados ha dado, además va a permitirnos entender un poco más las elecciones del modelo

Random Forest

Explicabilidad

- La importancia de las variables que ha usado el random forest es la siguiente, donde se ve que la variable que más ha contribuido es el redshift y la que menos delta y alpha:

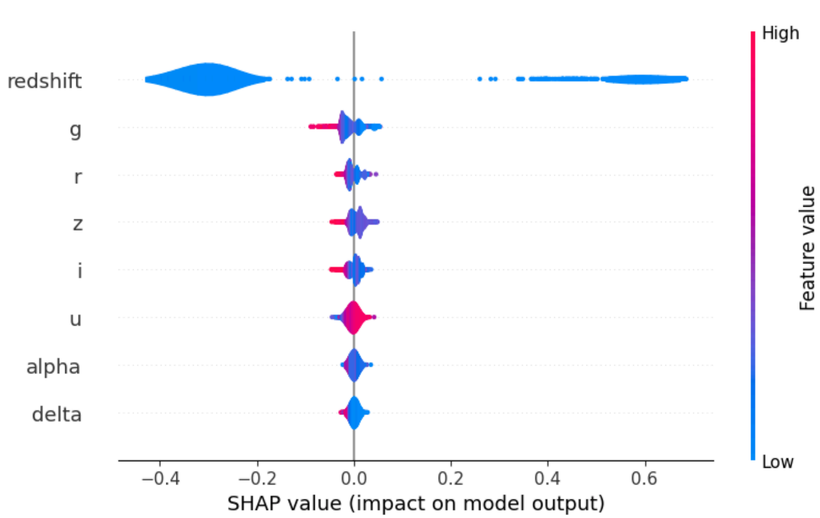
redshift	0.628066
z	0.100012
u	0.074340
g	0.073461
i	0.060123
r	0.050667
delta	0.006830
alpha	0.006501

- Esto indica que la posiciona en el cielo de los objetos no es muy importante para su clasificación, sin embargo el redshift sí que juega un papel relevante.
- Para entender un poco más cuál es el efecto de las variables en la clase que propone el modelo para cada uno de los objetos vamos a obtener los valores Shap

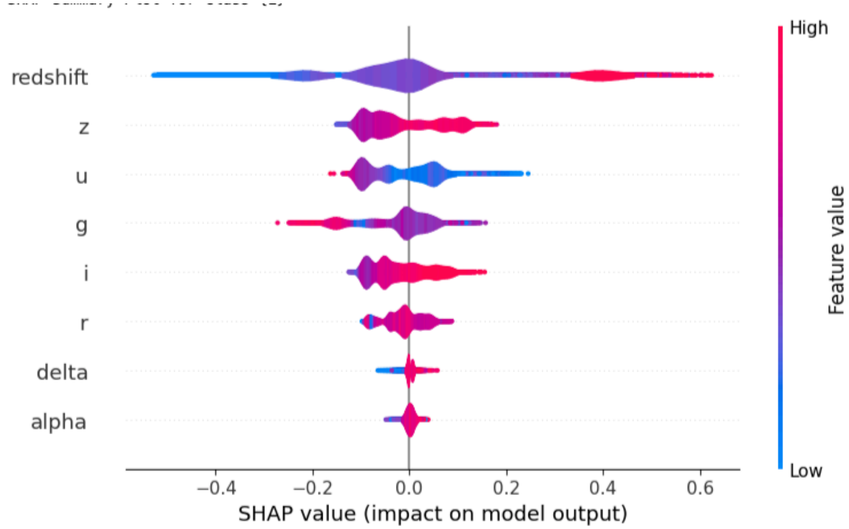
Random Forest

Valores shap globales

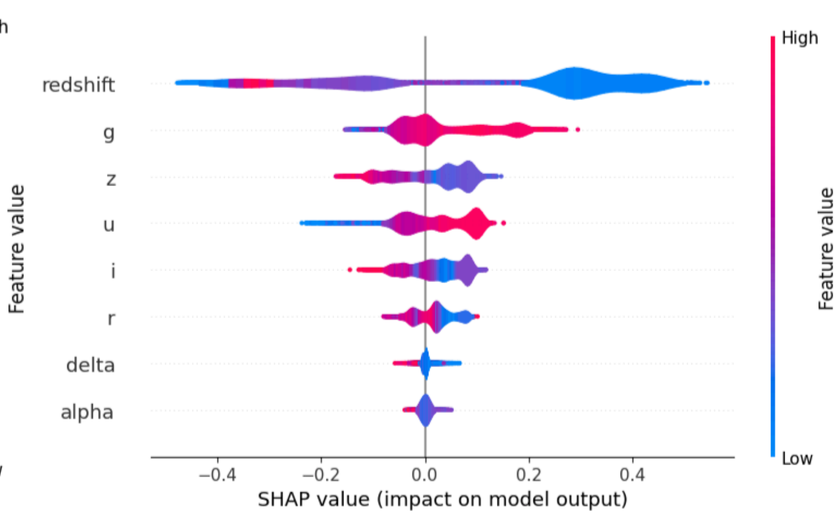
Clase estrellas



Clase galaxies



Clase QSO



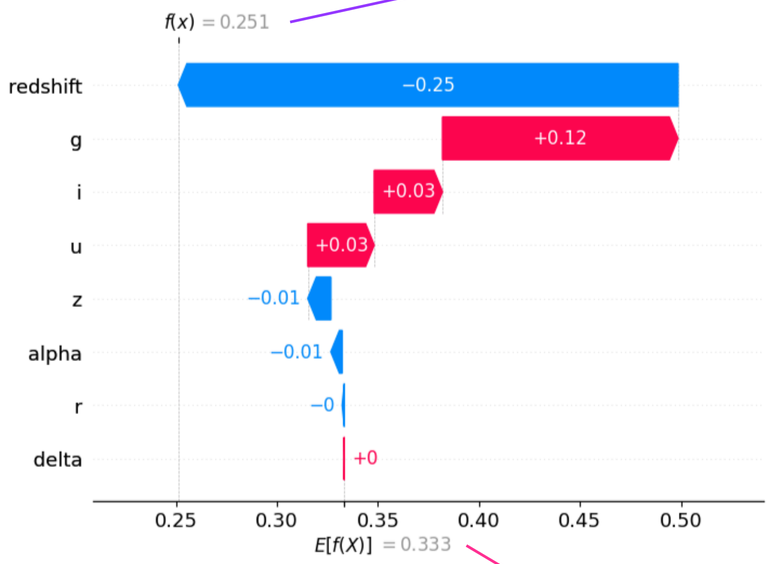
- Podemos ver como la variable redshift es la que más modifica el valor de la predicción

Random Forest

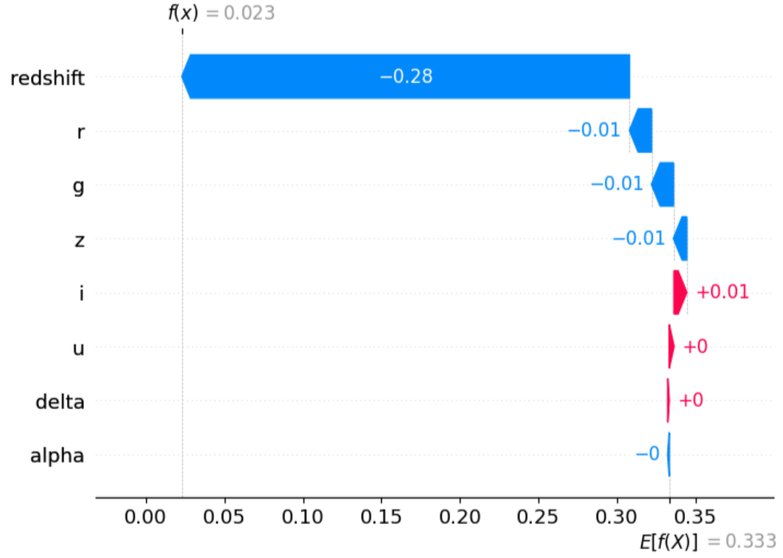
Valores shap individuales

Objeto QSO

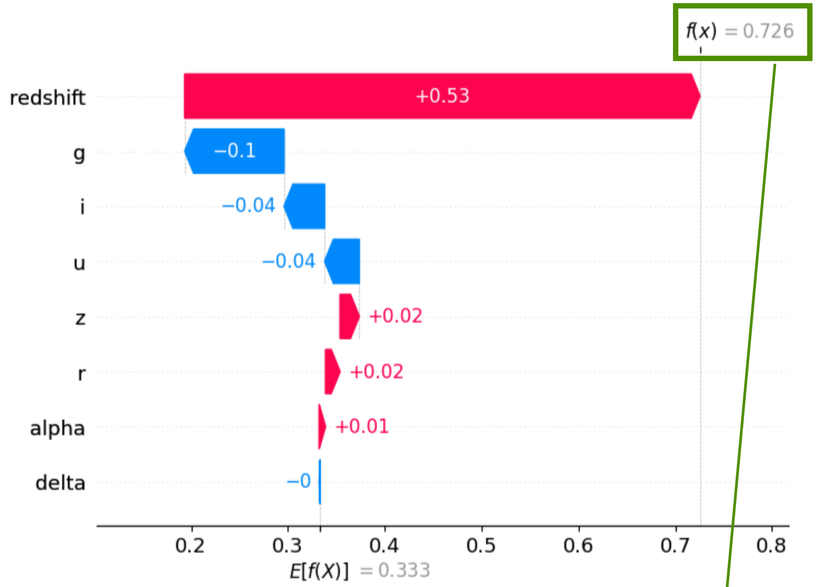
Probabilidad para esta clase



Clase estrellas



Clase galaxias



Clase QSO

Probabilidad media

Clase con mayor probabilidad, y por lo tanto la elegida por el modelo

Random Forest

Error por clase

- Vamos a echar un vistazo a la clasificación que se ha hecho por clase mirando la matriz de confusión

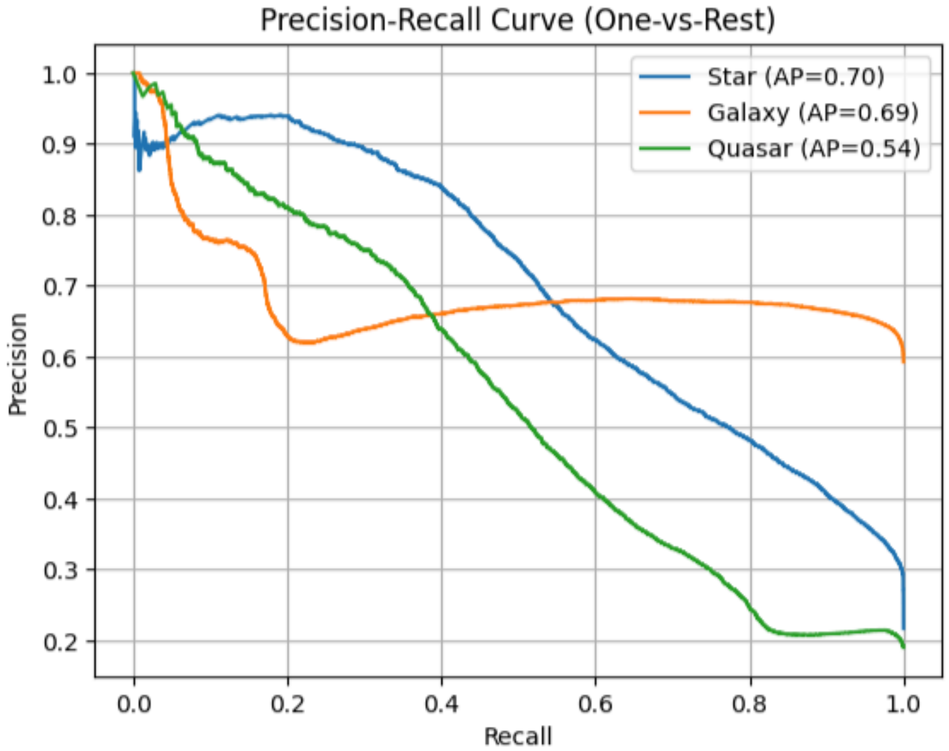
```
[[ 5421     0     0]
 [  134 14334   356]
 [    1   248  4506]]
```

- El modelo está cometiendo pocos errores en la clasificación de las estrellas, mientras que tiene más problemas con los quasars:
  - Tasa de error para estrellas: 0.000
  - Tasa de error para galaxias: 0.0331
  - Tasa de error para los quasars: 0.0524

Random Forest

Precisión-Recall Curve

- La precisión media (AP) muestra como las estrellas son efectivamente el objeto más fácil de clasificar, mientras que los quasars son el más difícil.
- La curva de los quasars muestra además como la precisión cae rápidamente según aumenta el recall, mostrando la existencia de falsos positivos



Random Forest

Modificar los umbrales de predicción

- El algoritmo utilizado tiene un valor por defecto para el umbral de clasificación. Como ese umbral está cometiendo errores en la clasificación de los quasares, vamos a intentar buscar si hay alguna opción que nos proporcione mejores resultados.
- Para ello vamos a utilizar el valor de f1, lo calcularemos para diferentes umbrales y nos quedaremos con los aquellos que maximicen f1.
  - Clase estrellas: umbral=0.828 f1=0.996
  - Galaxias: umbral =0.253 f1=0.978
  - QSO: umbral= 0.591 f1=0.94

Random Forest

Modificar los umbrales de predicción

- Con estos nuevos umbrales obtenemos los siguientes resultados
- acc=0.97488
  - Tasa de error para estrellas: 0.000
  - Tasa de error para galaxias: 0.0216
  - Tasa de error para los quasars: 0.0648

```
[[ 5421     0     0]
 [   71 14504   249]
 [    1   307  4447]]
```

- Estos cambios en los umbrales han llevado a disminuir el error que se cometía en la clasificación de las galaxias, sin embargo ha aumentado el que teníamos en los quasars.
- Elegir un caso u otro dependería de los objetivos finales de la investigación, pero al ser los quasars unos objetos más raros y de los cuales hay menos muestra en la base de datos, priorizaremos su correcta detección y nos quedaremos con los umbrales dados por defecto.

Conclusiones

- El model de random forest es el que mejores resultados ha proporcionado
- La variable que mayor impacta en la clasificación de los objetos es el redshift
- Los umbrales del modelo se pueden modificar para ajustar la forma de decidir que tiene el modelo y favorecer los objetos que presentan mayor interés
- Se desarrolla un script en python donde el usuario podrá elegir entre los dos modelos que mejor resultados han proporcionado (Random Forest y SVM) que permitirá guardar el modelo o ejecutar uno ya guardado, predecir sobre datos nuevos o evaluar el modelo sobre datos que se encuentran labeleados.