Marc Bardeli Gamez
1637718

Predicció de la forma d'un OVNI

Abstract

This work delves into sightings of UFOs reported by people worldwide over a very extensive period. Various models such as Random Forest, SVM, etc., have been employed to seek the desired results and predict the shape a UFO would take based on other reported data. Unfortunately, the results have not been satisfactory on this occasion. Despite observing various shapes in the dataset, the highest accuracy achieved was 20%, with very poor precision-recall and ROC curves. It's also worth noting that further exploration of the dataset yielded results of 50%, 60%, and even 80% accuracy by reducing possible values through category grouping or elimination, but these values are not deemed acceptable. In summary, with the available data in this dataset, making an accurate prediction of the shape of the next sighted UFO without excessive generalization is not feasible.

Paraules Clau: sightings, UFO, accuracy, category grouping, elimination, generalization .

Fig. 1. Imatge "albirament".

1.Introducció

L'albirament d'OVNIs és un tema amb opinions molt dividides, des de gent que no creu que existeixin els alienígenes fins als que afirmen que n'ha vist. En aquest treball es deixa de banda les opinions i s'intenta buscar si és possible predir la forma que tindrà un OVNI en funció d'unes dades d'albiraments força completes d'un període d'anys bastant extens sent 1949 el primer any del primer albirament del dataset. Tot i no centrar-nos en opinions, el dataset segueix sent subjectiu i cada albirament i per tant cada dada del dataset depenia de l'opinió i del que creia haver vist cada persona, per aquesta raó podem observar fins a 29 formes diferents possibles, d'aquestes 29, algunes ja es podien ajuntar amb d'altres directament ja que es podien interpretar de la mateixa forma o es podien eliminar ja que només tenien 1 o molt poques entrades en el dataset.

Les descripcions dels albirament no s'han tingut en compte ja que en moltes es deia directament la forma del OVNI i inutilitzava la finalitat d'aquest treball.

Light	11454
Circle	8997
Triangle	5648
Fireball	4348
Other	3923
Unknown	3879
Disk	3630
Oval	3095
Cigar	2321
Formation	1748
Changing	1331
Flash	908
Rectangle	889
Diamond	817
Chevron	672
Teardrop	544
Cross	152

Fig. 2. Diferents formes d'OVNI

2.Metodologia

Primer de tot s'ha visualitzat bé quines i quin tipus de dades té el dataset, en aquest pas ja s'ha vist diferents coses a tenir en compte com; que hi ha atributs redundants, és a dir, atributs que la seva informació ja està en altres atributs, per exemple, Date_time ens diu la data, any, mes, dia i hora exacta, tot això ja esta separat en altres atributs així que podem directament eliminar Date_time i Encounter_Duration que ens diu el temps en hores, minuts o segons i es troba en la mateixa situació. També s'ha descartat la data en la que es documenta l'albirament en la database, no és una dada pertinent per l'anàlisi a fer.
I per últim, així a priori veiem que tenim atribut pels països i un altre pel codi d'aquest, per tant un d'aquests atributs s'ignora.

Respecte al balanceig de dades, on se n'ha observat més és en l'atribut del país que majoritariament té com a valor Estats Units, per tant s'ha decidit o tenir-lo en compte, la resta de dades estan prou ben balancejades per no suposar cap problema

També s'ha comprovat si ens faltaven dades, i s'ha descobert que si, donat que en cap cas el percentatge era major a 3% s'ha optat per elimina senzillament les files que no tenen tots els valors.

I com a últim pas abans de començar a aplicar models les dades categòriques s'han passat a numèriques i s'ha triat una mètrica, accuracy, amb les dades que i l'objectiu que tenim per aquest dataset sembla la millor per analitzar els models.

Una vegada ja està tot preparat s'ha començat a aplicar diferents models, aquesta vegada són 4: DecisionTreeClassifier, linear_model, RandomForestClassifier i svc. Per buscar els millors resultats s'ha tingut en compte apart del propis resultats de la mètrica triada, el temps d'execució de cada mètode ja que han variat bastant.

3.Resultats i anàlisi

Els resultats no han estat bons en general amb cap mètode ni amb cap mètrica, si bé no és agradable veure com un treball com aquest acaba amb uns resultats així, no significa que el temps invertit hagi estat perdut.

En primer lloc cal comentar la primera classificació feta, aquesta era simplement intentar classificar només amb el preprocessing de les dades fet. Això ha donat uns resultats nefasts, un accuracy menor a un 20%. Aquí ja es pot començar a observar que el dataset sembla no tenir suficient informació per triar correctament la predicció.

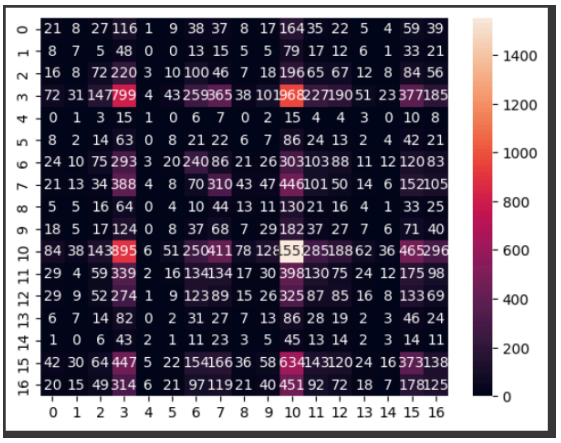


Fig. 3. Matriu de confusió inicial

Aquesta matriu de confusió correspon al RandomForestClassifier, donat que hi ha 16 classes diferents és difícil veure que passa, això ja és una primera cosa a tenir en compte, tot i així si es mira amb calma es pot observar que la classe 10, 3 i 15 son les que tenen més prediccions tant equivocades com correctes, aquestes 3 classes coincideixen amb les 3 classes amb més entrades al dataset.

A partir d'aquí, apart d'anar provant diferents atributs com a valors de X_train i X_test(que no van tenir gairebé cap efecte), es van començar a provar canvis més dràstics:

Fusionar el màxim nombre possible de tipus de forma en una sola.(Deixant més de dos formes diferents)

Així en queden unes 3, 4 formes depenent de quines es vulguin fusionar, tot i així l'accuracy només ha arribat fins al 37%.
Per tant, no ha estat suficient.

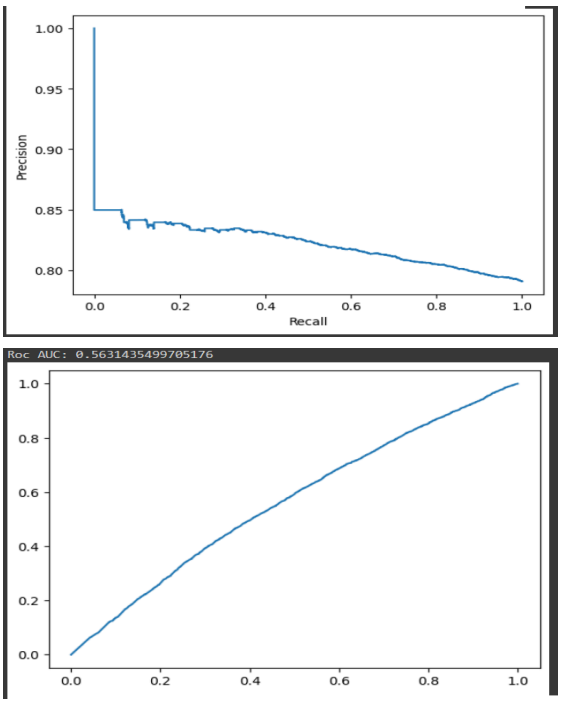
Fusionar el màxim nombre possible de tipus de forma en una sola.(Deixant només dos formes diferents)

Aquesta prova ha acabat en dos resultats diferents, hi ha més possibilitats però amb les observades ja s'extreuen conclusions.

El primer resultat d'aquesta idea dona una mica més de 50% d'accuracy, la fusió aquí s'ha fet ajuntant les formes cap a light o cap a other mentres es manté un balanceig de dades equilibrat

El segon resultat dona al voltant de 75% d'accuracy, arribant a 79% amb svc i tracta de fer la fusió amb les formes light i other però només afegint els valors a other, així creant un desbalanceig important.

Fig. 4. Corbes Roc i PR



Aquestes corbes corresponen al segon resultat, tot i tenir un bon accuracy, donat el gran desbalanceig i la forma de les pròpies corves, no es pot agafar com un resultat satisfactòri.

La següent matriu de confusió correspon al svc amb el 79% d'accuracy, s'observa que la majoria de prediccions correctes es troben en l'1 - 1, aquesta casella correspon als valors d'other(la classe desbalancejada), per tant, veiem que el bon accuracy surt d'aquí i no es pot agafar com a una mètrica correcte.

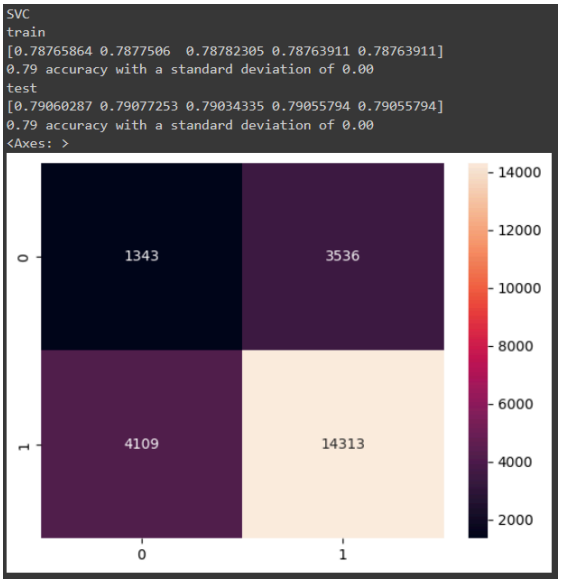


Fig. 5. Matriu de confusió desblanacejada

Eliminar totes les formes menys dues

Aquesta ha estat l'última prova i és la que a priori ha donat uns resultats una mica millors per analitzar-los com a acceptables.
S'ha ajuntat circle amb fireball i s'han eliminat totes les files que contenien al target un valor que no fos o fireball o light.

Això ha donat un target balancejat i a l'hora de classificar-ho amb els models tots han donat per sobre de 50% en el que accuracy respecta.

El millor ha estat RandomForestClassifier amb un 57% d'accuracy al test.

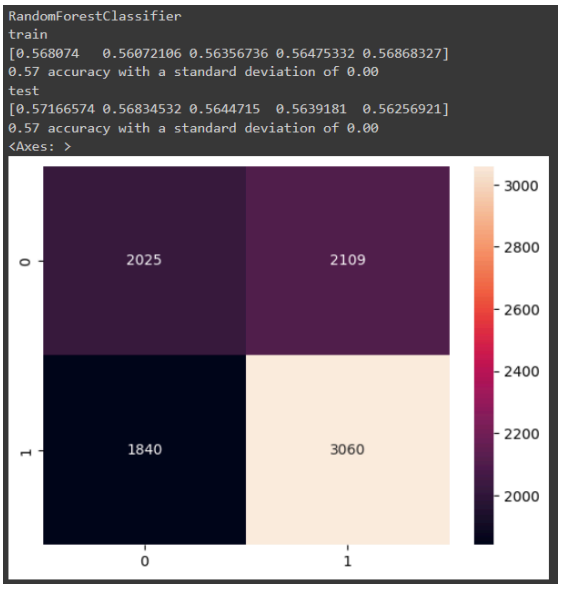


Fig. 6. Matriu de confusió dades balancejades(2 classes)

Veient la seva matriu de confusió i les corbes roc i pr, tot i semblar en un principi que els resultats havien de tenir una mica més de sentit, no ha estat així.

Cal destacar que en cap mètode una cerca d'hiperparametres ideal ha servit per millorar resultats.

4.Conclusions

Després de passar per tot l'anàlisi de tots els models i probes fetes es pot afirmar que no és possible predir quina forma tindrà el següent albirament d'un OVNI amb les dades que es proporcionen en el dataset utilitzat.

És un datatset molt complet en respecte a la quantitat de dades, aquestes si bé permeten una gran llibertat a l'hora d'analitzar visualment coses com on es fan més albiraments, com es veu en el seguent gràfic;

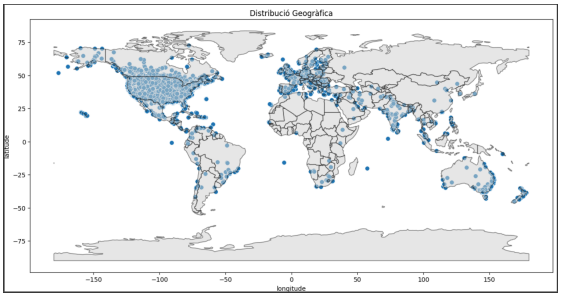


Fig. 7. Graf distribució geogràfica

A l'hora de classificar i haver de predir certs valors ja siguin la forma de l'OVNI o algun dels altres atributs, la tasca es complica degut a que, posant un exemple, dues formes d'OVNIs diferenets podrien tenir exactament els mateixos valors en tots els altres atributs ja que un mateix albirament es podria donar en el mateix lloc i moment per dues persones diferents i al tractar-se d'un tema subjectiu, acabar amb dos entrades del dataset exactament iguals excepte per la forma de l'OVNI.

5.Bibliografia

Dataset Utilitzat:
https://www.kaggle.com/datasets/willianoliveiragibin/ufo-sightings/data