Pokémon. Trobant els vertaders llegendaris amb models d'aprenentatge computacional.

Víctor Valero Carrascco

Abstract—Aquest projecte és un Cas Kaggle realitzat amb el dataset de Pokémon. Primer es farà un estudi del dataset, seguidament s'extreuran les columnes innecessàries, a continuació es comprovaran els models d'aprenentatge automàtic per aplicar Clustering al problema i finalment amb dos models es trobaran els que es consideren els vertaders llegendaris de la saga Pokémon, des de la generació 1 fins a la 7. Es tindran en consideració 72 llegendaris per model, ja que és el nombre exacta de llegendaris que hi ha al dataset. Els resultats són molt interessants i podrem observar si la empresa creadora de Pokémon, Nintendo, escull els llegendaris de forma arbitrària o hi ha quelcom al darrere que els identifica.

Keywords—Pokémon, Image Classification, Clustering, Legendary Pokémon, Machine Learning, Gaussian Mixture Model, Feature Engineering, Data Visualization, Kaggle Dataset, K-Means

1 INTRODUCCIÓ

Com a entusiasta de Pokémon, l'interès de realiztar el projecte va sorgir amb la fascinació que senteixo de la franquícia i la curiositat de comprendre millor les característiques dels Pokémon amb un enfoc analític. Aquesta investigació forma part d'un Cas Kaggle, una plataforma coneguda per la resolució de problemes mitjançant l'ús de dades i tècniques d'aprenentatfe automàtic, també està relacionat amb l'assignatura d'Aprenentatge Computacional i tracta sobre trobar els vertaders Pokémon llegendaris, fent referència a si els Pokémon llegendaris que hi ha al joc tenen característiques que els destaquen o són arbitraris. En aquesta anàlisi, la base de dades principal és el fitxer "pokemon.csv", no s'ha fet servir cap tipus de base de dades externa, així que el projecte està centrat exclusivament en la manipulació i anàlisi d'aquest conjunt de dades en concret. Encara que no s'ha utilitzat una bibliografia formal, s'ha consultat un projecte similar a Kaggle cocm a punt de partida. Aquesta referència ha servit com a guia inicial per entendre l'estructura i l'enfocament del projecte, tot i que posteriorment s'ha pres autonomia per desenvolupar el projecte de forma pròpia i original. També, s'ha fet ús de la intel·ligència artificial per poder tenir suport en moments on no sabia com seguir.

2 PROPOSTA/METODOLOGIA

Primer de tot, s'ha definit la variable OMP_NUM_THREADS a 4, ja que ens permetrà realitzar correctament l'aplicació de models d'aprenentatge automàtic. Seguidament, s'han separat les columnes del dataset en 3 subdatasets diferents: Columnes numèriques, columnes objecte i columnes amb NaNs. Hi ha 4 columnes amb NaNs que són el pes del pokémon, l'alçada, el tipus secundari i el percentatge de que sigui mascle. Primer m'he fixat en la columna del tipus secundari, ja que tenia 384 NaNs, el valor més elevat, i és degut a que hi ha Pokémon sense tipus secundari, per exemple Pikachu és de tipus elèctric només, i al no tenir cap tipus secundari el data set ho interpreta com a NaN. En aquest cas s'ha decidit mapejar tota NaN com a 'None', així li atribuïm valors. Les altres columnes s'han modificat més endavant al projecte, però ho explicarem ara mateix,per al pes i l'altura s'ha fet la mitja de la columna, i per el percentatge de ser mascle la columna s'ha eliminat durant l'anàlisi.

2.1 Anàlisi de columnes objecte

A continuació tenim l'anàlisi columna a columna per decidir si mantenir la columna al model final d'entrenament. Primer s'ha fet una anàlisi de totes les columnes objecte que són les columnes que tenen variables categòriques, és a dir que tenen text. De les 7 columnes objecte primer ens fixem en "capture rate", ja que es pot observar que hi ha tot nombres i no hauria de correspondre, però fent una cerca de valors únics veiem que hi ha un Pokémon que té diferent ratio de captura segons la forma que tingui, el que s'ha realitzat és mapejar la variable i deixar-la a 30, ja que és la més comú. Seguidament, s'ha tingut en compte les columnes de "Name" i "Japanese Name", aquestes columnes s'han eliminat ja que són identificatives de la fila i no aporten valor real a l'estudi. A continuació s'ha estudiat la columna "abilities", aquesta columna indica les possibles habilitats que pot tenir un Pokémon. S'ha observat el nombre d'habilitats que pot tenir un Pokémon normal i un legendari per separat i les habilitats més comuns en aquests dos grups, però s'ha decidit remoure-la també, ja que hi havia masses valors únics per tenir-la en compte. La columna "classification" també s'ha extret ja que és el nom que se li dona a cada Pokémon a l'hora de classificar-los i no era útil. Finalment, les dues columnes de "type 1" i "type 2" s'han mantingut, ja que són suficientment diferentes entre si i els tipus dels llegendaris amb la resta tenen una variació, per tant es pot considerar per l'aprenentatge. Per tant, de columnes objecte ens quedem només amb els tipus.

2.2 Anàlisi de columnes numèriques

Seguint amb les columnes numèriques hem hagut de dividir-les en altres tres subapartats, les columnes d'efectivitats, les d'estadístiques i les altres columnes. Les columnes d'efectivitat indiquen si el Pokémon és fort o feble contra un atac d'un determinat tipus. Aquestes columnes poden tenir els següents valors: 0,0.25,0.50,1,2 i 4. Aquest nombre és el nombre que s'utilitza per multiplicar-lo amb la potència base de l'atac que s'utilitzi, per tant, un atac amb 90 de potència multiplicat per 2 farà 180 de potència al Pokémon. S'ha decidit mantenir aquestes columnes, ja que estan relacionades amb els tipus i són gairebé úniques entre els Pokémon. Per les columnes estadístiques, són segurament les més importants per determinar si un Pokémon és legendari o no, així que s'han distribuït les columnes i extret els outliers positius. Al extreure els outliers positius, podrem veure si aquests pertanyen a Pokémon llegendaris o no, ja que són els que destaquen per la part positiva. Fent l'anàlisi estadística a estadística s'ha pogut observar que en gairebé totes les estadístiques al voltant del 40% dels outliers pertanyen a Pokémon llegendaris mentre que la resta no ho són. Però si ens

fixem en les estadístiques de poder base, tenim el 76% de Pokémon llegendaris, per tant en la suma d'estadístiques els Pokémon llegendaris guanyen. Finalment per les columnes numèriques tenim les columnes restants on comencem amb "percentatge male", aquí com s'ha esmentat prèviament teníem NaNs que es deuen al fet de que hi ha Pokémon sense gènere. Primer s'han comparat les dades de Pokémon no llegendaris sense gènere i Pokémon llegendaris sense gènere i s'obté que hi ha 63 Pokémon llegendaris i 35 normals. Sabent que hi ha 70 llegendaris tenim 63 sense gènere, un gran percentatge per tant és millor treure la columna. Seguidament està la columna "experience growth" i al fer un anàlisi de la columna, s'observa que tots els llegendaris tenen el mateix valor de creixemnt d'experiència menys 3, per tant millor treure la columna també. Les columna de "base happiness" i "base egg steps" passa com la columna anterior i és que els valors són poc diferents entre ells per tant és millor treure'les. Seguidament tenim les columnes de pes i altura on hem fet la mitja de les columnes per calcular les NaNs i s'han mantingut aquestes columnes perquè els valors són suficientment diferents entre ells i hi ha outliers interessants per fer l'entrenament. Finalment, la columna de generació s'ha tret perquè no es vol que els Pokémon de generacions amb més llegendaris tinguin.

2.3 Previ a selecció de model

Previament a fer la part de selecció de model, s'han de valorar totes les columnes que hem deixat amb una matriu de correlació, així podrem observar tots els valors que tinguin una gran correlació i agrupar-los o eliminar-los.

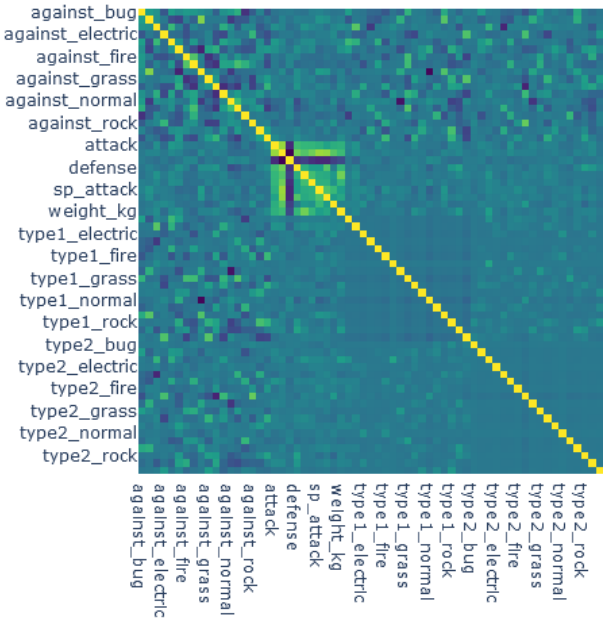

Fig. 1. Matriu de correlació de totes les columnas que s'han decidit incloure al model d'aprenentatge. Per observar bé els valors es recomana mirar el Jupyter Notebook.

A la Figura 1 es pot observar que hi ha una zona amb una gran diferència en el to del mapa de correlació. Això significa que les variables tenen una gran correlació entre elles i és millor o agrupar-les o eliminar-les. Aquestes variables són les de relació amb les estadístiques individuals i el poder base. Per tant, s'ha decidit eliminar totes les columnes d'estadística menys la de poder base i deixar aquesta última.

2.4 Selecció de model

Per avaluar el model s'ha fet servir la mètrica del Silhouette Score, això és degut a que les dades del projecte i el rendiment del clustering són més fàcils d'avaluar amb aquesta mètrica. Aporta una coherència interna, interpretació intuïtiva amb valors entre -1 i 1 i s'adapta a la forma i tamany dels Clusters. Amb aquesta mètrica s'han avaluat 4 models diferents i s'han obtingut els seus Silhouette Score pertinents, així que ens ha ajudat a escollir el millor model entre els possibles. S'han avaluat els models de K-Means, DBSCAN, Agglomerative i Bayesian Gaussian Mixture. Amb K-Means s'ha obtingut un score de 0.17, amb DBSCAN -0.21, amb Agglomerative 0.10 i amb Bayesian Gaussian Mixture 0.33. Poden semblar valors petits però aquest score indica si l'estructura dels clústers generats per aquests models són més robustes i millor definides, per tant tot i ser un valor relativament baix, no significa que siguin percentatges d'encerts o de errades.

Per tant, s'ha decidit realitzar un estudi amb Bayesian Gaussian Mixture ja que té el Silhouette Score més alt, però també s'ha decidit entrenar el model amb KMeans, ja que així es podran comparar els diferents resultats i observar similituds i possibles errors en uns i altres.

3 Experiments, resultats i anàlisi

Una vegada escollits els models i realitzat tot el filtratge de columnes que s'havia de fer, s' aplicarà els algorismes d'aprenentatge als dos models que s'ha comentat prèviament.

3.1 Algorisme K-MEANS

Per l'algorisme K-Means, primer s'han filtrat les columnes i s'han inserit a la variable X. Seguidament s'ha realitzat la codificació One-Hot per codificar les variables categòriques "type1" i "type2". Seguidament, s'han normalitzat les columnes i s'ha aplicat K-Means Clustering amb 2 clusters, que és el valor que millor Silhouette Score proporcionava. Seguidament s'ha definit un Umbral, en aquest cas era 8.8, ja que proporcionava un total de 72 llegendaris, una xifra molt propera al que és al data set, i per filtrar els Pokémon candidats a ser llegendaris s'han filtrat tots el Pokémon classificats al clúster 1 i que tinguin una distancia al centroide major a l'umbral.

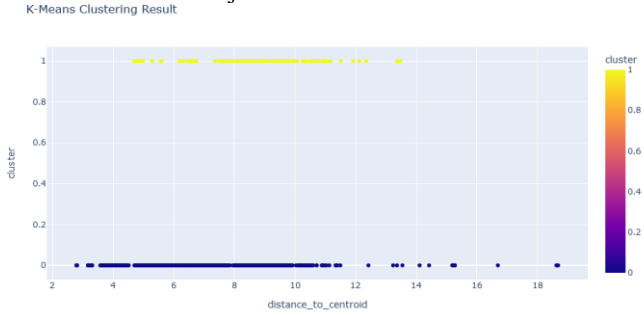

Fig. 2: Resultats del Clustering amb K-Means. Els punts superiors indiquen els Pokémon classificats al Cluster 1 I els inferiors els del Cluster 0. Per determiner els llegendaris són tots els punts superiors més a la dreta del 8.0.

Es pot observar a la següent figura els resultats de l'algorisme K-MEANS:

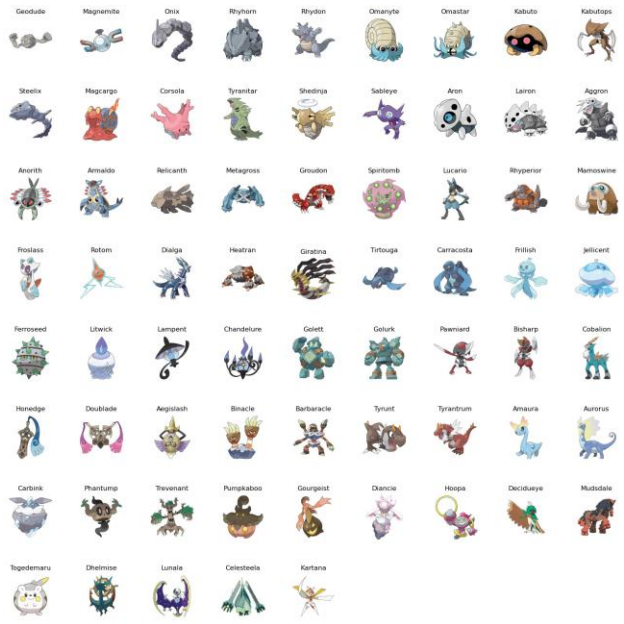

Fig. 3: Imatges de tots els Pokémon considerats llegendaris per l'algorisme K-MEANS.

Finalment, per concloure aquest algorisme, s'ha calculat el Silhouette Score final del model entrenat i dona un resultat de: 0.1728.

3.2 Algorisme Bayesian Gaussian Mixture

Primer de tot, per aquest algorisme s'han creat unes pipelines de les variables categòriques, una per les variables numèriques i una que les uneix. Seguidament s'ha aplicat l'algorisme amb un nombre de 30 components i s'ha comprovat que convergeixi, en aquest cas ha estat positiva la comprovació. Seguidament s'han calculat les densitats, s'ha considerat un umbral de 9 i s'han trobat les anomalies que están per sota de l'umbral.

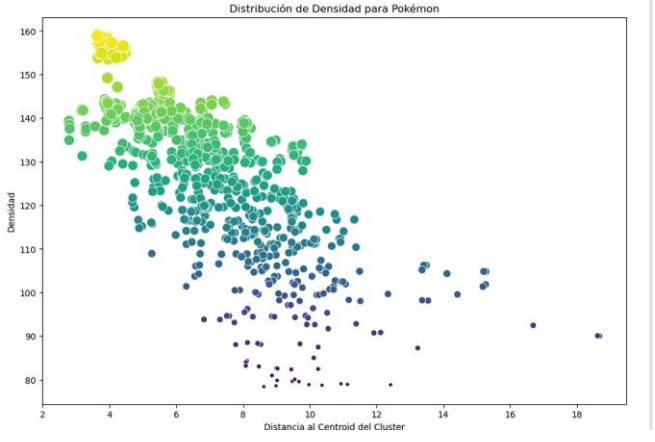

Fig. 4. Representació gràfica dels resultats del Algorisme de Bayesian Gaussian Mixture.

Podem observer a la següent imatge els Pokémon legendaris considerats per l'algorisme Bayesian Gaussian Mixture:

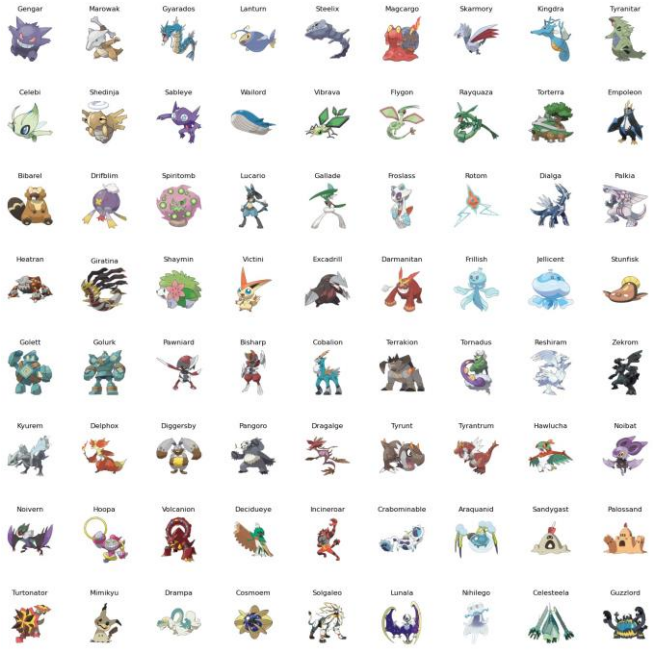

Fig. 5: Imatges de tots els Pokémon considerats llegendaris per l'algorisme Bayesian Gaussian Mixture.

Finalment, per concloure aquest algorisme, s'ha calculat el Silhouette Score final del model entrenat i dona un resultat de: 0.2541.

3.3 Conclusions dels 2 algorismes

Com podem observar hi ha molts Pokémon diferents entre els dos models, s'ha decidit fer una selección entre els dos models per veure quins Pokémon hi havia en comú. No com a objectiu de que siguin els vertaders llegendaris, però per veure'ls per pura curiositat.

Fig. 6: Imatges dels Pokémon considerats llegendaris que tenen en comú els dos algorismes.

Finalment, cal comentar que ens quedariem principalment amb els Pokémon que venen del Model de Bayesian Gaussian Mixture, ja que després d'avaluar les dues tècniques de clustering, s'observa que el model de Bayesian Gaussian Mixture ha obtingut un Silhouette Score més alt (0,254) en comparació amb el model de K-Means (0,173). Això indica una millor cohesió dins dels clusters i una millor separació entre ells amb el model de Bayesian Gaussian Mixture.

Un Silhouette Score superior indica que la estructura dels clusters generats pel model de Bayesian Gaussian Mixture és més robusta i millor definida. Aquest resultat suggereix que aquest model pot ser més adequat per a la classificació dels

Pokémon en comparació amb el K-Means.

Cal tenir en compte que el Silhouette Score no proporciona una mesura directa de la precisió en la classificació dels Pokémon com a llegendaris o no. Per tant, per realitzar un millor estudi es podrien tenir en consideració realitzar altres avaluacions i considerar altres factors i realitzar de nou l'aprenentatge.

5 CONCLUSIONS

Amb aquest treball s'ha aconseguit veure una nova perspectiva dels Pokémon llegendaris. Tot i que el model i l'execució tinguin molt marge de millora, es pot observar que la gran majoria de Pokémon que hem que l'Algorisme troba com a llegendaris, en realitat no ho són. I ens pot aportar una mica de perspectiva sobre com Nintendo decideix de forma arbitrària quin Pokémon ha de ser llegendari, ja que normalmente es pensaría que són els més forts o més únics als jocs però podem observar que no és així. Sempre, tenint en compte que el model és millorable.

BIBLIOGRAFIA

[1] ChatGPT: OpenAI."ChatGPT." 2023. https://platform.openai.com/models/chatgpt
[2] Rohan Asokan. "The Complete Pokemon Images Data Set." Kaggle,2018. https://www.kaggle.com/datasets/arenagrenade/the-complete-pokemon-images-data-set
[3] Ubiratan Filho. "Predicting the Real Legendary Pokémon." Kaggle,2019. https://www.kaggle.com/code/ubiratanfilho/predicting-the-real-legendary-pok-mon
[4] Rounak Banik. "Pokemon."Kaggle,2017. https://www.kaggle.com/datasets/rounakbanik/pokemon