Universitat Autònoma de Barcelona

MÉS ENLLÀ DELS PODERS: ANÀLISI ESTADÍSTICA DE L'UNIVERS DELS SUPERHEROIS

Autor: Marco Antonio Giles López

Tutor: Joan Gasull Jolís

14 de juny de 2024

Índex

1 Resum i Abstract

1.1 Resum

En aquest estudi s'ha realitzat una anàlisi estadística exhaustiva de l'univers dels superherois utilitzant tècniques estadístiques. L'objectiu principal ha estat identificar les diferències entre els poders dels superherois segons el gènere i en quina mesura influeixen, a més de determinar els factors que influencien el Tier (semblant al nivell) dels superherois per posteriorment crear models predictius de regressió i classificació. Prèviament a l'anàlisi estadística s'han preprocessat les dades i s'ha dut a terme una anàlisi descriptiva d'algunes variables importants per l'estudi.

Després de l'estudi de la influència dels poders al gènere dels superherois mitjançant un model de regressió logística, s'ha estudiat que pels superherois de gènere femení els dos superpoders més influents són Acrobatisme i Manipulació del Cabell i pels de gènere masculí són Nou poder i Uni-Poder.

Per a la predicció del Tier dels superherois, s'ha utilitzat un model de Random Forest tant per regressió com per classificació.
El model de classificació ha mostrat un accuracy del 85.1%, com a variables més influents el nivell de poder, la durabilitat, la velocitat.
D'altra banda, el model de regressió ha obtingut com a estimadors un MSE de 0.35 i un coeficient de determinació de 0.94, tenint com a variables més influents la durabilitat, el nivell de poder i la intel·ligència.

Aquest estudi proporciona una visió detallada de les característiques dels superherois i dels factors que determinen el seu nivell, així com els poders que influeixen en el seu gènere, oferint una base sòlida per a futurs treballs en aquest camp.

El llenguatge de programació per a realitzar la part pràctica de l'estudi ha estat Python, mitjançant Jupyter Notebook. [6] [7] La part de redactat de l'informe s'ha realitzat amb el llenguatge LaTeX, mitjançant Overleaf.

1.2 Abstract

In this study, an exhaustive statistical analysis of the universe of superheroes has been carried out using statistical techniques. The main objective has been to identify the differences between the powers of superheroes according to the gender and to what extent they influence, in addition to determining the factors that influence the Tier (similar to the level) of superheroes to later create predictive models of regression and classification. Before this statistical analysis, the data has been pre-processed and a descriptive analysis of some important variables for the study has been done.

After the study of the influence of powers on the gender of superheroes through a logistic regression model, it has been studied that for female superheroes the two most influential superpowers are Acrobatism and Hair-Manipulation and for male gender ones are New Power and Uni-Poder.

For the Tier prediction of superheroes, a Random Forest model has been used for both regression and classification.

The classification model has shown a accuracy of 85.1%, as the most influential variables are the power, durability and speed.

On the other hand, the regression model has obtained as estimators an MSE of 0.35 and a coefficient of determination of 0.94, having as most influential variables the durability, the power and the intelligence.

This study provides a detailed view of the characteristics of superheroes and the factors that determine their level, as well as the powers that influence their gender, offering a solid basis for future work in this field.

The programming language to perform the practical part of the study has been Python, by Jupyter Notebook. [6] [7] The part of the report was written using the LaTeX language, using Overleaf.

2 Presentació

2.1 Motivació

L'univers dels superherois és molt present a la societat des de fa dècades, des de les primeres aparicions en còmics fins a les adaptacions cinematogràfiques. Ídols d'infants i admirats pels adults els superherois mouen masses de tota mena de públic.

El fet de trobar un conjunt de dades amb tants registres de superherois i amb tantes característiques va fer que tingués molt clar que volia realitzar aquesta anàlisi estadística.

El meu objectiu principal era detectar insights interessants d'aquest univers com per exemple diferències entre gèneres o quins aspectes defineixen el nivell d'un superheroi.

2.2 Les dades

2.2.1 Font i Descripció

El conjunt de dades de l'estudi s'ha obtingut del repositori Kaggle, anomenat "Superhero Characters and Powers".

La font primària de les dades és la pàgina web *SuperHeroDB* i, mitjançant web scrapping, l'autor Baraa Zaid ha obtingut el conjunt de dades final.

El conjunt de dades conté informació sobre diversos superherois de diferents universos. Cada fila representa un superheroi individual i inclou una varietat d'atributs que descriuen les seves característiques i capacitats.

Els atributs inclouen característiques físiques, com per exemple de quina espècie és o de quin gènere és. També inclouen característiques de combat, com la durabilitat al combat, la velocitat o la força. A més, proporcionen informació com el creador del superheroi i l'univers al qual pertany, entre d'altres.

És important saber que en aquest conjunt de dades apareixen diferents versions del mateix superheroi. És a dir, no només hi ha un Spider-Man, com a exemple. El conjunt de dades diferència cadascuna de les versions del mateix personatge, per exemple l'Spider-Man de Tom Holland de l'Spider-Man de Tobey Maguire. És per aquest motiu que el conjunt de dades és tan gran tot i existir només 11000 personatges diferents aproximadament.

2.2.2 Les variables

Variable	Descripció
Alignment	L'alineació moral del superheroi (ja sigui "Bo"o "Dolent")
Alter_Egos	Qualsevol altra identitat o nom pel qual es coneix el superheroi
Base	La ubicació de la base d'operacions del superheroi
Character	El nom del superheroi (per exemple, Batman)
Class_Value	Un valor numèric que representa el nivell de poder o classe del superheroi
Collections	Qualsevol grup o col·lecció a la qual el superheroi pertany
Combat	La competència del superheroi en combat, valorada en una escala de 1-100
Creator	L'entitat o persona que va crear el superheroi
Durability	La capacitat del superheroi per resistir danys, valorada en una escala de 1-100
Equipment	Qualsevol equip especial o arma que utilitzi el superheroi
Eye_color	El color dels ulls del superheroi
Formerly	Qualsevol identitat o grup anterior del qual formava part
Full_Name	El nom complet del superheroi (per exemple, Bruce Wayne)
Gender	El gènere del superheroi
Hair_color	El color del cabell del superheroi
Height	L'alçada del superheroi
IQ	El quocient intel·lectual del superheroi
Intelligence	La intel·ligència del superheroi, en escala de 1-100
Leader	Grups dels quals el superheroi és líder
Level	El nivell de poder o classe del superheroi
Member	Qualsevol grup o organització del qual és membre
Name	El nom del superheroi
Occupation	L'ocupació o feina principal del superheroi
Omnipotent	Les habilitats omnipotents del superheroi (en una escala de 0 a 100)
Omnipresent	Les habilitats omnipresents del superheroi (en una escala de 0 a 100)
Omniscient	Les habilitats omnisciants del superheroi (en una escala de 0 a 100)
Place_of_Birth	El lloc de naixement del superheroi
Power	El nivell general de poder del superheroi, valorat en una escala de 1-100
Relatives	Qualsevol parent conegut del superheroi
Species	L'espècie o raça del superheroi
Speed	La velocitat del superheroi, en una escala de 1-100
Speed_velocity	La velocitat màxima que el superheroi pot assolir, mesurada en m/s2
Strength	La força del superheroi, en una escala de 1-100

Variable	**Descripció**
Strength_Force	La quantitat màxima de força que pot exercir
Super_powers	Llista d'habilitats o poders especials que posseeix el superheroi
Tier	La categoria o classe de poder del superheroi
Universe	L'univers o continuïtat al qual pertany el superheroi
Weight	El pes del superheroi
History	La història del superheroi

Taula 1: Descripció de les variables del conjunt de dades

3 Preprocessament de les dades

Eliminació de duplicats

Durant el procés de preprocessament, s'ha portat a terme una revisió exhaustiva per identificar i eliminar registres duplicats de les dades. Això evita biaixos en les anàlisis de l'estudi.
S'han eliminat en total 3449 registres.

Eliminació de variables

Algunes variables es van considerar irrellevants per als objectius de l'estudi o presentaven una alta correlació amb altres variables. Les variables eliminades inclouen:

S'han eliminat algunes variables segons alguns criteris:

- **Variables amb més de 10000 valors faltants:** En aquest cas s'ha realitzat una anàlisi variable per variable per poder detectar no només els valors faltants, sinó també els valors que no aportaven informació a la variable, com -". A més, algunes de les variables no proporcionaven informació valuosa per l'estudi.

  Les variables eliminades segons aquest criteri han estat: *Height, Weight, Name, Eye_color, Hair_color, Omniscient, Omnipotent, Omnipresent, Alter_Egos, Equipment, Leader, Formerly, History, Base, Collections, Full_name, Member, Occupation, Place_of_birth, Relatives,History.*

- **Variables perfectament correlacionades:** S'ha decidit eliminar la variable *IQ*, ja que proporcionava exactament la mateixa informació que *Intelligence* però sense escala de l'1 al 100.

Eliminació de casos amb valors faltants

Per l'estructura de l'estudi i els objectius de l'anàlisi plantejats les variables que s'han considerat com a variables resposta en els models estadístics són *Gender* i *Tier*.
Per tal de no introduir biaixos als models ni distorsionar-los, s'ha decidit eliminar els casos on es presentava un valor faltant en aquestes variables.

A més, per tal de realitzar una anàlisi descriptiva adequada i considerant que imputar els superpoders d'un superheroi causaria un gran impacte (seria com canviar totalment el superheroi) s'ha decidit eliminar també els casos on es presentava un valor faltant a *Super_powers*

S'han eliminat en total 7686 registres.

Recodificació de variables

Algunes variables han estat recodificades per facilitar una millor anàlisi i per reduïr el conjunt de valors possibles. Les variables recodificades han estat:

- **Creator:** S'han deixat els creadors més comuns: Marvel Comics, DC Comics, Shueisha i Nintendo; els altres s'han codificat com Other.
- **Species:** S'han deixat les espècies més comunes: Human, Alien, God / Eternal i Demon; les altres s'han codificat com Other.

Imputació de valors faltants

Les variables amb menys de 10000 valors faltants han estat *Creator, Species, Alignment i Universe*. Aquestes variables comparteixen la característica que són variables categòriques, i per això es descarten tècniques d'imputació com la d'imputació per la mitjana, per exemple.

Finalment, s'ha optat per realitzar la tècnica d'imputació pels K-Veïns més propers (KNN). En aquest cas pels 5 veïns més propers. Per realitzar correctament la tècnica s'han codificat cadascuna de les variables en valors numèrics i s'han pres com a valors de referència els valors de les variables numèriques del conjunt de dades. És a dir, s'ha imputat la variable Creator segons les combinacions de valors de les variables numèriques, com per exemple Strength.

Variable	Recodificació	Imputació NA	Eliminació NA	Sense canvis
Alignment		X		
Character				X
Class_value				X
Combat				X
Creator	X	X		
Durability				X
Gender			X	
Intelligence				X
Level				X
Power				X
Species	X	X		
Speed				X
Speed_velocity				X
Strength				X
Strength_force				X
Super_powers			X	
Tier			X	
Universe		X		

Taula 2: Variables finals després del preprocessament

4 Anàlisi Descriptiva

A continuació es mostren alguns gràfics on podrem visualitzar principalment la distribució de les variables que s'han estudiat com a variable resposta dels nostres models.

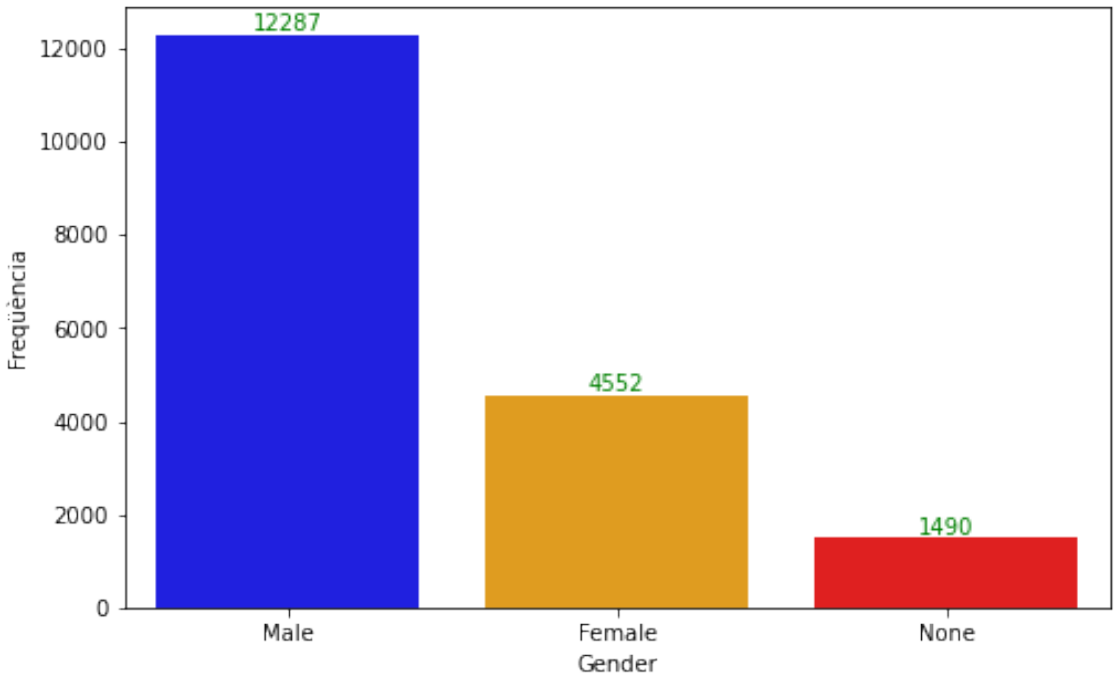

Figura 1: Distribució de Tier

En la Figura 1 es pot observar un clar desbalanceig en la variable *Gender*.
Hi ha gairebé el triple de superherois identificats com a gènere Masculí (12287) que superherois identificats com a gènere Femení (4552). També trobem 1490 superherois sense gènere identificat.
Aquest desbalanceig pot afectar molt negativament a la modelització de l'estudi, per això posteriorment aplicarem tècniques de sobremostreig per estudiar la relació del gènere amb els superpoders.

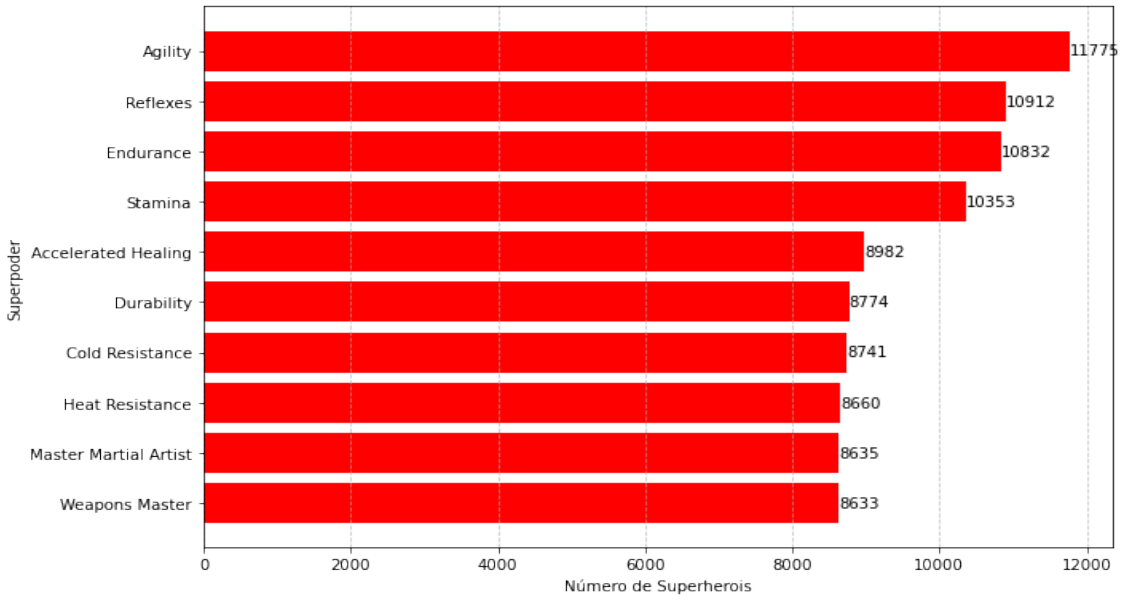

Figura 2: Top 10 superpoders més comuns

La Figura 2 mostra els 10 superpoders més comuns de tots els superherois del conjunt de dades. Els superpoders més comuns i la seva definició segons *SuperHeroDB* són:

1. **Agilitat**: Capacitat de respondre o canviar ràpidament adaptant la seva configuració estable inicial.
2. **Reflexos**: La capacitat de reaccionar més ràpidament del normal.
3. **Resistència**: La capacitat de suportar tensions difícils o desagradables més enllà dels límits i capacitats dels humans més grans.
4. **Aguant**: El poder de funcionar durant llargs períodes de temps sense cansar-se ni esgotar-se.
5. **Curació Accelerada**: La capacitat de curar-se ràpidament, tot i que varia en diferents nivells.
6. **Durabilitat**: Resistència a lesions físiques.
7. **Resistència al Fred**: El poder de ser resistent o immune a una, a algunes o a totes les formes de fred.
8. **Resistència a la Calor**: Virtualment immune als efectes de la majoria o totes les formes de calor.
9. **Mestre en Arts Marcials**: Un mestre de diferents tècniques d'atac i autodefensa, amb el propòsit d'autosuperació física i/o espiritual.
10. **Mestre d'Armes**: Capacitat alta de maneig d'armes.

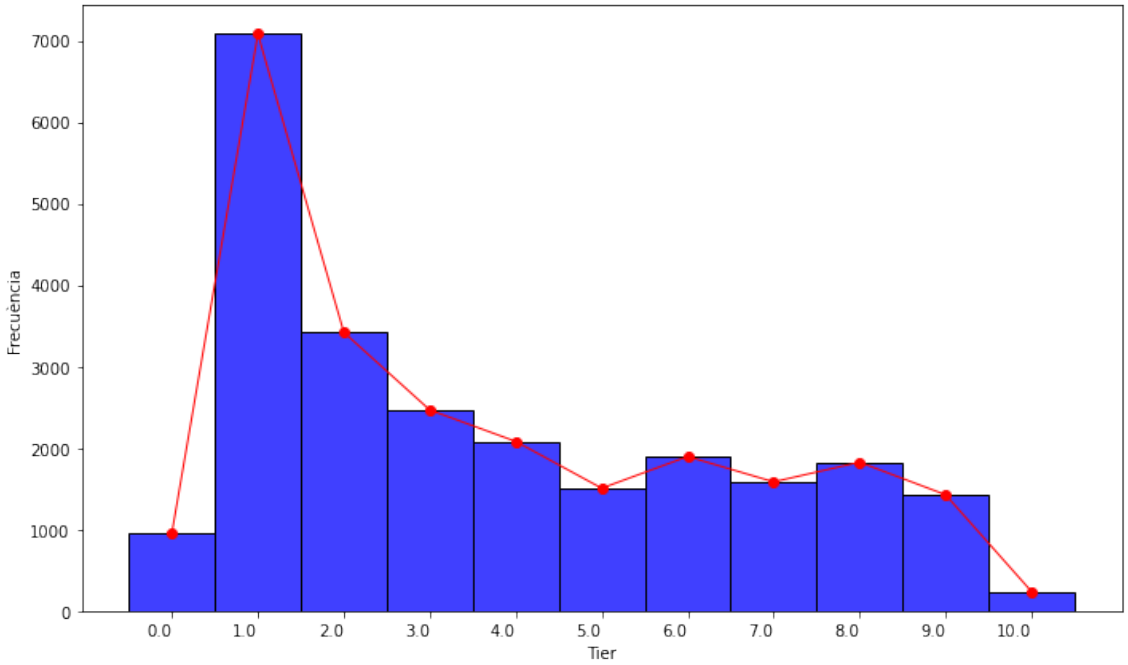

Figura 3: Distribució de Tier

La Figura 3 mostra la distribució de la variable Tier per tots els superherois del conjunt de dades.
Important: La variable Tier valora la categoria o classe de poder del superheroi, del 0 al 10. No és un ranking, és a dir, els superherois amb millor Tier són els que estan catalogats com Tier 10 i els pitjors com Tier 0.

S'observa com clarament hi ha un Tier predominant i és l'1, on aproximadament 7000 superherois estan catalogats. A més, els extrems 0 i 10 són les classes menys predominants, destacant el 10.
Un exemple de superherois catalogats com a millor i pitjor Tier és una de les versions de Loki i Baby Groot, respectivament.

Figura 4: Baby Groot (Tier 0) i Loki (Tier 10)

5 Anàlisi Estadística

5.1 La influència dels poders al gènere

L'objectiu principal d'aquesta part de l'estudi era poder establir relacions entre els diferents poders i el gènere d'un superheroi.

Al conjunt de dades hi ha un total de 485 superpoders diferents i hi ha superherois que posseeixen desenes de poders alhora.

L'estudi d'aquesta relació gènere-superpoders s'ha realitzat només amb els superherois de gènere masculí o femení, deixant de banda els que no tenen gènere per qüestió d'interès en l'anàlisi. A més, s'ha categoritzat el gènere dels superherois com 1 (masculí) i 0 (femení).

Un dels aspectes negatius de la variable Gender d'aquest conjunt de dades és el desbalanceig entre el gènere Masculí i Femení. Com hem observat a l'anàlisi descriptiva la descompensació entre gèneres és clara:

- Masculí: 12287 superherois
- Femení: 4552 superheroïnes

Per poder solucionar aquest aspecte de cara a una millor modelització de la variable Gender s'ha aplicat una tècnica estadística anomenada *SMOTE (Syntetic Minority Over-sampling Technique)*.

La tècnica SMOTE, com el seu propi nom indica, és una tècnica de sobremostreig que es basa en la creació de mostres sintètiques de la variable minoritària (en aquest cas el Gender Femení) mitjançant un algoritme que es centra en l'espai de característiques per generar noves instàncies amb l'ajuda de la interpolació entre les instàncies positives que es troben juntes. En combinació amb l'algorisme dels K veïns més propers (KNN), SMOTE busca millorar encara més el rendiment del model. [8]

Per a poder resoldre aquest problema de classificació binària (Male/Female) el tipus de model que s'ha escollit ha estat el de la *regressió logística*.

La regressió logística és un model estadístic per estudiar les relacions entre un conjunt de variables qualitatives Xi i una variable qualitativa Y. Es tracta d'un model lineal generalitzat que utilitza una funció logística com a funció d'enllaç. Un model de regressió logística també permet predir la probabilitat que passi un esdeveniment (valor d'1) o no (valor de 0) a partir de l'optimització dels coeficients de regressió. En el cas de l'estudi, l'esdeveniment es ser de gènere masculí (1) o ser de gènere femení (0). [1]

En el conjunt de dades emprat per l'estudi els superpoders de cada superheroi es trobaven en la mateixa variable separats per comes, pel que per poder fer la modelització hem separat aquests superpoders en variables dummies. Aquest conjunt de variables dummies són les que hem modelat com variables predictores.

La fórmula de la regressió logística és: [1]

$$\hat{y} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

- $\hat{y}$ representa la probabilitat de gènere Masculí.
- β_0 és l'intercept
- $\beta_1, \beta_2, \ldots, \beta_n$ són els coeficients dels predictors (dummies de superpoders) $x_1, x_2, \ldots, x_n$.
- $x_1, x_2, \ldots, x_n$ són els valors dels predictors (absència/presència de superpoder)

Els coeficients obtinguts del model pels 10 superpoders amb el coeficient més alt en valor absolut han estat els següents:

Intercept
0.231

Superpoder	**Traducció Catalana**	**Coeficient**
Acrobatics	Acrobatisme	-0.323
Hair Manipulation	Manipulació del Cabell	-0.278
New Power	Nou Poder	0.250
Soul Resistance	Resistència de l'Ànima	-0.249
Uni-Power	Uni-Poder	0.233
Omega Effect	Efecte Omega	0.226
Nigh-Omniscience	Nigh-Omnisciència	0.219
Regeneration Negation	Negació de Regeneració	-0.215
Molecular Combustion	Combustió Molecular	-0.198
Fear Manipulation	Manipulació de la Por	0.197

Es pot observar els 10 superpoders més influents en el gènere, ordenats de major a menor segons el valor absolut del seu coeficient del model. Els valors negatius dels coeficients indiquen una major probabilitat Cada coeficient representa el canvi en el logaritme de la raó de probabilitats (odds ratio) del gènere quan el superheroi té el superpoder o no el té.

Per tant, un coeficient positiu indica que la presència del superpoder augmenta la probabilitat que el gènere sigui masculí, mentre que un coeficient negatiu indica que la presència del superpoder augmenta la probabilitat que el gènere sigui femení. I quant més gran és el valor absolut del coeficient, més influent en el gènere.

Veiem que el superpoder amb major influència en el fet de tenir gènere Femení és Acrobatisme (tècnica de realitzar maniobres d'equilibri, destresa, agilitat i coordinació) amb coeficient -0.323 i el superpoder amb major influència en el fet de tenir gènere Masculí és New Power (poder antic i combinació còsmica de potència) amb coeficient 0.250.

Es poden observar de manera més gràfica aquests resultats:

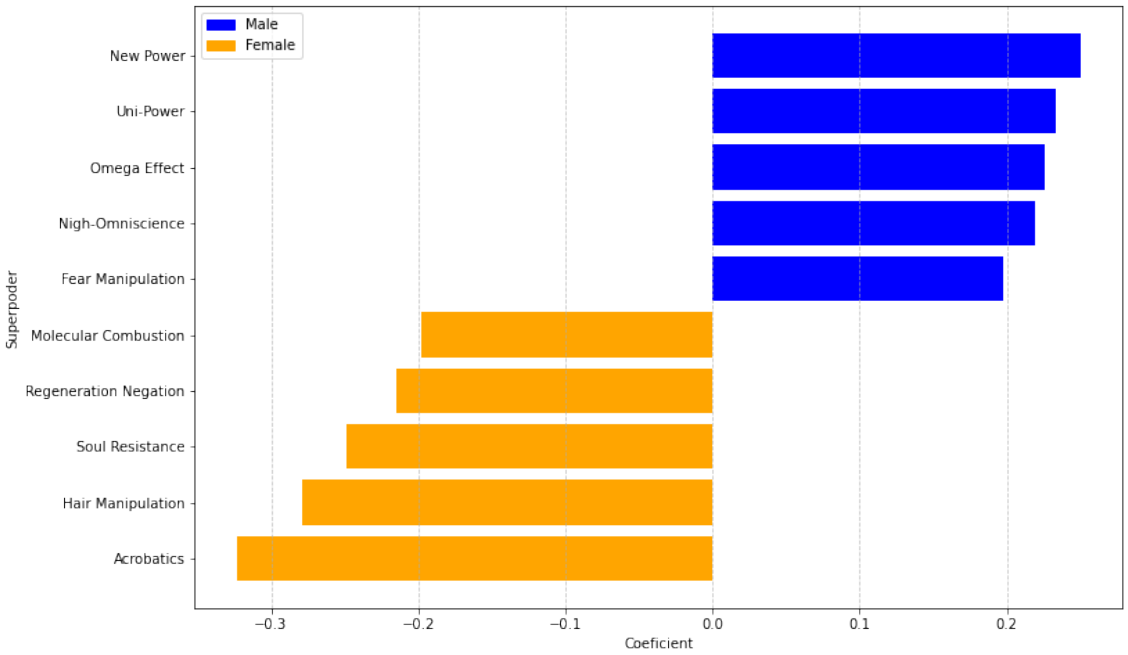

Figura 5: Top 10 Superpoders influents en el Gènere

Per tant, i com a exemple, la probabilitat $\hat{y}$ que el superheroi tingui gènere Masculí d'un superheroi que té com a poders Acrobatism, Manipulació del Cabell i Resistència de l'Ànima seria:

$$\hat{y} = \frac{1}{1 + e^{-(intercept + \beta_1 acrobatisme + \beta_2 manipulaciodelcabell + \beta_3 resistenciadel'anima)}}$$

$$\hat{y} = \frac{1}{1 + e^{-(0.231 + -0.323 \cdot 1 - 0.278 \cdot 1 - 0.249 \cdot 1)}}$$

$$\hat{y} = \frac{1}{1 + e^{0.619}}$$

La probabilitat de ser de gènere Masculí d'un superheroi amb aquests poders seria de 0.35 (probabilitat molt baixa per tenir els superpoders més influents en el gènere Femení).

5.2 Predicció del Tier dels superherois

El principal objectiu d'aquesta part de l'estudi ha estat crear dos models predictius de tipus Random Forest per a predir el Tier dels superherois en base a altres variables del conjunt de dades.

El Random Forest és una tècnica de Machine Learning que es basa en arbres que s'anomenen **arbres de decisió**.

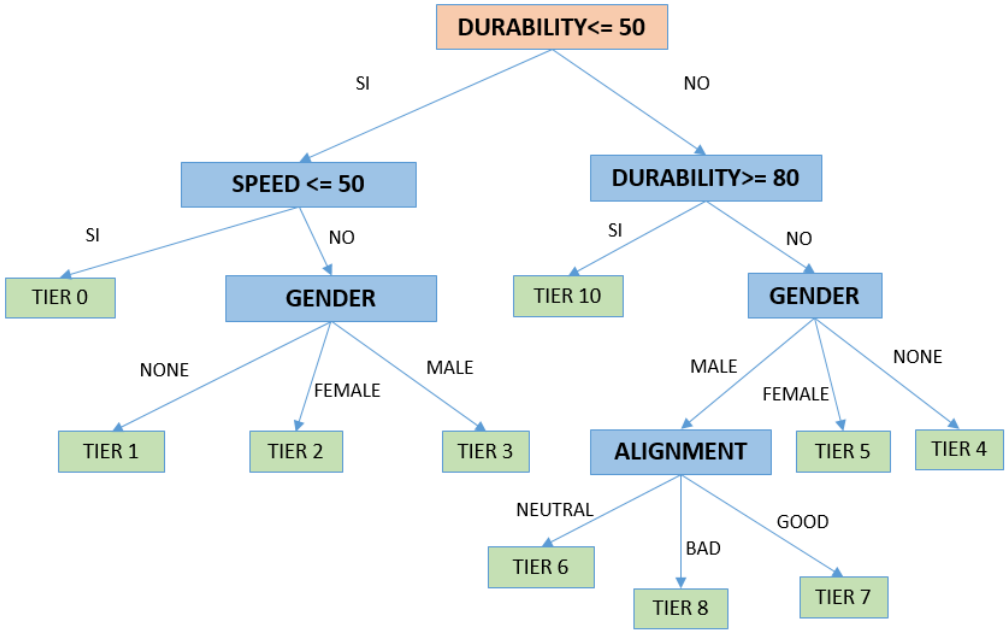

Figura 6: Exemple aleatori de representació d'un arbre de decisió

Els arbres de decisió estan formats per nodes i la seva lectura es realitza de dalt cap avall. Dins d'un arbre de decisió distingim diferents tipus de nodes: [2] [4]

- **Arrel**: En ell es produeix la primera divisió en funció de la variable més important.
- **Intern**: Després de la primera divisió trobem aquests nodes, que tornen a dividir el conjunt de dades en funció de les variables.
- **Fulla**: Se situa a la part inferior de l'esquema i la seva funció és indicar la classificació definitiva.

Un arbre de decisió ajuda a prendre una decisió gràcies a una sèrie de preguntes/tests la resposta binària dels quals portarà a la decisió final. Cada pregunta és un node i en funció de la resposta a la pregunta es pren un camí que portarà definitivament a una fulla.

Un random forest està compost per diferents arbres individuals relacionats amb les decisions. A partir d'aquest grup de decisions, es realitza una votació i els vots majoritaris es prenen per a obtenir una mitjana o una classificació.

Alguns dels avantatges de Random Forest són: [5]

- És una solució per a problemes de regressió i de classificació.
- Evita el sobreajustament gràcies a la combinació de resultats de diferents arbres.
- Permet controlar l'espai de característiques i calcular la seva importància al model.

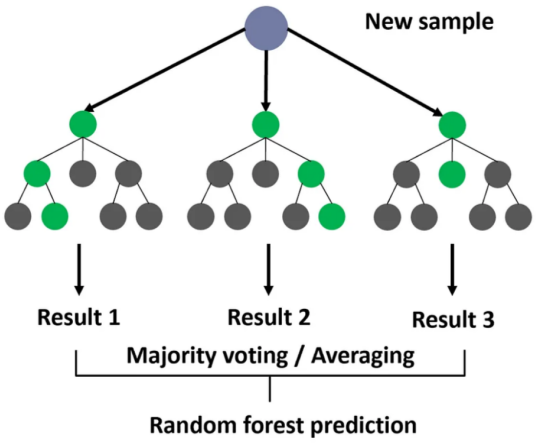

Figura 7: Representació d'un Random Forest

Per a problemes de classificació es pren la predicció de cada arbre i s'aplica el criteri de la majoria de vot i per a problemes de regressió es pren la mitjana de les prediccions de tots els arbres com a predicció final.

A l'hora d'entrenar el model predictiu Random Forest existeixen uns paràmetres que no s'aprenen a partir de les dades, anomenats **hiperparàmetres**. Aquests s'han d'establir abans d'entrenar el model.
Els hiperparàmetres que s'han treballat a l'estudi: [3]

- **n_estimators**: Nombre d'arbres del bosc.
- **max_depth**: Profunditat màxima de cada arbre
- **min_samples_split**: Nombre mínim de mostres necessàries per a dividir un node intern.
- **min_samples_leaf**: Nombre mínim de mostres necessàries per a estar en una fulla. Està molt associat amb min_samples_split.
- **bootstrap**: True/False. Indica si es considera o no reemplaçament per files a l'hora de construir cada arbre.

La biblioteca de Python Sklearn proporciona una classe anomenada **GridSearchCV** que s'utilitza per poder ajustar un model amb diferents combinacions d'hiperparàmetres. Els possibles valors d'aquests hiperparàmetres s'han de definir prèviament, i en el cas d'aquest estudi han estat:

- n_estimators: [100, 200, 300, 400]
- max_depth: [10, 20, 30, None],

- min_samples_split: [2, 5, 10],
- min_samples_leaf: [1, 2, 4],
- bootstrap: [True, False]

GridSearchCV permet poder visualitzar i ajustar el model amb la millor combinació d'-hiperparàmetres segons la puntuació obtinguda després de validació creuada K-fold Cross Validation.

En l'estudi s'ha validat el model amb *5-fold Cross Validation*. Aquesta és una tècnica per a la validació del model on, en aquest cas, es divideix el conjunt de dades en 5 parts (folds). Posteriorment s'entrena el model 5 vegades i en cada iteració una de les parts es pren com a dades test/validació i les altres 4 com a dades d'entrenament. Els resultats finals s'avaluen un cop realitzades totes les iteracions.

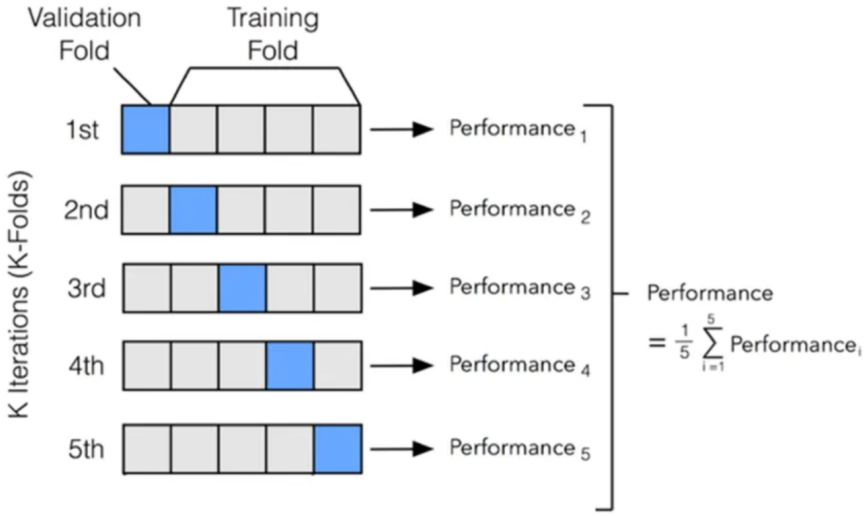

Figura 8: Representació gràfica de 5-fold Cross Validation

A més, en aquest estudi s'han seleccionat les 10 variables més importants per ajustar el model. És important saber que aquesta selecció s'ha realitzat un cop passades totes les variables categòriques a dummies, pel que una variable de les considerades 10 més importants podria ser per exemple *Species_Human*, variable que indica si el superheroi és d'espècie humana o no (binària).

La variable **Tier** té una característica important i és que es pot tractar com a categòrica i també com a numèrica, pel que s'ha realitzat l'estudi de modelització d'ambdues maneres aprofitant que Random Forest permet realitzar problemes de regressió i de classificació.

A continuació es mostren els resultats obtinguts.

5.2.1 Regressió amb Random Forest

Pel model de regressió amb Random Forest per a predir el Tier dels superherois no s'han inclòs les variables *Class_value* i *Level* per la seva correlació i semblança en el significat amb la variable resposta i tampoc *Speed_velocity* i *Strength_force* perquè ja tenim aquesta informació escalada en les variables *Speed* i *Strength*, respectivament.

Les 10 variables amb major importància del model han estat:

Variable	Importància
Durability	0.762
Power	0.092
Intelligence	0.025
Speed	0.016
Strength	0.012
Combat	0.007
Super_powers_Nigh-Omnipotence	0.002
Species_Alien	0.001
Creator_DC Comics	0.001
Creator_Other	0.001

Observem com la variable Durability té una importància molt destacada sobre la resta i, com s'ha comentat anteriorment, trobem algunes variables dummies com a variables importants.

Els hiperparàmetres òptims considerats:

Hiperparàmetre	Valor
Bootstrap	True
max_depth	None
min_samples_leaf	1
min_samples_split	2
n_estimators	400

Això vol dir que el millor model té 400 arbres al bosc, sense profunditat màxima, 1 mostra necessària per estar a una fulla i 2 mostres necessàries per dividir un node intern. A més es considera reemplaçament per files a l'hora del mostreig als arbres.

Els estimadors calculats:

Estimador	Valor
Mean Squared Error	0.35
$R^2 Score$	0.94

Un valor de MSE més baix indica un millor ajust del model a les dades. En aquest cas, un MSE de 0.355 indica que, en mitjana, hi ha aquesta desviació de les prediccions als valors reals.

El coeficient de determinació mesura la variabilitat explicada pel model, pel que un valor més alt indica un millor ajust el model. Un coeficient de 0.94 és un indicador de bon ajust, ja que el 94% de la variabilitat de les dades està explicada pel model.

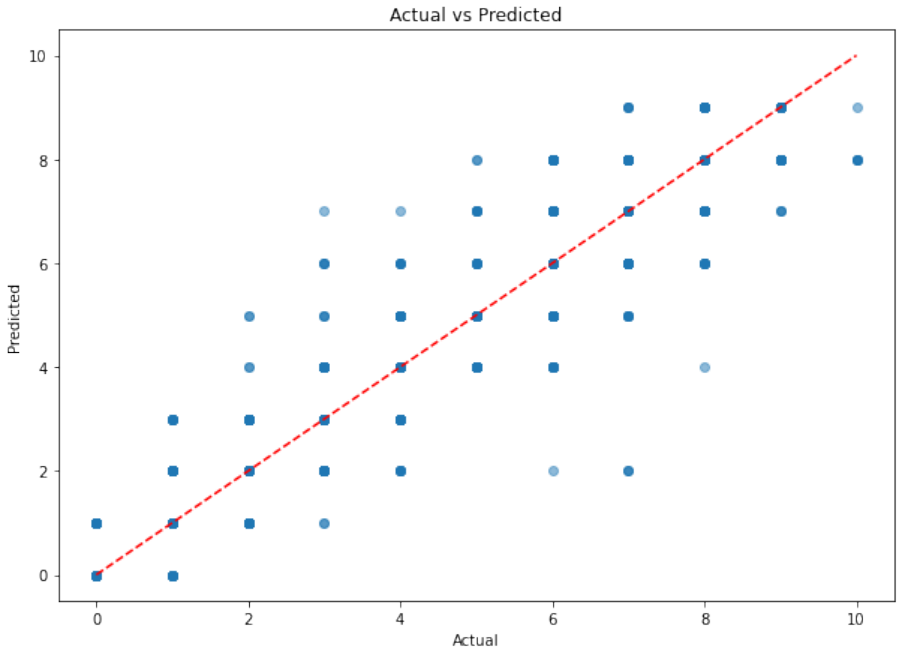

Figura 9: Actual vs Predicted Random Forest Regressió

A la Figura 9 es pot observar un gràfic de punts on a l'eix X es representen els valors reals i a l'eix Y els valors predits.
S'inclou al gràfic la recta $y=x$, ja que els punts sobre aquesta recta indiquen prediccions perfectes, punts per sobre indiquen sobreprediccions i punts per sobre subprediccions.

La dispersió és considerable al voltant de la recta vermella tot i que hi ha punts molt llunyans a ella, que representen errors significatius. S'observa clarament com la pitjor predicció és pels superherois de Tier 10 ja que són casos minoritaris i no s'aconsegueix una bona precisió per aquesta classe.

5.2.2 Classificació amb Random Forest

Pel model de classificació amb Random Forest per a predir el Tier dels superherois, de la mateixa manera que quan s'ha realitzat el model de regressió i amb els mateixos criteris, no s'han inclòs les variables *Class_value* i *Level* i tampoc *Speed_velocity* i *Strength_force*.

Les 10 variables amb major importància del model de classificació han estat:

Variable	Importància
Power	0.074
Durability	0.074
Speed	0.054
Strength	0.051
Intelligence	0.036
Combat	0.033
Species_Human	0.011
Species_Other	0.008
Alignment_Good	0.007
Alignment_Bad	0.006

Observem com les variables Power i Durability tenen una importància destacada sobre la resta, tot i que no tan destacada com al model de regressió. També trobem algunes variables dummies entre les més importants.

Els hiperparàmetres òptims considerats:

Hiperparàmetre	Valor
Bootstrap	True
max_depth	None
min_samples_leaf	1
min_samples_split	2
n_estimators	400

Això vol dir que el millor model té 400 arbres al bosc, sense profunditat màxima, 1 mostra necessària per estar a una fulla i 2 mostres necessàries per dividir un node intern. A més, es considera reemplaçament per files a l'hora del mostreig als arbres.

L'estimador calculat:

Estimador	Valor
Accuracy	0.851

L'accuracy indica el percentatge de prediccions correctes realitzades pel model sobre el conjunt de dades de prova. En aquest cas, un accuracy de 0.851 indica que el model encerta en el 85.1% de les seves prediccions.

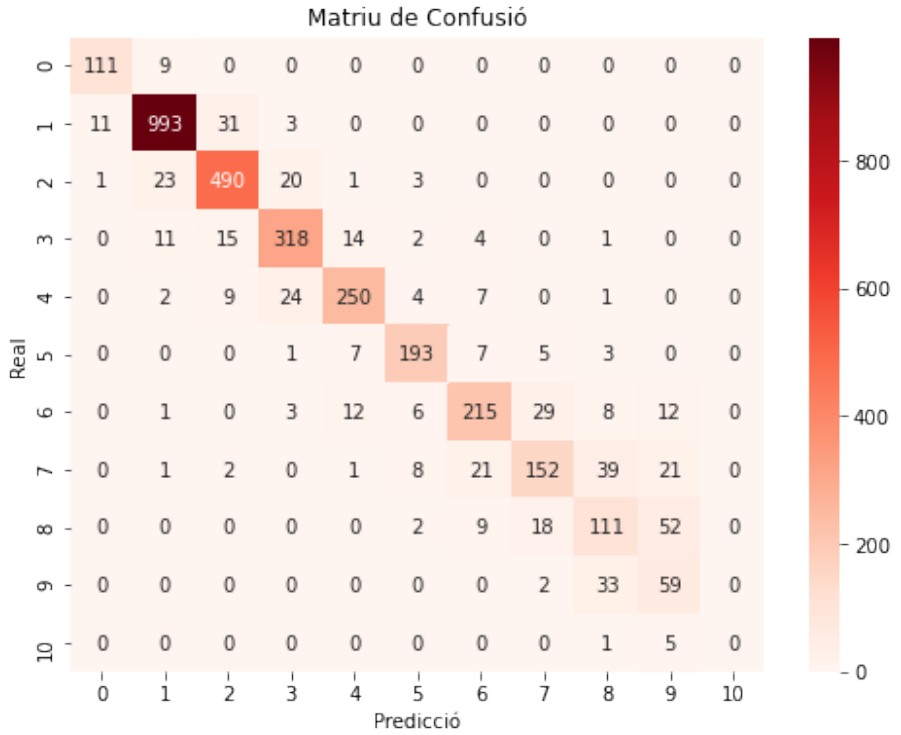

Figura 10: Matriu de Confusió Random Forest Classificació

La Figura 10 mostra la matriu de confusió del model de classificació, on a l'eix X es representen les prediccions i a l'eix Y els valors reals.

La matriu mostra un bon rendiment en general al model, sobretot a les classes 1 i 2. Les classes 8,9 i 10 mostren un rendiment inferior.
Els errors (punts fora de la diagonal) no semblen ser molt grans, estan a prop de la diagonal, sobretot en les classes amb millor rendiment predictiu.

Una idea per a un estudi futur on sigui més complicat poder obtenir un bon model com el plantejat i on la matriu de confusió presenti errors de major mesura que en la d'aquest estudi, seria poder aplicar majors penalitzacions a errors més grans.
Per exemple, per a un superheroi de Tier 5, penalitzar més si el model prediu que és de Tier 8 que si prediu que és de Tier 4, ja que es considera un error més gran.

6 Conclusions

Després de l'anàlisi de les qüestions plantejades en l'estudi mitjançant tècniques estadístiques, les conclusions principals són les següents:

1. **Influència dels Poders en el Gènere**:
   - S'ha observat quins són els poders més influents per a cada gènere i la seva influència a nivell numèric gràcies a un model de regressió logística.
   - Pels superherois de gènere femení els dos superpoders més influents són Acrobatisme i Manipulació del Cabell i pels de gènere masculí són Nou poder i Uni-Poder.

2. **Predicció del Tier dels Superherois**:
   - La utilització del model Random Forest ha demostrat ser efectiva per a la predicció del Tier dels superherois.
   - Tant per a problemes de regressió com de classificació, Random Forest ha demostrat ser una tècnica robusta, evitant problemes de sobreajustament.
   - S'han identificat hiperparàmetres òptims per als models de classificació i regressió, com ara la utilització de 400 arbres de decisió, sense una profunditat màxima, i altres paràmetres relacionats amb el mostreig i la divisió dels nodes.
   - El model de classificació Random Forest ha mostrat un accuracy del 85.1%, demostrant la seva bona capacitat predictiva.
   - El model de regressió Random Forest ha obtingut com a estimadors un MSE de 0.35 i un coeficient de determinació de 0.94, el que indica un bon ajust del model.

7 Bibliografia

Referències

[1] DataScientest. *¿Qué es la regresión logística?* 2024. URL: https://datascientest.com/es/que-es-la-regresion-logistica.

[2] Máxima Formación. *¿Qué son los árboles de decisión y para qué sirven?* 2024. URL: https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/.

[3] Panamá Hitek. *Optimización de hiperparámetros en Machine Learning*. 2024. URL: https://panamahitek.com/optimizacion-de-hiperparametros-machine-learning/?utm_content=cmp-true.

[4] Joaquín Amat Rodrigo. *Random Forest con Python*. 2020. URL: https://cienciadedatos.net/documentos/py08_random_forest_python.

[5] Inesdi Digital Business School. *Random Forest: ¿Qué es y cómo funciona?* 2024. URL: https://www.inesdi.com/blog/random-forest-que-es/.

[6] Charles R. Severance. *Python for Everybody: Exploring Data Using Python 3*. Scotts Valley, CA: CreateSpace Independent Publishing Platform, 2016.

[7] Jake VanderPlas. *Python Data Science Handbook*. O'Reilly Media, 2017.

[8] Analytics Vidhya. *Overcoming Class Imbalance Using SMOTE Techniques*. 2020. URL: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/.

8 Annex

Al repositori de Github TFGMarcoAntonioGilesLopez1599965 es pot trobar la documentació annexa següent:

- Conjunt de dades *superheroes2.csv* de l'estudi
- Fitxer .ipynb amb el codi Python
- Fixer .html amb el codi Python