Universitat Autònoma de Barcelona
Facultat de Ciències

Cas Kaggle: Classificació de categories de pel·licules i series de Netflix

Autor:

Gerard Lahuerta

1601350

16 de Desembre del 2022

Índex

1 Introducció

1.1 Presentació del treball

L'objectiu d'aquesta pràctica és, mitjançant la interfície proporcionada per Jupyter Notebook, estudiar i classificar una serie o pl·licula de l'empressa Netflix segons el seu rang de classificacions[1].
Les dades han sigut proporcionades per la web de Kaggle, concretament, la base de dades de series i pel·licules de Netflix.

El dataset que s'utilitza es pot trobar al següent enllaç:

https://www.kaggle.com/shivamb/netflix-shows.

[1] El rang de classificació de les series i pel·licules esta explicat a l'aprtat 2.1

1.2 Llibreries i importacions

Per tal de poder dur a terme aquesta tasca és imprescindible tenir intal·lades les següents llibreries, ja que s'utilitzen les funcions següents (d'entre altres).

Llibreria	**Funció utilitzada**
sklearn.datasets	make_regression
pandas (as pd)	read_csv DataFrame
matplotlib pyplot (as plt)	figure plot hist scatter
seaborn (as sns)	heatmap
sklearn.linear_model	LogisticRegression LogisticRegressionCV
sklearn.tree	DecisionTreeClassifier
sklearn.metrics	confusion_matrix ConfusionMatrixDisplay precision_recall_curve average_precision_score roc_curve auc precision_score
sklearn.model_selection	cross_val_score cross_validate train_test_split
sklearn.ensemble	RandomForestClassifier
sklearn.neighbors	KNeighborsClassifier
sklearn.pipeline	make_pipeline
sklearn.preprocessing	StandardScaler
sklearn.svm	SVC LinearSVC
sklearn.cluster	KMeans
sklearn.mixture	GaussianMixture
warnings	filterwarnings
numpy (as np)	meshgrid concatenate

Taula 1: Llibreries i funcions utilitzades

1.3 Criteris i assumcions

S'han tingut en compte els següents criteris per a la correcta classificació de les dades:

1. Degut a la gran quantitat d'informació que s'obté en el proces de tractament de les dades; només s'utilitzarà amb les més rellevant.

2. Es preferible etiquetar un element en una classe que no hi pertany que no etiquetar un element d'una classe a la que hi pertany.

2 Gestió i estudi del Dataset

2.1 Explicació del Dataset

El dataset tracta sobre les series i pel·licules que hi disposa l'empressa Netflix per al seu consúm.
L'objectiu del treball és, mitjançant les dades disposades, poder general models capaços de classificar, mitjançant el minim nombre d'atributs disponibles, les categories a les que pertany les noves series i pel·licules (apartir d'ara ens referirem a les series i pel·licules com: *producció*) que hi afegeixen en la plataforma.

El dataset en qüestió té una mida de 8807 x 12 (files x columnes).
Els 12 atributs recollits de les produccions són:

Atribut	**Explicació**	**Tipus de dada**
type	serie o pel·licula	string
title	titól de la producció	string
director	nom del director de la producció	string
cast	actors de la producció	string
country	països on és grabada la producció	string
data_added	data quan va ser afegida la producció al catèleg	string
release_year	any quan va ser estrenada la producció	int
rating	etiqueta de destinació de la producció (TV/PG)	string
duration	durada de la producció (minuts o temporades)	string
listed_in	generes de la producció	string
description	synopsis de la producció	string

Taula 2: Explicació dels atributs i el seu rang de valors que assoleixen

Cal destacar que la gran majoria d'atributs son de tipus string o llistes d'string; per aquest motius es decideix utilitzar un one-hotter-coding dels atributs de tipus string.

També, cal mencionar, que part dels atributs poden ser omesos per que de manera lógica és veu que no hi tenen relació entre el genere i el valor de l'atribut (a l'hora de classificar futures produccions que s'incloguin al cataleg) o per la dificultat que hi pot existir per a utilitzar el valor de l'atribut de manera generica per a classificar la producció. Alguns d'aquests atributs que és poden ometre son: data_added[2] i director[3].

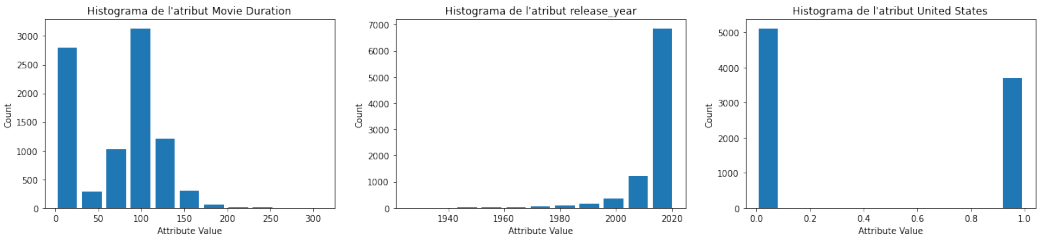
[2]S'observa de manera clara com el any que va ser afegida al cataleg una producció no afecta a quin genere pertany.

[3]El director (així com el repart de la producció), si bé es pot utilitzar per a classificar la producció ja que molt artistes/directors graben generes semblants, la quantitat de paràmetres que ens generaria el one-hotter-coding per aquests atributs seria tant gran que dificultaria la seva gestió i utilització en la classificació.

2.2 Distribució de les dades

S'iniciarà l'estudi del dataset observant la distribució de les dades per intuïr relacions senzilles des d'on començar a plantejar els primers models, així com crivar els atributs rellevants per a fer la classificació.

Es mostren ara alguns dels histogrames generats, així com *scatter-plots* del *price_range* respecte les variables.



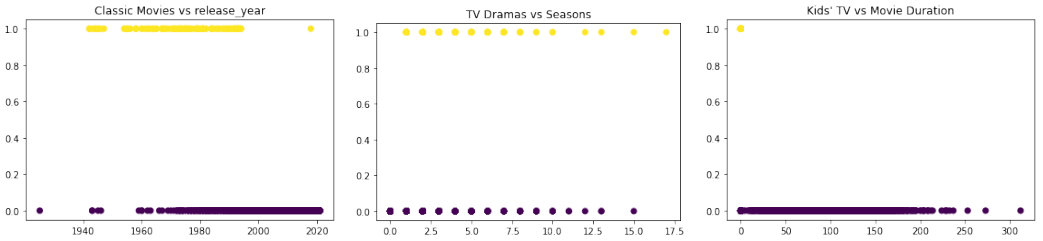
(a) *Movie Duration* (b) *Release Year* (c) *United State of America*

Figura 1: Mostra dels histogrames generats per l'estudi inicial

S'observen les següents característiques dels histogrames[4]:

- Degut al one-hotter-coding la majoria d'atributs tenen distribucions binaries.
- S'observa com no hi existeix una uniformitat en el nombre de produccions de diferents generes, sent més populars uns que altres. Degut aquesta no uniformitat, hi pot haber problemes en la classificació dels generes. És discutirà aquest tema més endavant.

Per a obtenir millors conclusions, es decideix veure les relacions dels atributs representant els objectius respecte la resta d'atributs. Mostrem ara alguns exemples dels resultats obtinguts



(a) Clasic Movies vs Release Year (b) TV Drama vs Seasons (c) Kid's TV vs Movie Duration

Figura 2: Mostra dels *scatter-plots* generats per l'estudi inicial

És conclou dels *scatter-plots* [5] que no es percep de manera intuïtiva cap relació entre les variables.

És procedeix a estudiar les correlacions de variables per trobar relacions entre elles i els nostres atributs objectius.

[4] És pot consultar el conjunt de histogrames al notebook entregat conjuntament amb aquest informe
[5] El conjunt sencer d'*scatter-plots* són al notebook entregat conjunt aquesta memória

2.3 Correlació de les variables

S'ha decidit estudiar la correlació entre els atributs que conté la base de dades per tal d'analitzar la importància entre ells per poder trobar els millors paràmetres per classificar i decidir com tractar les incongruències de les dades exposades anteriorment a l'apartat 2.1.

Gènere	**Atribut**	**Correlació**
Documentaries	Seasons	-0.149
International TV Shows	Movie Duration	-0.572
	Taiwan	0.178
	South Korea	0.229
	Japan	0.173
	United States	-0.314
	Seasons	0.312
TV Dramas	Movie Duration	-0.414
	Seasons	0.340
TV Mysteries	Movie Duration	-0.143
	Seasons	0.148
Romantic TV Shows	Movie Duration	-0.282
	Taiwan	0.250
	South Korea	0.239
	United States	-0.127
	Seasons	0.151
Spanish-Language TV Shows	Movie Duration	-0.191
	Colombia	0.309
	Argentina	0.131
	Spain	0.186
	Mexico	0.260
⋮	⋮	⋮

Taula 3: Correlacions d'alguns atributs

Mostrem a la taula els atributs que tenen una correlació millor que el 70% d'elles; és a dir, $cor_i \geq \mu + \sigma$ on μ és ma mitjana de les correlacions per al genere escollit i σ la desviació estandard.

A partir dels resultats obtinguts es pot deduïr que hi existeixen molt poques bones correlacions entre els atributs.

Per aquest motius és procedira a analitzar visualment les interacions dels atributs més rellevants entre ells i com afecten a l'etiquetatge.

S'observa, mitjançant les dades obtingudes[6], com hi han atributs rellevants per a cert gèneres que en altres no hi son pas.

Per aquest motius és pot teoritzar que és pot parametritzar diferents models de predicció per a cada gènere i, per tant, podem crear un model que categoritci les produccions mitjançant un conjunt de models diferents per a cada gènere.

Mencionar, a més, que s'obté atributs que no tenen correlacions significatives, això és per la falta de dades de produccions d'aquell tipus; per aquest motius és decideix no treballar amb aquestes dades ja que no s'obtindrà una classificació decent al no tenir la suficient informació com per a testejar i entrenar debidament (i la creació de noves dades mitjançant les que ja és disposa és desestimat degut a que tampoc és disposa amb dades suficients com per a crear un model de regressió amb bones prediccions capaz de generar noves dades ficticies).

Utilitzarem doncs aquesta idea com a primer pas en la cerca del nostre millor model

[6]És pot consultar la totalitat de les correlacions rellevants en el notebook entregat conjuntamnet amb la memória

2.4 Anàlisi dels atributs rellevants

Analitzant en més detall les distribucions dels atributs rellevants trobats en l'analisis de les correlacions de variables.
Per l'analisi més exaustiu representem de manera gràfica en $\mathbb{R}^2$ i $\mathbb{R}^3$ els atributs més rellevants de cada variable objectiu per trobar així relacions menys evidents:

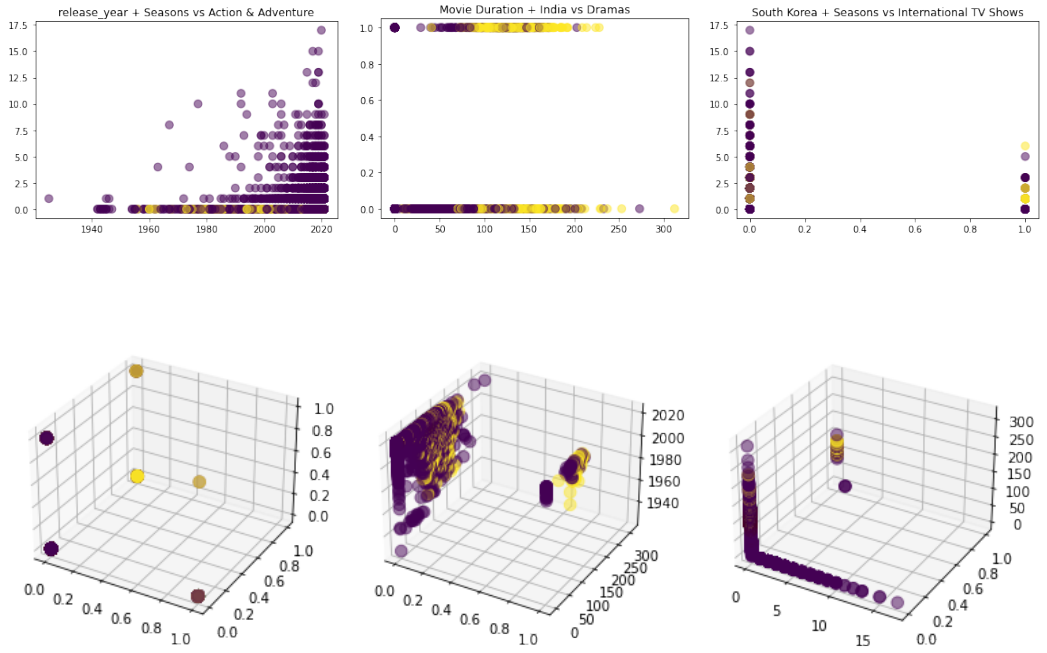


Figura 3: Mostra dels *scatterplots* en $\mathbb{R}^2$ i $\mathbb{R}^3$ generats per l'analisis exahustiu

A partir de les imatges[7] generades[8] es pot deduïr la no existencia de relacions dels atributs amb la variable objectiu.

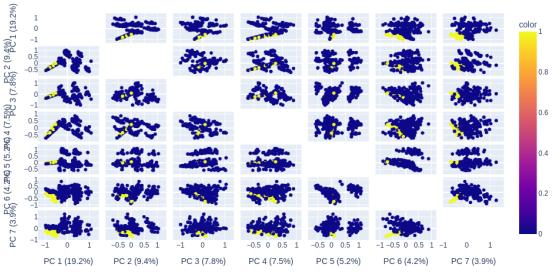


Figura 4: PCA de la variable objectiu *Comedies TV* amb els seus atributs més rellevants

Per tal de poder obtenir un millor enteniment de les distribucions de les daes i les relacions entre elles i la variable objectiu, és procedeix a fer una PCA de les variables objectius respecte als atributs més rellevants que hi corresponen.
En les imatges obtingudes com la PCA no obté resultats resenyables o minimament decents per a obtenir una bona classificació.
És procedeix a estudiar les interacions de diversos classificadors amb les variables objectius.

[7] La resta d'imatges es podenc onsultar al notebook entregat conjuntament amb la memória
[8] A la imatge es representa la variable objectiu com el color (lila o groc, no pertany o pertany respectivament) i els atributs utilitzats com a eixos

3 Classificació del dataset

3.1 Logistic Regressor

3.1.1 Estudi de la precisió

S'inicia l'estudi observant la precisió del model logístic amb paràmetres estàndard.
El resultat d'aplicar un regressor logístic (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

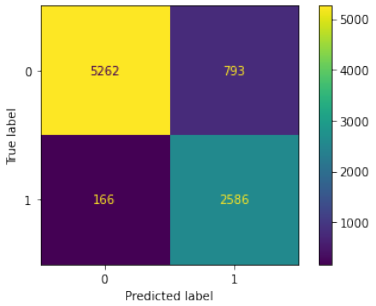


Figura 5: *Confusion matrix* del model logístic estàndard per *International Movie*

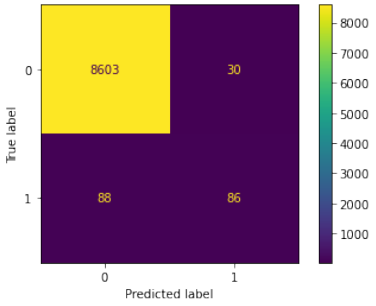


Figura 6: *Confusion matrix* del model logístic estàndard per *Spanish-Language TV Shows*

Gènere	**Atribut**	**Precissió**
International TV	Taiwan United States Movie Duration	0.945
Dramas	Movie Duration India	0.447
International Movie	India United States Seasons	0.945
British TV	United Kingdom Movie Duration Seasons	0.889
Spanish-Language	Spain Movie Duration Mexico	0.494
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 4: Mostra de la recall de les variables objectius mitjançant un classificador logístic estandard

Es representen en una taula els resultats obtinguts pel classificador Logístic per aquelles variables objectius que no obté una precissió superior al 30%.

És conclou que el classificador logístic té, inicialment, bona capacitat de classificació. Pel que és recurreix a fer un cross validation per tal de corrovorar els resultats obtinguts.

Mencionar que les matrius de confució mostrades són un exemple de les obtingudes en la cerca del millors atributs per a classificar la variable objectiu mencionada com a peu d'imatge.

3.1.2 Cross Validation

Al aplicar un *cross validation* (amb 10 subdivisions del dataset) s'obté les següents precissions (mostrem només les que tenen una precissió major a 30%):

Gènere	**Atribut**	**Precissió**
International TV	United States Movie Duration	0.945
Dramas	Movie Duration India	0.447
International Movie	United States Seasons	0.94
British TV	United Kingdom Movie Duration	0.889
Spanish-Language	Spain Movie Duration Mexico	0.492
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 5: Mostra de la recall de les variables objectius mitjançant un CrossValidation

És conclou que el classificador logístic permet de manera molt eficient classificar algunes variables objectiu; ja que tot i aplicar un CrossValidation amb 10 Kfolds obté molt bona recall i no cambia gairé be res respecte a la recall obtinguda amb totes les dades, no ésta tenint overfitting.

És prosegueix l'analisis dels classificadors provant altres models per millorar les prediccions del model logístic o obtenir més resultats eficients que el model logístic.

3.2 K-Nearest Neighbors

3.2.1 Estudi de la precisió

Es continua l'estudi observant la precisió del model KNN amb paràmetres estàndards, ja que al tractar-se d'un dataset amb dades molt compactes podria obtenir bons resultats. El resultat d'aplicar aquest model (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

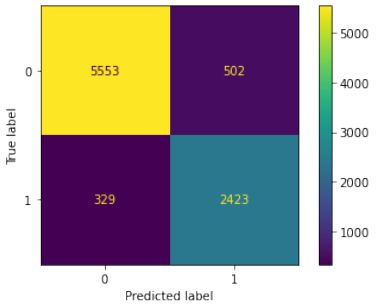


Figura 7: *Confusion matrix* del model K-Nearest Neighbour estàndard (K = 5) per *International Movie*

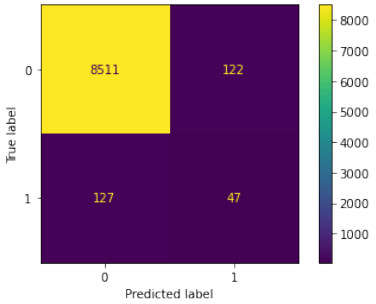


Figura 8: *Confusion matrix* del model K-Nearest Neighbour estàndard (K = 5)per *Spanish-Language TV Shows*

Gènere	**Atribut**	**Precissió**
TV Dramas	Seasons Movie Duration	0.7
Romantic TV	Seasons South Korea Taiwan	0.33
TV Comedies	Seasons Movie Duration Taiwan	0.349
Dramas	Movie Duration India Seasons	0.548
International Movie	India United States Seasons	0.945
British TV	United Kingdom Movie Duration Seasons	0.889
Spanish-Language	Spain Colombia Mexico	0.632
Classic Movies	Released Year	0.491

Taula 7: Mostra de la recall de les variables objectius mitjançant un classificador KNN estandard

Es representen en una taula els resultats obtinguts pel classificador KNN per aquelles variables objectius que no obté una precissió superior al 30%.

És conclou que el classcadorifi KNN té, inicialment, bona capacitat de classificació. Pel que és recurreix a fer un cross validation per tal de corrovorar els resultats obtinguts.

A més, s'observa com té variables objectius que classifica de manera decent diferents a les obtingudes amb el regressor logístic, pel que és disposa a cambiar la dinàmica (de fer un model diferents per variable objectiu) a fer un classificador diferents per cada model (segons classifiqui millor).

Mencionar que les matrius de confució mostrades, al igual que en el regressor logístic, són un exemple de les obtingudes en la cerca del millors atributs per a classificar la variable objectiu mencionada com a peu d'imatge.
En futures explicacions de models son mostrades també exemples trobats.

3.2.2 Cross Validation

Al aplicar un *cross validation* (amb 10 subdivisions del dataset) s'obté les següents precissions (mostrem només les que tenen una precissió major a 30%):

Gènere	**Atribut**	**Precissió**
International TV	United States India Movie Duration	0.945
Dramas	Movie Duration Seasons	0.5
International Movie	United States Seasons	0.94
British TV	United Kingdom Movie Duration	0.889
Spanish-Language	Spain Movie Duration Mexico	0.487
Classic Movies	Released Year	0.38
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 8: Mostra de la recall de les variables objectius mitjançant un CrossValidation

És conclou que el classificador KNN permet de manera molt eficient classificar algunes variables objectiu; ja que tot i aplicar un CrossValidation amb 10 Kfolds obté molt bona recall.

Tot i així, s'observa com cambia la recall respecte obtinguda amb totes les dades en algunes variables de manera significativa, pel que pot estar tenint overfitting; és tindrà cura aquesta circunstacia en futurs estudis més exahustius del mètode.

3.3 Support Vector Classification

3.3.1 Estudi de la precisió

Es continua l'estudi observant la precisió del model SVC amb paràmetres estàndards, ja que al tractar-se d'un dataset que ha funcionat de manera decent amb un classificador logístic podria obtenir bons resultats.
El resultat d'aplicar aquest model (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

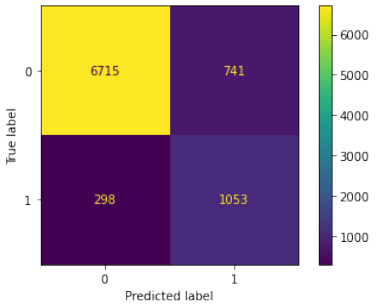


Figura 9: *Confusion matrix* del model SVC per *International TV Show*

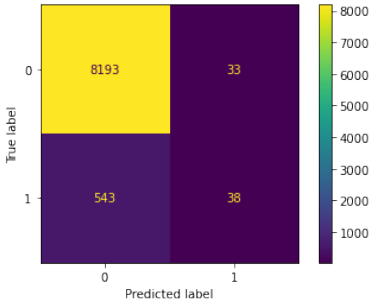


Figura 10: *Confusion matrix* del model SVC per *Comedies TV Shows*

Gènere	**Atribut**	**Precissió**
International TV	Taiwan United States Movie Duration	0.945
Dramas	Movie Duration Seasons	0.437
International Movie	India United States Seasons	0.94
British TV	United Kingdom Movie Duration Seasons	0.889
Spanish-Language	Spain Movie Duration Mexico	0.483
Classic Movies	Released Year	0.397
Anime Series	Movie Duration Japan	0,813
Korean TV	Movie Duration South Korea	0,874

Taula 11: Mostra de la recall de les variables objectius mitjançant un classificador SVC estandard

Es representen en una taula els resultats obtinguts pel classificador SVC per aquelles variables objectius que no obté una precissió superior al 30%.

És conclou que el classificador SVC té, inicialment, bona capacitat de classificació. Pel que és recurreix a fer un cross validation per tal de corrovorar els resultats obtinguts.

Cal descatar com s'intueix (mitjançant totes les dades per ara analitzades) que hi existeix una tendencia, tant en la recall com en els atributs) en els models a l'hora de classificar els gèneres. Aquesta tendencia a tenir valors semblant indica que, efectivament, les variables que s'estàn utilitzant són les que millor hi representen la classe; tot i que també reflexa un cert límit que podem assolïr amb elles i, a menys que en futurs classificadors utilitcin altres o classifiquin de millor manera que fins els ara vist, per tant pot haber variables objectius que siguin incapaces de ser classificades de forma decent per la falta de dades o d'informació.

3.3.2 Cross Validation

Al aplicar un *cross validation* (amb 10 subdivisions del dataset) s'obté les següents precissions (mostrem només les que tenen una precissió major a 30%):

Gènere	**Atribut**	**Precissió**
International TV	United States Movie Duration	0.945
Dramas	Movie Duration Seasons	0.443
International Movie	United States Seasons	0.94
British TV	United Kingdom Movie Duration	0.889
Spanish-Language	Spain Movie Duration Mexico	0.481
Classic Movies	Released Year	0.381
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 12: Mostra de la recall de les variables objectius mitjançant un CrossValidation

És conclou que el classificador SVC permet de manera molt eficient classificar algunes variables objectiu; ja que tot i aplicar un CrossValidation amb 10 Kfolds obté molt bona recall.

Cal esmentar com, es pot conjeturar l'hipotesis abans proposada que hi existeix un cert límit que hi podem arribar a obtenir com a recall per als atributs mitjançant les dades que hi disposem.
Per tal d'intentar evitar aquest cas i superar aquest hipotétic umbral, és procedeix a probar models de classificació basants en altres tecniques de classificació.

3.4 Decision Tree Classificator

3.4.1 Estudi de la precisió

Es continua l'estudi observant la precisió del model Decission Tree Classifier amb paràmetres estàndards, ja que al tractar-se d'un dataset que ha funcionat de manera decent amb un classificador logístic podria obtenir bons resultats.
El resultat d'aplicar aquest model (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

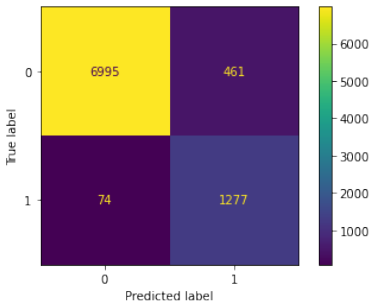


Figura 11: *Confusion matrix* del model Decission Tree per *International TV Show*

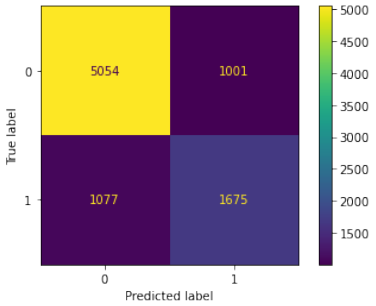


Figura 12: *Confusion matrix* del model Decission Tree per *International Movies*

Gènere	**Atribut**	**Precissió**
International TV	Taiwan United States Movie Duration	0.945
Romantic TV	Taiwan South Korea Seasons	0.332
Dramas	Movie Duration Seasons	0.55
International Movie	India United States Seasons	0.94
British TV	United Kingdom Movie Duration Seasons	0.889
Spanish-Language	Spain Movie Duration Mexico	0.494
Classic Movies	Released Year	0.457
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 11: Mostra de la recall de les variables objectius mitjançant un classificador Decission Tree estandard

Es representen en una taula els resultats obtinguts pel classificador Decission Tree per aquelles variables objectius que no obté una precissió superior al 30%.

És conclou que el classificador Decission Tree té, inicialment, bona capacitat de classificació. Pel que és recurreix a fer un cross validation per tal de corrovorar els resultats obtinguts.

3.4.2 Cross Validation

Al aplicar un *cross validation* (amb 10 subdivisions del dataset) s'obté les següents precissions (mostrem només les que tenen una precissió major a 30%):

Gènere	**Atribut**	**Precissió**
International TV	United States Movie Duration	0.945
Dramas	Movie Duration Seasons	0.498
International Movie	United States Seasons	0.94
British TV	United Kingdom Movie Duration	0.889
Spanish-Language	Spain Movie Duration Mexico	0.487
Classic Movies	Released Year	0.38
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 12: Mostra de la recall de les variables objectius mitjançant un CrossValidation

És conclou que el classificador Decission Tree permet de manera molt eficient classificar algunes variables objectiu; ja que tot i aplicar un CrossValidation amb 10 Kfolds obté molt bona recall.

També mencionar que no ha pogut predir de manera més eficient que la resta de classificadors probats fins ara els atributs que menys recall obtenen, pel que és probarà mitjançant un ensemble de Decission Trees dur a terme aquesta tasca.

3.5 Random Forest Classificator

3.5.1 Estudi de la precisió

Es continua l'estudi observant la precisió del model Random Forest Classifier amb paràmetres estàndards, ja que al tractar-se d'un dataset que ha funcionat de manera decent amb un classificador Decission Tree, podria obtenir bons resultats.
El resultat d'aplicar aquest model (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

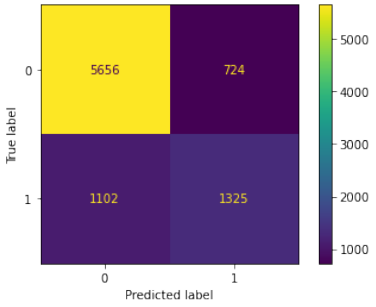


Figura 13: *Confusion matrix* del model Random Forest per *Dramas*

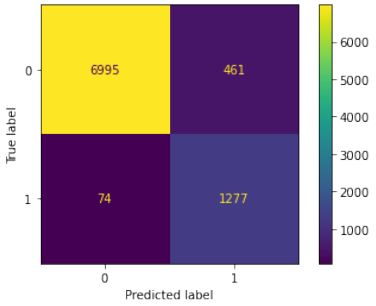


Figura 14: *Confusion matrix* del model Random Forest per *International TV Show*

Gènere	**Atribut**	**Precissió**
International TV	Taiwan United States Movie Duration	0.945
Romantic TV	Taiwan South Korea Seasons	0.332
Dramas	Movie Duration Seasons	0.592
International Movie	India United States Seasons	0.94
British TV	United Kingdom Movie Duration Seasons	0.889
Spanish-Language	Spain Movie Duration Mexico	0.494
Classic Movies	Released Year	0.422
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 13: Mostra de la recall de les variables objectius mitjançant un classificador Random Forest estandard

Es representen en una taula els resultats obtinguts pel classificador Decission Tree per aquelles variables objectius que no obté una precissió superior al 30%.

És conclou que el regressor Random Forest té, inicialment, bona capacitat de classificació. Pel que és recurreix a fer un cross validation per tal de corrovorar els resultats obtinguts.

3.5.2 Cross Validation

Al aplicar un *cross validation* (amb 10 subdivisions del dataset) s'obté les següents precissions (mostrem només les que tenen una precissió major a 30%):

Gènere	**Atribut**	**Precissió**
International TV	United States Movie Duration	0.945
Dramas	Movie Duration Seasons	0.503
International Movie	United States Seasons	0.94
British TV	United Kingdom Movie Duration	0.889
Spanish-Language	Spain Movie Duration Mexico	0.487
Classic Movies	Released Year	0.414
Anime Series	Movie Duration Japan	0.813
Korean TV	Movie Duration South Korea	0.874

Taula 14: Mostra de la recall de les variables objectius mitjançant un CrossValidation

És conclou que el classificador Random Forest permet de manera molt eficient classificar algunes variables objectiu; ja que tot i aplicar un CrossValidation amb 10 Kfolds obté molt bona recall.

També mencionar que no ha pogut dur a terme la tasca de obtenir bones prediccinos per a altres atributs als plasmats a les taules durant l'analisis dels models vist fins ara, pel que es recorrirà a altres mètodes de classificació.

3.6 Kmeans

3.6.1 Estudi de la precisió

Es continua l'estudi observant la precisió del model K-means Classifier amb paràmetres estàndards.
El resultat d'aplicar aquest model (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

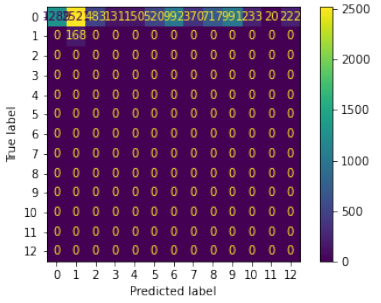


Figura 15: *Confusion matrix* del model K-means per *Action & Adventures*

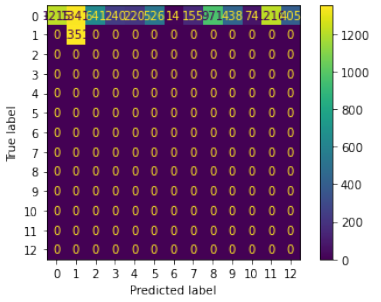


Figura 16: *Confusion matrix* del model K-means per *International TV Show*

Gènere	**Atribut**	**Precissió**
Documentaries	Seasons	1.00
Dramas	India Seasons	0.727
International Movie	Seasons United States	0.94
Action & Adventure	China Hong Kong Seasons	0.888
Anime Series	Japan	0.859

Taula 15: Mostra de la recall de les variables objectius mitjançant un classificador K-means estandard

Es representen en una taula els resultats obtinguts pel classificador Decission Tree per aquelles variables objectius que no obté una precissió superior al 30%.

S'observa de forma evident que el model pateix d'un fort overfitting ja que només és capaç de classificar 13 classes (ja que és el paràmetre estandard del K-means).

A més, observant la matriu de confusió, és plasma de manera evident la poca capacitat de classificació eficient que té el model.

És conclou que el classificador K-means té, inicialment, bona capacitat de classificació peró no amb el paràmetre estandard.

No és procedirà amb el CrossValidation i, és retornarà a estudiar el cas més endabant en la cerca de hiperparàmetres per si consegueix així corretgir el comportament.

És procedeix doncs amb un altre model similar al K-means per a intentar obtenir resultats millors amb aquest sistema de classificació.

3.7 Gaussian Mixture

3.7.1 Estudi de la precisió

Es continua l'estudi observant la precisió del model Gaussian Mixture Classifier amb paràmetres estàndards.
El resultat d'aplicar aquest model (per a cadascuna de les variables objectius) per a classificar el dataset (sense utilitzar part del mateix per a validar) és:

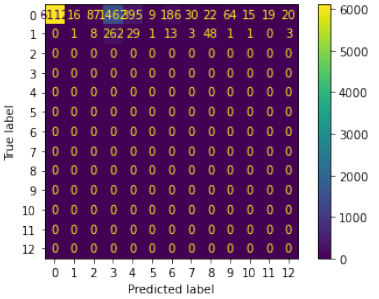


Figura 17: *Confusion matrix* del model Gaussian Mixture per *Romantic TV Show*

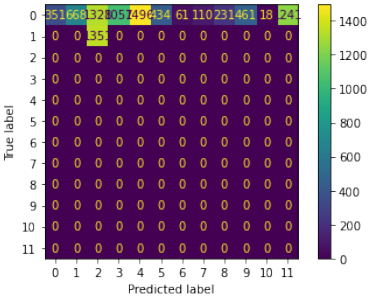


Figura 18: *Confusion matrix* del model Gaussian Mixture per *International TV Show*

Gènere	Atribut	Precissió
International TV	Japan United States	0.840
International Movie	India United States	0.631
Anime Series	Japan	0.859

Taula 17: Mostra de la recall de les variables objectius mitjançant un classificador Gaussian Mixture estandard

Es representen en una taula els resultats obtinguts pel classificador Gaussian Mixture per aquelles variables objectius que no obté una precissió superior al 30%.

S'observa de forma evident que el model pateix d'un fort overfitting i d'incapacitat per a classificar de manera adient.

També es pot observar aquesta poca capacitat de classificació eficient que té el model mitjançant les matrius de confució.

És conclou que el classificador Gaussian Mixture té, inicialment, bona capacitat de classificació peró no amb el paràmetre estandard.

No és procedirà amb el CrossValidation i, és retornarà a estudiar el cas més endabant en la cerca de hiperparàmetres per si consegueix així corretgir el comportament.

És procedeix doncs a probar altres mètodes més sofisticats per classificar.

3.8 Models de xarxes neural

3.8.1 Estudi de la precisió

Es continua l'estudi observant la precisió de diversos models mitjançant xarxes neural deferents.
S'ha definit 4 models diferents que tenen com a diferencia unica el *hidden layers*.
És mostre ara els resultats de cada model:

Gènere	**Model**	**Precissió**
International TV	4	0.926
Dramas	3	0.567
International Movie	3	0.913
British TV	2	0.868
Anime Features	3	0.809
Anime Series	3	0.887
Korean TV	3	0.689
Stand-Up Comedy	3	0.34

Taula 18: Mostra de la recall de les variables objectius mitjançant xarxes neurals

Es representen en una taula els resultats obtinguts pels models de xarxes neurals proposats per aquelles variables objectius que obté una precissió superior al 30%.

S'observa que el model no té una bona precissió (o almenys en comparació a altres models ja analitzats); tot i així, obté bons resultats en atributs que no han estat classificats de forma adient fins ara.

Mencionar, que els models proposats han estat creats de forma intuïtiva i que, per tant, és pot refinar la seva estructura per a obtenir millors classificacions.

És conclou que la xarxa neural model 3 té, inicialment, bona capacitat de classificació per certs atributs fins ara no classificats de forma adient.

És procedirà a estudiar les xarxes neurals amb estructura similar a la del model 3 per tal de trobar altres atributs que puguin ser classificables amb aquests mètodes de classificació.

Model	**Estructura/distribució en la *hidden layer***
Model 1	1 capa de 5 nodes
Model 2	2 capes de 10 nodes cadascuna
Model 3	2 capes de 20 nodes cadascuna
Model 4	3 capes de 50 nodes cadascuna

Taula 19: Models proposats com a xarxes neurals de classificació

3.9 Estudi del model de xarxa neural 3

És procedeix a fer un estudi de la recall del model de xarxa neural 3 amb diferents valors de *hidden layer*.
L'estructura de dues capes és manté peró és variarà el nombre de nodes de cadascuna amb valors que oscilin entre 15 i 31. Els resultat d'aquest estudi son:

Gènere	**Nodes**	**Precissió**
International TV	20	0.936
Romantic TV	24	0.307
Dramas	23	0.582
International Movie	31	0.918
British TV	22	0.868
Spanish-Language TV	29	0.596
Anime Features	20	0.809
Anime Series	16	0.886
Korean TV	24	0.889
Stand-Up Comedy	24	0.408

Taula 20: Mostra de la recall de les variables objectius mitjançant xarxes neurals de model 3

Es representen en una taula els resultats obtinguts pels difersos models de tipus 3 de xarxa neural proposat anteriorment per aquelles variables objectius que obté una precissió superior al 30%.

S'observa clarament com el model obté precissions similars a les obtingudes per altres models més sencills plantejats anteriorment.

Per altra banda, també s'observa com el model amb 20 nodes és capaz de classificar de manera molt eficaç el gènere *Anime Features* que fins ara no ha sigut classificat de manera eficient per cap classificador.

En l'estudi s'ha dividit el dataset en train i test per així evitar que el model memorítzi o pateixi d'overfitting.

Com que ha obtingut bons resultats el model, es proba a utilitzar els models abans testejats d'una manera més complexa per probar si, també, poden servir per a classificar altres gèneres.

3.10 Estudi del model de xarxa neural 1

Recuperant l'idea abans esmentada de crear un classificador que aprofiti les dades que ha obtingut altres classificador, s'aprofita els models de xarxes neurals descartats anteriorment per a testejar si tenen algun comportament que, amb les noves dades obtingudes mitjançant els classificadors que ja disposem, pot classificar altres gèneres.

El model que millor comportament ha tingut dels abans esmentats és el model de xarxa neural 1.

L'unic resultat resenyable és el de la variable objectiu *TV Dramas* que obté amb el model 1 una recall de 0.593

Tot i no haver-se obtingut millors ćlassificacions i més classificacions de variables objectiu, s'ha obtingut una certa millora al poder ara classificar mitjançant models de xarxes neurals 1 un gènere fins ara no classificat decentment mai.

4 Plantejament del classificador

Mitjançant totes les dades obtingudes, és conclou utilitzar el millor classificador per a cada variable i, a l'hora de classificar, ajuntar les classificacions fetes per cada classificador.

L'estudi del millor classificador obté els següents resultats:

Variable	**Classificador**	**Atributs**	**Precissió**
International TV	Logístic	Movie Duration & U.S.A.	0.94
Dramas	Random Forest	Movie Duration & Seasons	0.56
International Movies	Logístic	U.S.A. & Seasons	0.94
British TV	Logístic	Movie Duration & U.K.	0.89
Spanish-Lang. TV	Logístic	Spain & Mexico & Movie Duration	0.51
Anime Series	Logístic	Japan & Movie Duration	0.82
Korean TV	Logístic	South Korea & Movie Duration	0.87
Anime Features	Neural Network	*ALL*	0.81
TV Dramas	Neural Network	*Classified types*	0.59

Taula 21: Millor combinació d'atributs per al millor classificador de cada variable

Mencionar que, en la taula, *ALL* fa referencia a utilitzar tot els atributs i *Classified types* fa referencia a les etiquetes proposades pels classificadors anterior a ella; per aquest motiu és separa la variable *TV Dramas* de la resta, perque precissa de la classificació de les altres variables per classificar-la.

S'bserva com, tot i tenir poques correlacions rellevants per classificar, els classificadors de manera estandard són capaços d'obtenir una bona classificació en alguns atributs que, a més, són els que més dades tenen.
Per aquest motiu, és confirma la sospita que era casi impossible predir de forma eficient certes variables objectius per falta de dades.

És limitarà doncs a augmentar la precissió dels mètodes que hi disposem mitjançant una cerca d'hiperparàmetres per tal de trobar aquells que fan que la precissió sigui la més alta possible.

Per altra banda, surgeix l'idea de utilitzar el mateix concepte que amb la xarxa neural del model 1 amb la variable *TV Dramas* per tal de classificar altres variables de la següent manera:

- Utilitzar els classificador sencills per a predir els resultats més sencills.
- Utilitzar xarxes neurals especifiques per a etiquetar les variables més dificils de classificar.
- Utilitzar una xarxa neural que (iterada sobre ella mateixa K cops) rectifiqui els error d'etiquetatge i classifiqui millor les produccions.

Així doncs, l'idea de l'algorísme seria la següent:

1. Classificar mitjançant els classificadors més sencills les variables més facils de classificar
2. Afegir les classificacions a una llista d'etiquetes
3. Mentres quedin variables sense etiquetar:
   (a) Classificar mitjançant una xarxa neural una variable no etiquetada amb totes les dades que s'hi disposen i la llista d'etiquetes obtingudes.
   (b) afegir la classificació a la llista d'etiquetes
4. Durant K-iteracions:
   (a) Utilitzar una xarxa neural que amb totes les classificacions fetes i totes les dades reclassifiqui les etiquetes.
   (b) Substituïr la llista d'etiquetes inicial per la obtinguda en la xarxa neural.
5. Classificar la producció com la llista d'etiquetes obtinguda

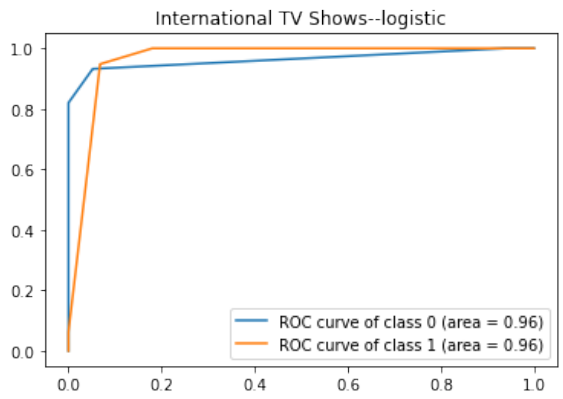
Aquesta idea d'algorisme no ha sigut testejada per falta de temps i no és pot assegurar que funcioni.

Tot i així, s'hagues intentat testejar aquest l'algorisme si s'hagues disposat de més temps.

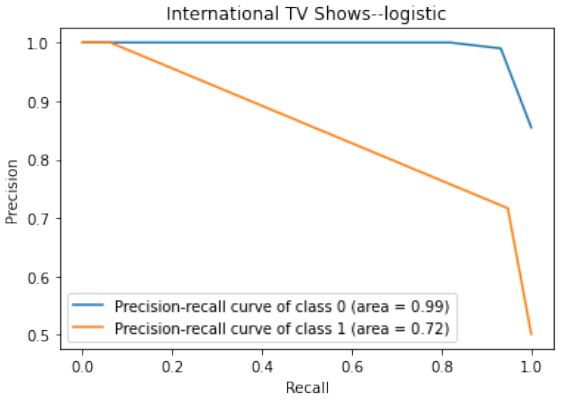
És procedeix doncs, a estudiar les curves dels mètodes de classificació que hi disposem abans de fer la cerca d'hiperparàmetres per tal de, una vegada duta a terme la cerca, comparar-les.

5 Estudi de les curves Precission-Recall i ROC dels models

5.1 Precission-Recall i ROC curve dels models Logístics



(a) *ROC curve*

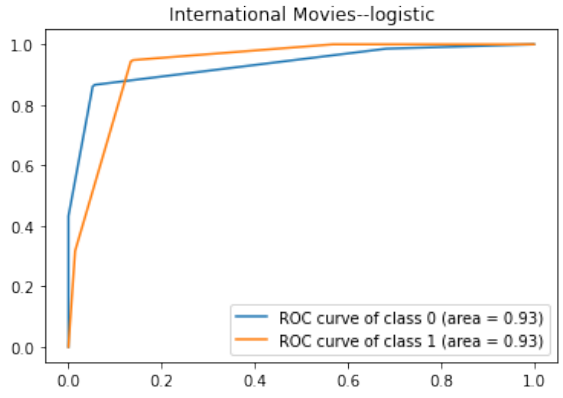


(b) *Precission-Recall curve*

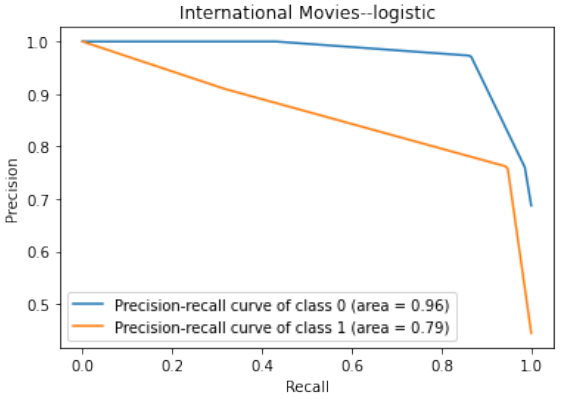
Figura 19: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *International TV Shows*

S'observa de les curves ROC que el model te una bona sensibilitat, peró també s'observa de les curves Precission-Recall com hi existeix una tendencia a no tenir molt bona recall precissió i com la recall de les classes decau de forma lineal fins a un valor proper al 70% en la classificació de la variable objectiu.

Tot i així te un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valor propers a 1.



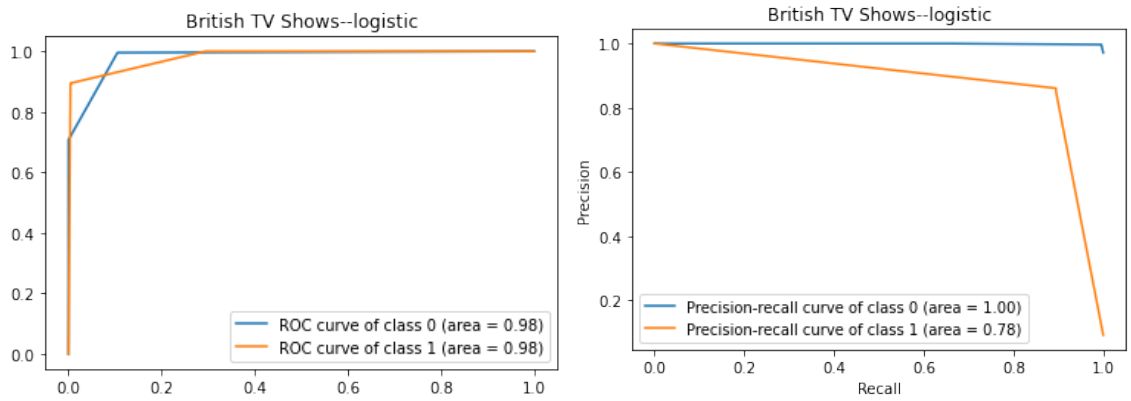
(a) *ROC curve*



(b) *Precission-Recall curve*

Figura 20: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *International Movies*

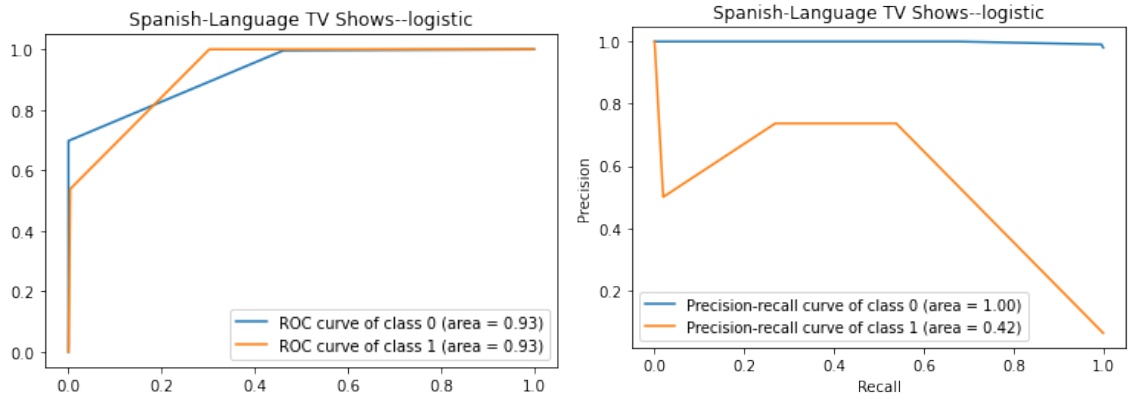
S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 21: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *British TV Shows*

S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.

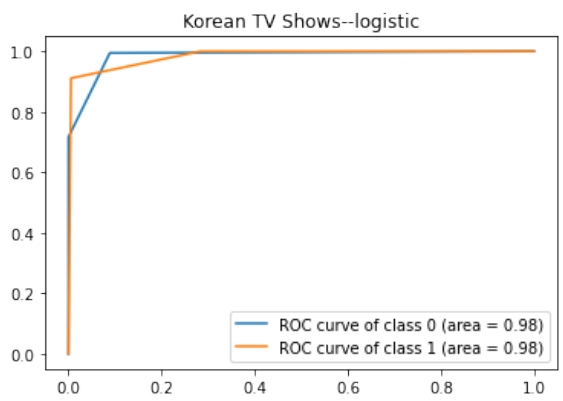


(a) *ROC curve* (b) *Precission-Recall curve*

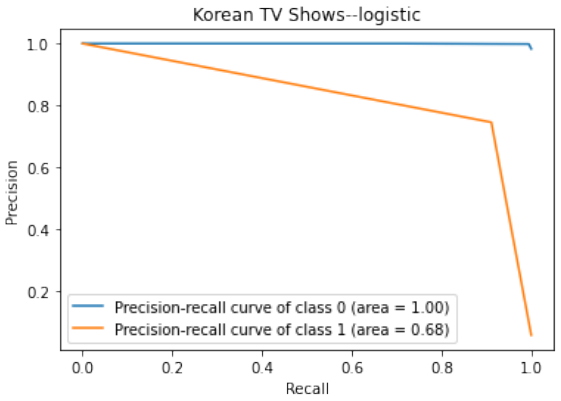
Figura 22: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *Spanish-Language TV Shows*

S'observa com el model de classificació logístic te problemes per distingir si una producció és o no de gènere *Spanish-Language TV Show*; tot i així, dona bons resultats quan s'utilitza i és basntant eficient.

Es conclou doncs que té un resultat bo pero millorable i amb la cerca d'hiperparàmetres és milloraran aquests resultats.



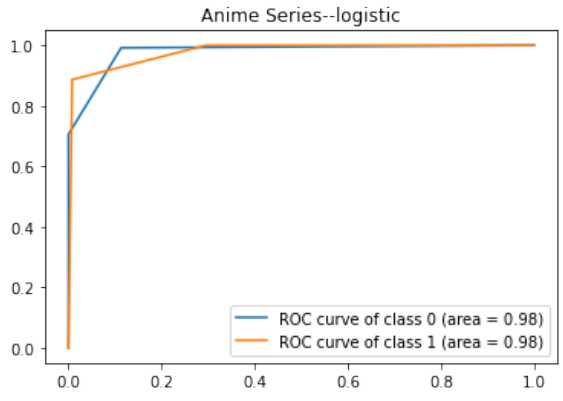
(a) *ROC curve*



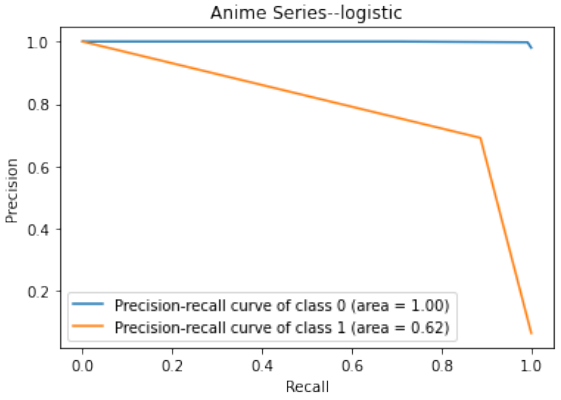
(b) *Precission-Recall curve*

Figura 23: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *Korean TV Shows*

S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.



(a) *ROC curve*

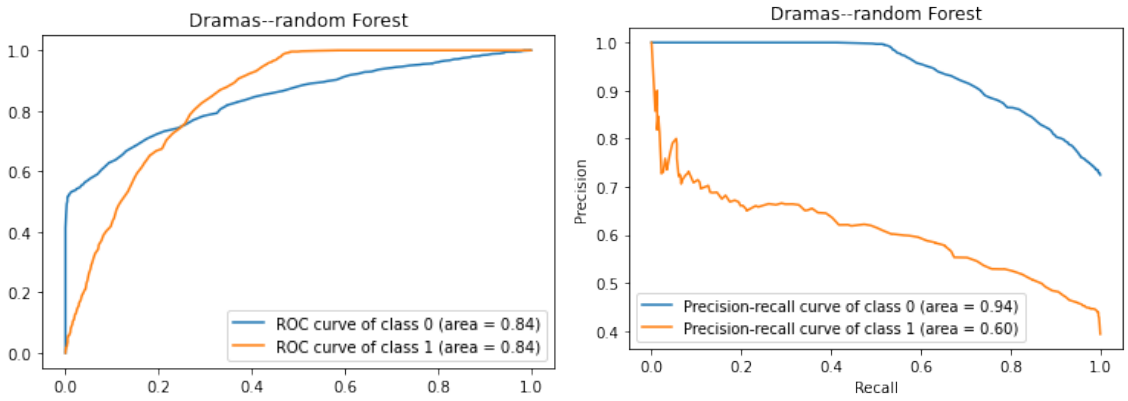


(b) *Precission-Recall curve*

Figura 24: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *Anime Series*

S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.

5.2 Precission-Recall i ROC curve del Random Forest



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 25: Corba *ROC* i *Precission-Recall* del Random Forest per la variable *Dramas*

S'observa, a partir de la corba ROC, com la sensitibitat del model és bona tot i que és pitjor que les que s'han obtingut en les variables objectius classificades amb els models logístic.

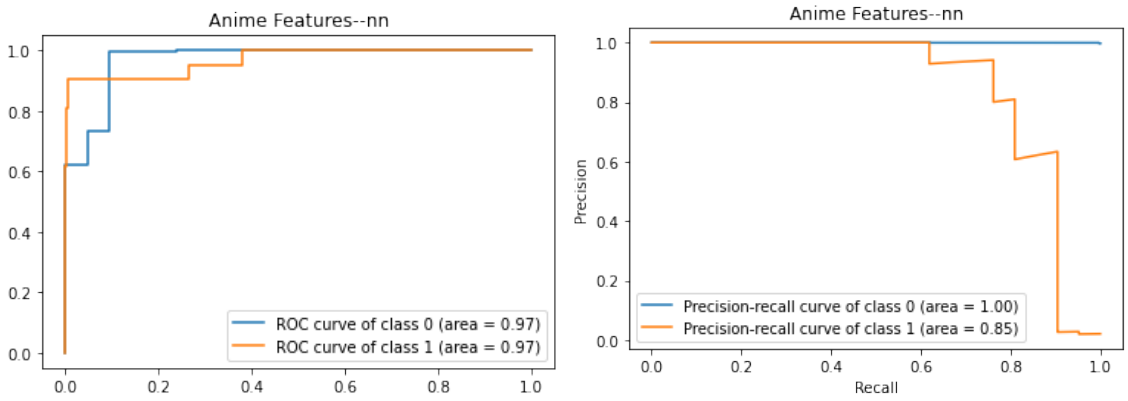
Per altra banda, preocupa la precissió del model, tal i com és pot observar en la corva Precission-Recall. El model tendeix a tenir problemes de precissió en la classificació de les produccions que tenen de gènere *Dramas*.

Aquest comportament pot ser degut a la falta de dades i seria mitigat si es pogués introduïrne de més.

Tot i així, el model obté bones prediccion i té un bon resultat d'area sota la corba Precission-Recall que, segurament, seran millorats amb la implementació del millors hiperparàmetres.

Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.

5.3 Precission-Recall i ROC curve de les Xarxes Neurals



(a) *ROC curve* (b) *Precission-Recall curve*

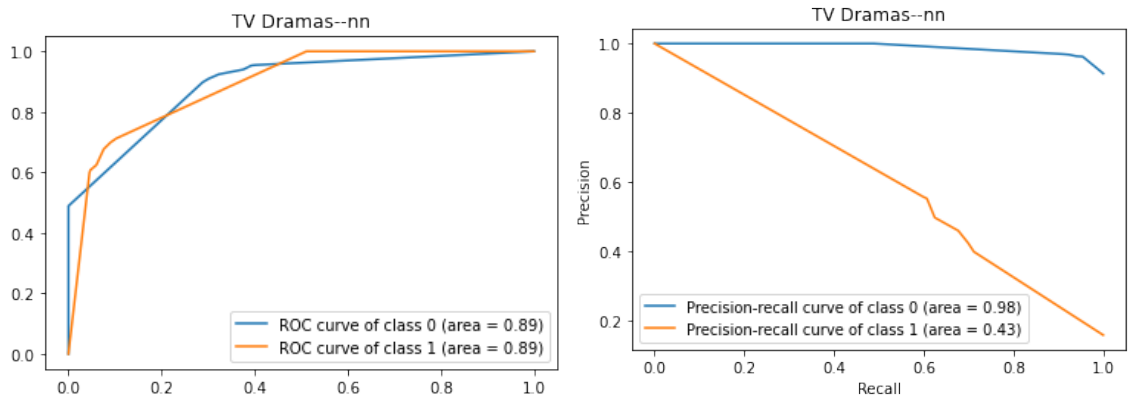
Figura 26: Corba *ROC* i *Precission-Recall* del model 3 de xarxa neural per la variable *Anime Features*

S'observa com el model de xarxa neural 3 és molt eficient a l'hora de classificar les produccions segons el gènere *Anime Features*.

Els valors obtinguts tant en la corba ROC com Precission-Recall son molt propers a 1 o iguals, pel que reafirma la seva bona capacitat de classificació.

Tot i així el classificador té problemes de precissió una vegada la recall supera el 0.8.

És Conclou que el model de xarxa neural 3 és un molt bon classificador.



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 27: Corba *ROC* i *Precission-Recall* del model 1 de xarxa neural per la variable *TV Dramas*

S'observa, com el model de xarxa neural 1 te problemes per classificar les produccions segons el gènere *TV Drama*.

Enfatisar que aquest problemes son degut segurament a la falta de fiabilitat de les dades, ja que es tracten de les obtingudes pels mètodes que, ja de per si, obtenen certs errors.

Per altra banda, s'observa de les corbes ROC com té una bona sensitibitat, peró analitzant les corbes Precission-Recall és conslou que el model classifica pitjor que una classificació aleatória.

Recalcar que, tot i semblar que el model classifica erroniament, a la practica dona molt bons resultats, hi els resultats que obtenim en les corbes de Precission-Recall segurament son degudas a que hi existeixen poques produccions a la base de dades que tinguin gènere *Drama*.

És conclou doncs que el model de xarxa neural 1 és bo tot i que s'espera millors resultats una vegada aplicada la cerca d'hiperparàmetres.

6 Cerca d'hiperparàmetres

6.1 Hiperparàmetres dels models

Exposem ara els millors hiperparàmetres per model i per atribut: Recarcar que, els parà-

Model	Atributs	Hiperparàmetres	Precissió
Logístic	International TV	C: K penalty: l2 solver: lbfgs random_state: 0	0.95
	International Movies		0.94
	British TV		0.89
	Spanish-Lan. TV		0.53
	Anime Series		0.87
	Korean TV		0.91
Random Forest	Dramas	n_estimators: 450 max_features: $log_2(n)$ criterion: entropy random_state: 0	0.52
Neural Network	Anime Features	activation: relu alpha: α_1 learnung_rate: constant learning_rate_init: L_1 solver: lbfgs random_state: 0 hidden_layer_sizes: (20, 20)	0.81
	TV Drama	activation: identity alpha: α_2 learnung_rate: invscaling learning_rate_init: L_2 solver: lbfgs random_state: 0 hidden_layer_sizes: (20,20)	0.60

Taula 22: Taula amb els millors paràmetres de cada model per cada variable objectiu i la seva recall

metres que associem a constants $K, \alpha_1, \alpha_2, L_1$ i L_2 és per la gran quantitat de decimals que contenen les constants trobades. Concretament, valen:

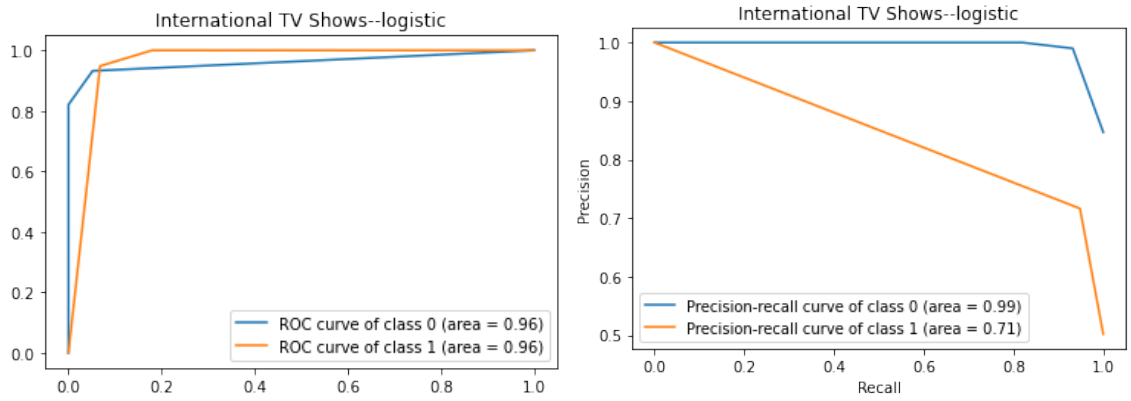
- $K = 2.5835764522666245$
- $\alpha_1 = 0.6458941130666561$
- $\alpha_2 = 0.02021839744032572$
- $L_1 = 0.2975346065444723$
- $L_2 = 0.7781567509498505$

Per altre banda, s'observa com la precissió augmenten/estabilitcen en valors més elevats que abans, pel que podem afirmar que la cerca d'hiperparàmetres ha funcionat correctament.

Analitcem ara les corbes ROC i Precission-Recall de cada model amb els hiperparàmetres.

6.2 Resultats d'aplicar els hiperparàmetres

6.2.1 Precission-Recall i ROC curve dels models Logístics



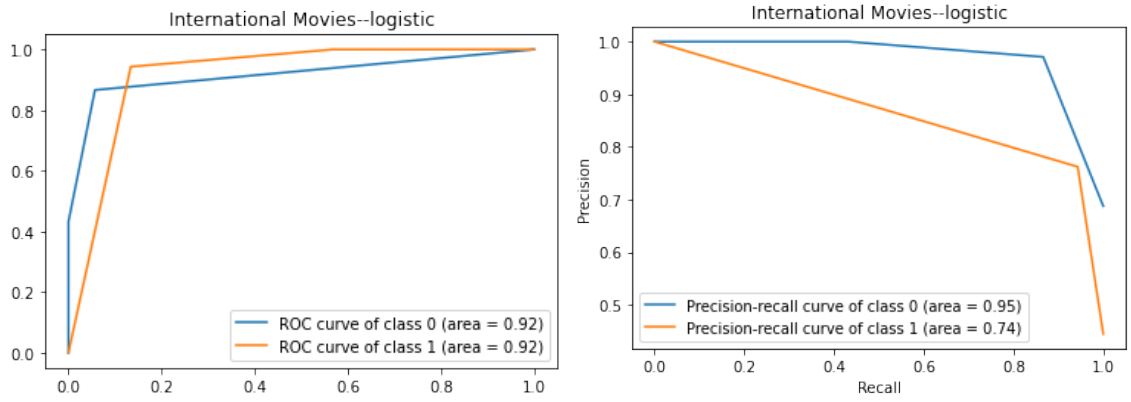
(a) *ROC curve* (b) *Precission-Recall curve*

Figura 28: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *International TV Shows*

S'observa de les curves ROC que el model te una bona sensibilitat.

A més, la tendencia a perde recall ara és menys pronunciada pero perd recall abans; fent així que perdi valor d'area respecte abans de la cerca d'hiperparàmetres.

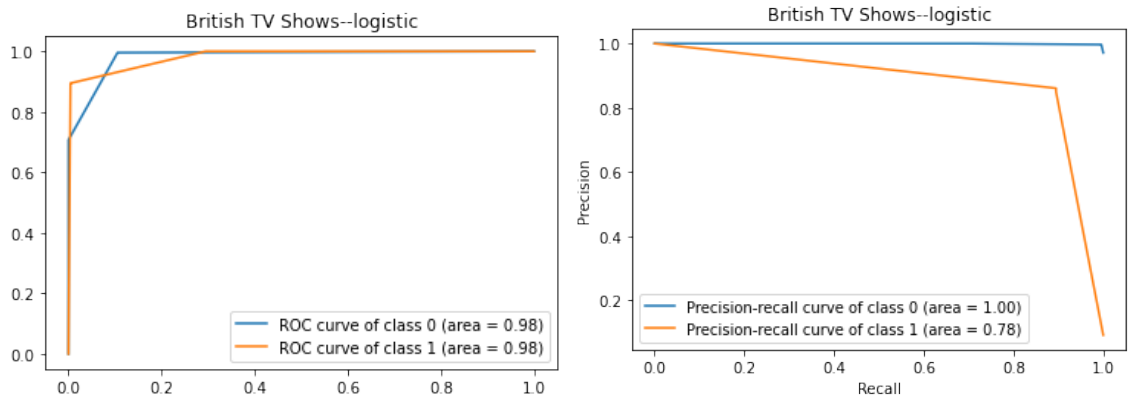
Tot i així te un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valor propers a 1.



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 29: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *International TV Shows*

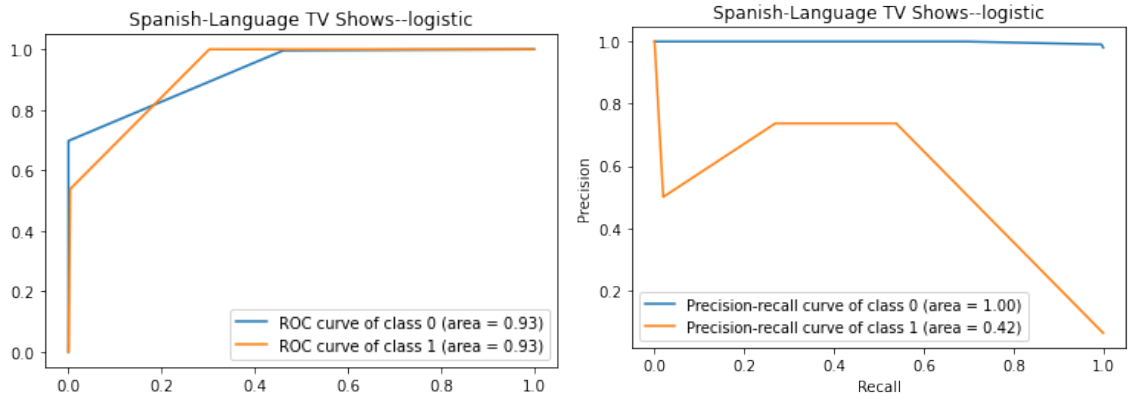
S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les corves obtenen valors propers a 1.



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 30: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *British TV Shows*

S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.
S'observa com el model de classificació logístic te problemes per distingir si una producció



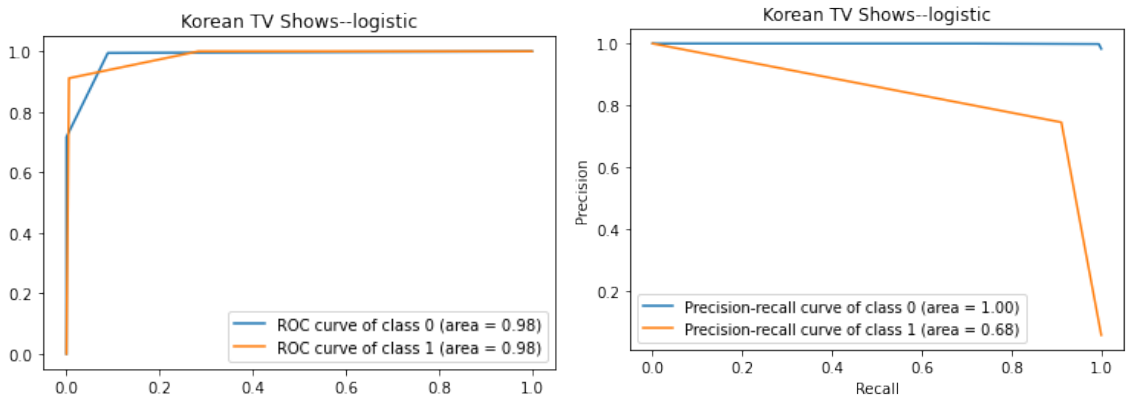
(a) *ROC curve* (b) *Precission-Recall curve*

Figura 31: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *Spanish-Language TV Shows*

és o no de gènere *Spanish-Language TV Show*; tot i així, dona bons resultats quan s'utilitza i és basntant eficient.

També afegir, que els cambis en la forma de la corba ROC i Precission-Recall han sigut infims.

Es conclou doncs que té un resultat bo peró s'esperaba una millora més significativa.

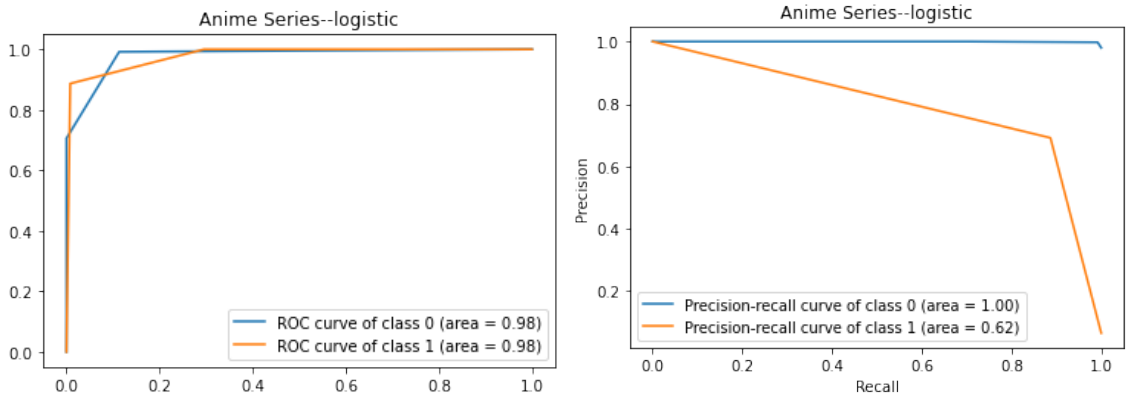


(a) *ROC curve* (b) *Precission-Recall curve*

Figura 32: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *Korean TV Shows*

S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.

S'observa de les corbes ROC i Precission-Recall un resultat similar al obtingut amb la



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 33: Corba *ROC* i *Precission-Recall* del model Logístic per la variable *Anime Series*

variable *International TV Show*.
Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.

Comentari General sobre el model Logístic:

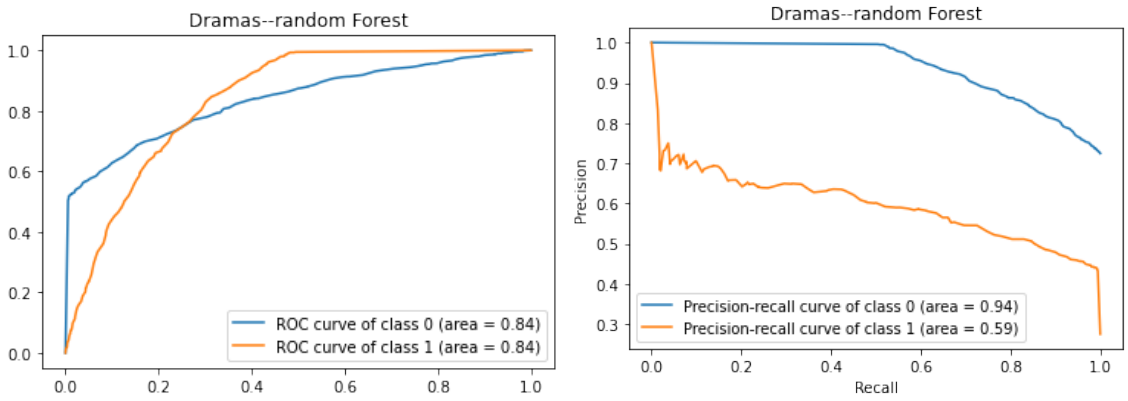
S'esperaba una millora més significativa en les corbes Precission-Recall una vegada optimitzats els hiperparàmetres.

Exposem les idees de que ha pogut succeir per a que obtingues les millores esperables:

- El regressor ja, una vegada optimitzat, pot augmentar la seva precissió a l'hora de classificar, pero les caracteristiques entre la Precission i la Recall és manté constant.
- El mètode utilitzat[9] per a trobar la millor combinació d'hiperparàmetres no és el millor i ha donat una combinació d'hiperparàmetres que millora la precissió només en alguns casos.

[9]S'ha utilitzat, degut a la gran quantitat d'hiperparàmetres a cecar, el mètode *RandomizedSearchCV* de la llibreria *sklearn*

6.2.2 Precission-Recall i ROC curve del Random Forest



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 34: Corba *ROC* i *Precission-Recall* del Random Forest per la variable *Dramas*

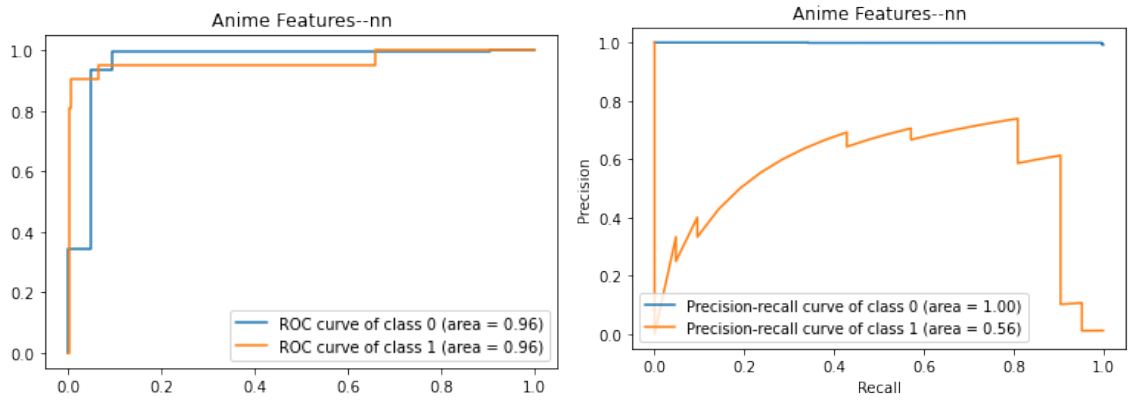
S'observa, a partir de la corba ROC, com la sensitibitat del model és bona tot i que és pitjor que les que s'han obtingut en les variables objectius classificades amb els models logístic.

Recarcar que, no ha patit grans cambis en la distribució/forma de les corbes ROC i Precission-Recall, pel que pot corroborar les hipótesis abans plantejades com a comentaris del model Logístic.

Tot i així, el model obté bones prediccion i té un bon resultat d'area sota la corba Precission-Recall.

Es conclou doncs que té un bon resultat ja que les dues corves (ROC i Precission-Recall) obtenen valors propers a 1.

6.2.3 Precission-Recall i ROC curve de les Xarxes Neurals



(a) *ROC curve* (b) *Precission-Recall curve*

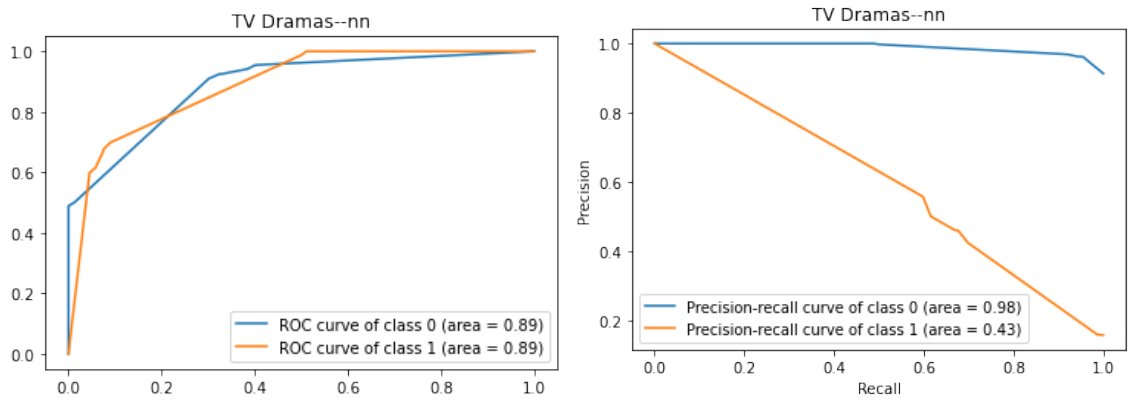
Figura 35: Corba *ROC* i *Precission-Recall* de la xarxa neural per la variable *Anime Features*

S'observa com, ara si, les corves han cambiat respecte les obtingudes anteriorment.

De les corbes ROC corroborem que el model de xarxes neurals és bo per a la classificació de l'atribut *Anime Features*. Per altra banda, les corbes de Precission-Recall indiquen que el classificador té problemes per d'obtenir bones Preccions i Recall a l'hora.

Aquesta diferencia de resultats respecte a les corbes Precission-Recall obtingudes pot ser deguda perque en aquest cas, s'ha repetir varies vegades l'experiment i ha mostrar que, si bé el model és bo classificant, no es tant bo com s'esperaba degut a possibles overfitting i processos de memorització del model.

És conclou que el model és prou bo.



(a) *ROC curve* (b) *Precission-Recall curve*

Figura 36: Corba *ROC* i *Precission-Recall* de la xarxa neural per la variable *TV Dramas*

S'observa, com el model de xarxa neural no ha variat respecte el ja obtingut anteriorment.

Aquest efecte segurament és degut a que, com ja hem explicat anteriorment, no és pot augmentar més la precissió del model degut a les poques dades i la dependencia que té respecte als resultats obtinguts pels altres classificadors.

Tot així obté bons resultats pel que és conclou que és un bon classificador.

7 Conclusions

Concluïm l'estudi mostrant les capacitats del classificador proposat, que esta format pels classificadors següents.

Model		**Temps d'execució (s)**		**Precission**	**Recall**	**Accuracy**
Clas.	Atribut	30% data	$25 \cdot 10^4$ prod.			
Logístic	International TV	0.01	0.02	0.72	0.95	0.82
	International Mov.	0.01	0.02	0.76	0.94	0.84
	British TV	0.01	0.02	0.86	0.89	0.88
	Spanish-Lan. TV	0.01	0.03	0.74	0.54	0.62
	Anime Series	0.01	0.02	0.69	0.89	0.78
	Korean TV	0.01	0.03	0.75	0.91	0.82
Rnd Forest	Dramas	1.11	36.85	0.59	0.52	0.55
NN	Anime Features	0.44	0.94	0.74	0.81	0.77
	TV Drama	0.95	0.9	0.56	0.6	0.58

Recalcar, finalemnt, que aquest classificador (tot i ser el millor testejat) no pot classificar tot els 31 gèneres que disposa la base de dades degut a que no disposem de suficients dades de la resta de gèneres. Per a poder obtenir un millor classificador s'hauria d'introduïr més exemples de produccions de la resta de gèneres no classificats.