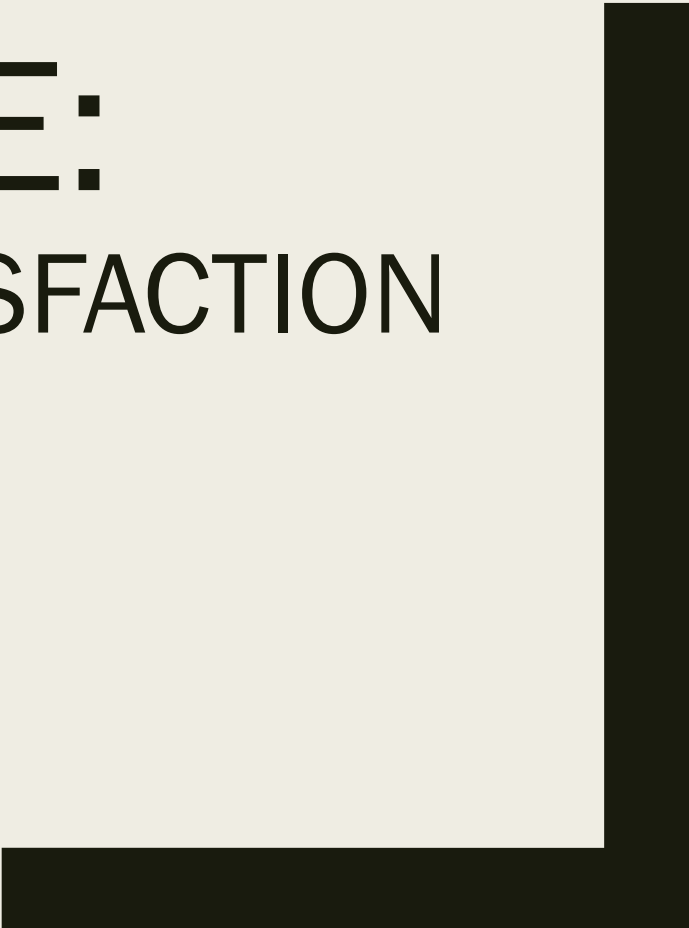




CAS KAGGLE:

AIRLINE PASSENGER SATISFACTION

Albert Company Macarrón
1490992



Introducció

L'objectiu d'aquesta pràctica consisteix en trobar el millor model per a representar les dades del nostre dataset.

El nostre Dataset :

- Informació de passatgers d'una aerolínea.
- Conté un dataset pel train i un pel test
 - Train: 103.904 files
 - Test: 25.976 files
- El nostre objectiu serà predir si el client está o no satisfet amb el vol.

Anàlisi de les dades

- Dataset amb 24 atributs:

id	103904	non-null	int64
Gender	103904	non-null	object
Customer Type	103904	non-null	object
Age	103904	non-null	int64
Type of Travel	103904	non-null	object
Class	103904	non-null	object
Flight Distance	103904	non-null	int64
Inflight wifi service	103904	non-null	int64
Departure/Arrival time convenient	103904	non-null	int64
Ease of Online booking	103904	non-null	int64
Gate location	103904	non-null	int64
Food and drink	103904	non-null	int64
Online boarding	103904	non-null	int64
Seat comfort	103904	non-null	int64
Inflight entertainment	103904	non-null	int64
On-board service	103904	non-null	int64
Leg room service	103904	non-null	int64
Baggage handling	103904	non-null	int64
Checkin service	103904	non-null	int64
Inflight service	103904	non-null	int64
Cleanliness	103904	non-null	int64
Departure Delay in Minutes	103904	non-null	int64
Arrival Delay in Minutes	103594	non-null	float64
satisfaction	103904	non-null	object

Atribut objectiu: satisfaction

Tipus d'atributs:

- Int
- Float
- Strings → Int

Eliminem columnes irrelevantes com el id

Eliminem files amb valors nulls
(Departure Delay in Minutes)

Anàlisi de les dades

Age	1	0.049	0.12	0.1	0.017	0.024	0.023	0.21	0.16	0.076	0.057	0.04	-0.048	0.035	-0.05	0.053	0.14
Type of Travel	0.049	1	0.49	0.27	0.1	0.13	0.063	0.22	0.12	0.15	0.056	0.14	0.031	-0.017	0.022	0.079	0.45
Class	0.12	0.49	1	0.43	0.023	0.094	0.077	0.3	0.21	0.18	0.21	0.2	0.16	0.16	0.16	0.13	0.45
Flight Distance	0.1	0.27	0.43	1	0.0071	0.066	0.057	0.22	0.16	0.13	0.11	0.13	0.063	0.073	0.057	0.093	0.3
Inflight wifi service	0.017	0.1	0.023	0.0071	1	0.72	0.13	0.46	0.12	0.21	0.12	0.16	0.12	0.043	0.11	0.13	0.28
Ease of Online booking	0.024	0.13	0.094	0.066	0.72	1	0.032	0.4	0.03	0.047	0.039	0.11	0.039	0.011	0.035	0.016	0.17
Food and drink	0.023	0.063	0.077	0.057	0.13	0.032	1	0.23	0.57	0.62	0.059	0.032	0.035	0.087	0.034	0.66	0.21
Online boarding	0.21	0.22	0.3	0.22	0.46	0.4	0.23	1	0.42	0.29	0.16	0.12	0.083	0.2	0.074	0.33	0.5
Seat comfort	0.16	0.12	0.21	0.16	0.12	0.03	0.57	0.42	1	0.61	0.13	0.11	0.075	0.19	0.069	0.68	0.35
Inflight entertainment	0.076	0.15	0.18	0.13	0.21	0.047	0.62	0.29	0.61	1	0.42	0.3	0.38	0.12	0.41	0.69	0.4
On-board service	0.057	0.056	0.21	0.11	0.12	0.039	0.059	0.16	0.13	0.42	1	0.36	0.52	0.24	0.55	0.12	0.32
Leg room service	0.04	0.14	0.2	0.13	0.16	0.11	0.032	0.12	0.11	0.3	0.36	1	0.37	0.15	0.37	0.096	0.31
Baggage handling	-0.048	0.031	0.16	0.063	0.12	0.039	0.035	0.083	0.075	0.38	0.52	0.37	1	0.23	0.63	0.096	0.25
Checkin service	0.035	-0.017	0.16	0.073	0.043	0.011	0.087	0.2	0.19	0.12	0.24	0.15	0.23	1	0.24	0.18	0.24
Inflight service	-0.05	0.022	0.16	0.057	0.11	0.035	0.034	0.074	0.069	0.41	0.55	0.37	0.63	0.24	1	0.089	0.24
Cleanliness	0.053	0.079	0.13	0.093	0.13	0.016	0.66	0.33	0.68	0.69	0.12	0.096	0.096	0.18	0.089	1	0.31
satisfaction	0.14	0.45	0.45	0.3	0.28	0.17	0.21	0.5	0.35	0.4	0.32	0.31	0.25	0.24	0.24	0.31	1
Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Ease of Online booking	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	satisfaction	

Correlació entre les dades:

Per la Variable Objectiu
(satisfaction)

Correlacions més altes :

- Online Boarding (0,5)
- Inflight entreteiment (0,4)
- Type of Travel (0,45)
- Class (0,45)

Mètode d'aprenentatge

Al tractar-se de un target binari (satisfets o no satisfets) he implementat diferents classificadors

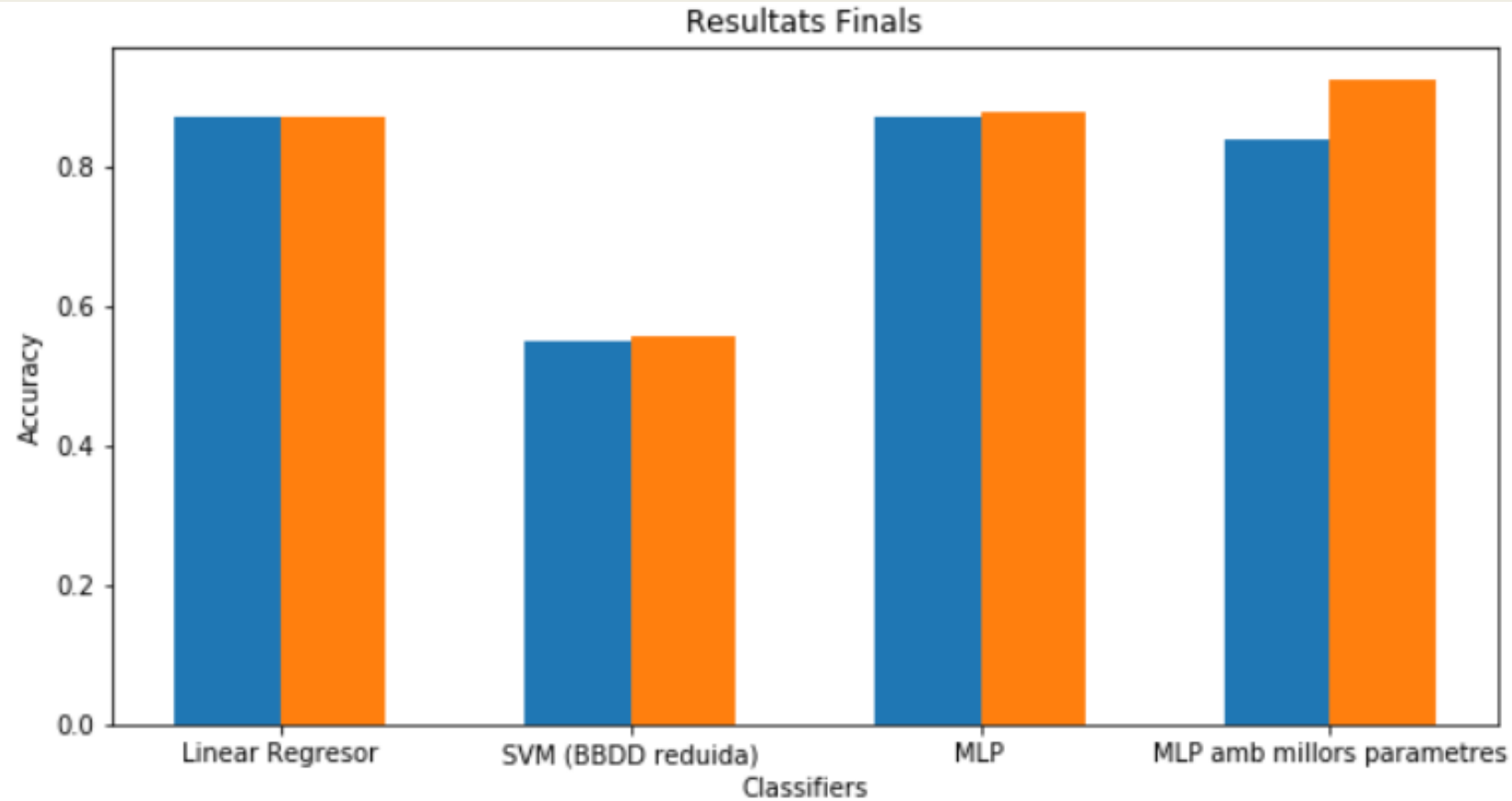
- Regresor Lineal
- SVM (Support Vector Machine)
- MLP (Multi-Layer Perception)

Mètrica escollida: Accuracy

- Cap classificació erronia és especialment perjudicial

Veurem la difència de resultats amb les dades escalades i sense

Resultats



Millor model:
MLP amb paràmetres

Accuracy:
92,60%

Conclusions

- He aconseguit una millora substancial en el MLP al trobar els millors parametres per aplicar-lo ($\alpha=0.001$ i $\text{learning rate}=0.001$)
- Degut a la dimensió de la BBDD no he pogut aplicar una SVM a tot el dataset, només a 5.000 files, i això ha fet que la accuracy fos molt Baixa
- Tots els models han obtingut un millor accuracy amb les dades escalades, menys el Regresor Lògistic que es manté igual
- Un accuracy final prou elevat ($\sim 93\%$), tot i que hi en aquest dataset s'ha arribat a resultats algo millors ($\sim 96\%$)