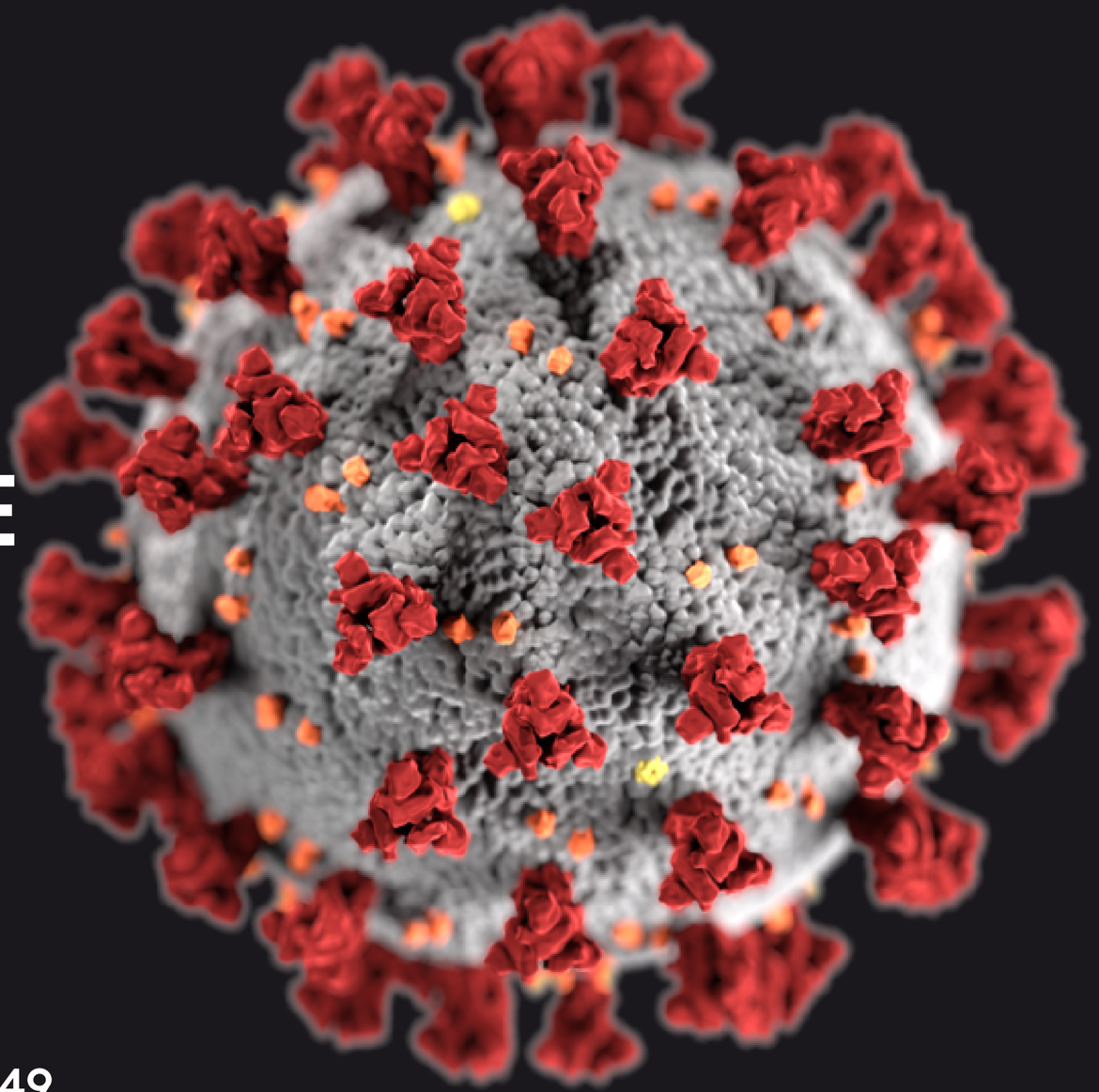
01

ANALISYS PANDEMIA COVID-19: MODELLI DI REGRESSIONE

PROGETTO PER IL CORSO DI DATA MINING
2019/2020



DOCENTI:
SERGIO GRECO
ANTONIO CALIÒ

STUDENTE:
MATTIA GATTO

MATRICOLA: 216649

Contesto

- La COVID-19, o malattia respiratoria acuta da SARS-CoV-2 e più semplicemente malattia da coronavirus 2019 o anche morbo da coronavirus 2019, è una malattia infettiva respiratoria causata dal virus denominato SARS-CoV-2 appartenente alla famiglia dei coronavirus.

- I primi casi sono stati riscontrati durante la pandemia di COVID-19 del 2019-2020.

- Ad oggi la situazione globale risulta:
  -11.046.917 casi confermati nel mondo dall'inizio dell'epidemia
  - 526.465 morti

MOTIVAZIONI ALLA BASE DELLO STUDIO

- Ogni giorno dal 22-01-2020 sono stati registrati record per ogni singolo paese nel mondo, tali record contengono varie informazioni su quel paese nella giornata ad una specifica ora dalla protezione civile.

- Mentre i valori relativi ai morti ,ai guariti ,agli ammalati ed ai ricoverati vengono di giorno in giorno sommati al valore del giorno precedente, i valori relativi ai nuovi casi,ai nuovi morti ai nuovi ricoverati ed ai nuovi guariti sono invece relativi alla singola giornata.

- OBIETTIVO :
  Studiamo il Dataset cercando di determinare secondo quali valori è possibile predire attraverso dei modelli di regressione il numero relativo ai Nuovi casi in un sottoinsieme di paesi, che nel mio caso risultano essere i 10 Paesi con un maggior Numero di conferamati fino al 23-06-2020, cercando di rilevare un minor errore possibile attraverso l'uso di metriche di valutazione.
  La predizione viene rivolta verso Delta[i] giorni in avanti, dove Delta è un insieme di 8 valori che determinano di quanti giorni guardare in avanti.

DESCRIZIONE DEL DATASET(I)

IL DATASET PRESENTA 28798 RIGHE E 10 COLONNE, DI CUI:

-7 COLONNE SONO DI TIPO INTERO(SONO RELATIVE AI VALORI DI CONFIRMED, ACTIVE, DEATHS, RICOVERED, NEW CASES,NEW DEATHS, NEW RECOVERED)

-1 COLONNA È DI TIPO DATE(OSSIA IL SINGOLO GIORNO)

-2 COLONNE SONO DI TIPO OBJECT (COUNTRY/REGION E WHO REGION)

DESCRIZIONE DEL DATASET(II)

	Date	Country/Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	WHO Region
0	2020-01-22	Afghanistan	0	0	0	0	0	0	0	Eastern Mediterranean
1	2020-01-22	Albania	0	0	0	0	0	0	0	Europe
2	2020-01-22	Algeria	0	0	0	0	0	0	0	Africa
3	2020-01-22	Andorra	0	0	0	0	0	0	0	Europe
4	2020-01-22	Angola	0	0	0	0	0	0	0	Africa

PRE PROCESSING: DATA CLEANING

01

Phase 1

Riformulo il Dataset costruendo un DataFrame che come indice possiede la data per sfruttare le proprietà derivanti dalle Time-Series

02

Phase 2

Gli attributi che ultizzerò nel Dataset Saranno le colonne originarie escludendo però la colonna relativa alle Regioni poichè possiedo quella relativa agli Stati/Paesi.

03

Phase 3

Non possiedo alcun dato nulli, mancanti o duplicati poichè i dati sono specificamente indicati e precisi poichè definiti dala protezione civile.

04

Phase 4

Per integrare il mio Dataset e Definire una nuova proprietà, ossia il Lockdown adattato da ogni singolo paese, uso un Dataset esterno che indicherà per ogni paese uno dei tre livelli possibili di Lockdown Adattato:

-TOTALE
-PARZIALE
-NULLO

VISUALIZZAZION(I)

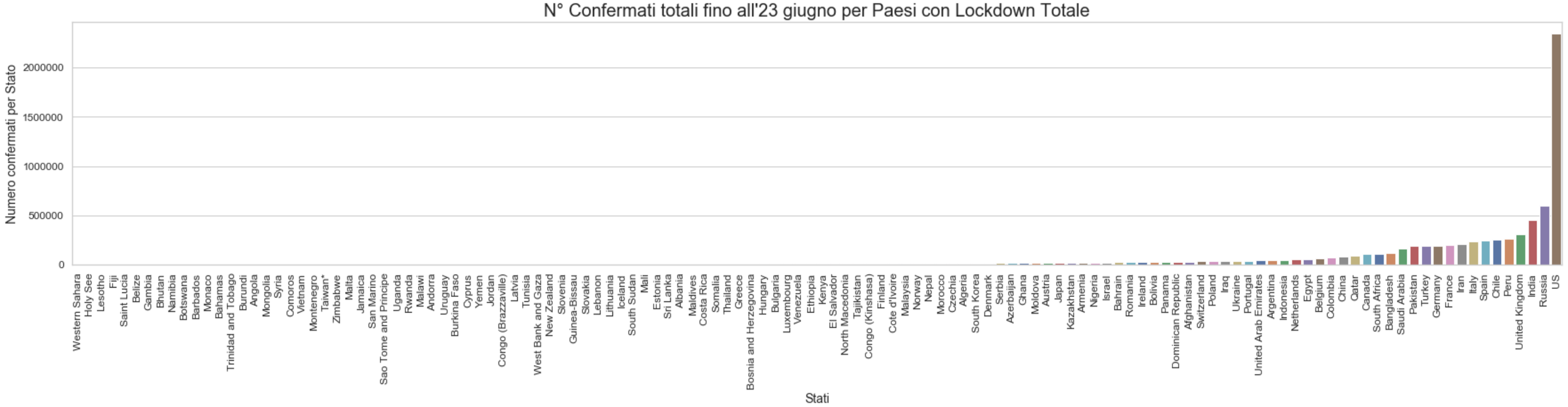
	Paese	Confermati	Nuovi casi	Ricoverati	Nuovi Ricoverati	Morti	Nuovi morti	Lockdown
2020-01-22	Afghanistan	0	0	0	0	0	0	TOTALE
2020-01-22	Albania	0	0	0	0	0	0	TOTALE
2020-01-22	Algeria	0	0	0	0	0	0	TOTALE
2020-01-22	Andorra	0	0	0	0	0	0	TOTALE
2020-01-22	Angola	0	0	0	0	0	0	TOTALE
...	...	...	...	...	...	...	...	...
2020-01-27	Finland	0	0	0	0	0	0	TOTALE
2020-01-27	France	3	0	0	0	0	0	TOTALE
2020-01-27	Gabon	0	0	0	0	0	0	NESSUNO
2020-01-27	Gambia	0	0	0	0	0	0	TOTALE
2020-01-27	Georgia	0	0	0	0	0	0	PARZIALE

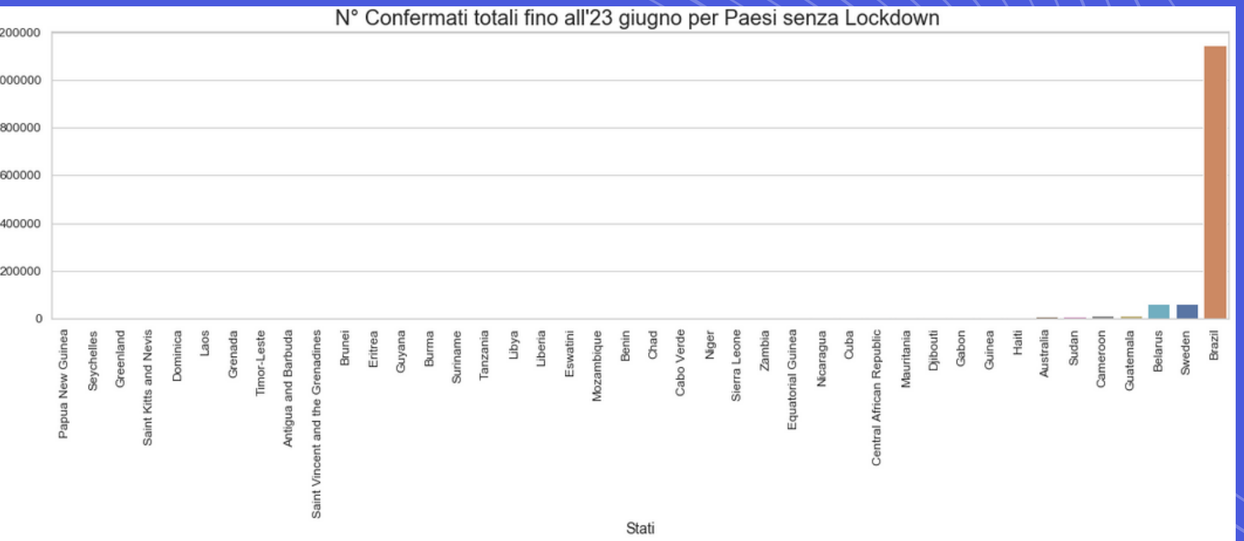
1000 rows × 8 columns

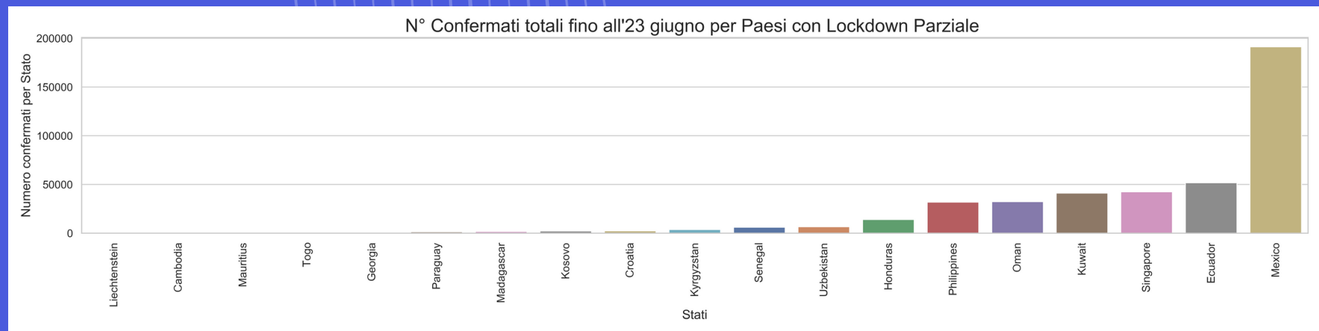
ECCO UN ESEMPIO DELLE PRIME 1000 COLONNE DOPO LE MODIFICHE DI PREROCESSING ADATTATE.

VSUALIZZAZION(II)

Andmaneto grafico dei Confermati totali, parizionando i paesi per Lockdown adattato.

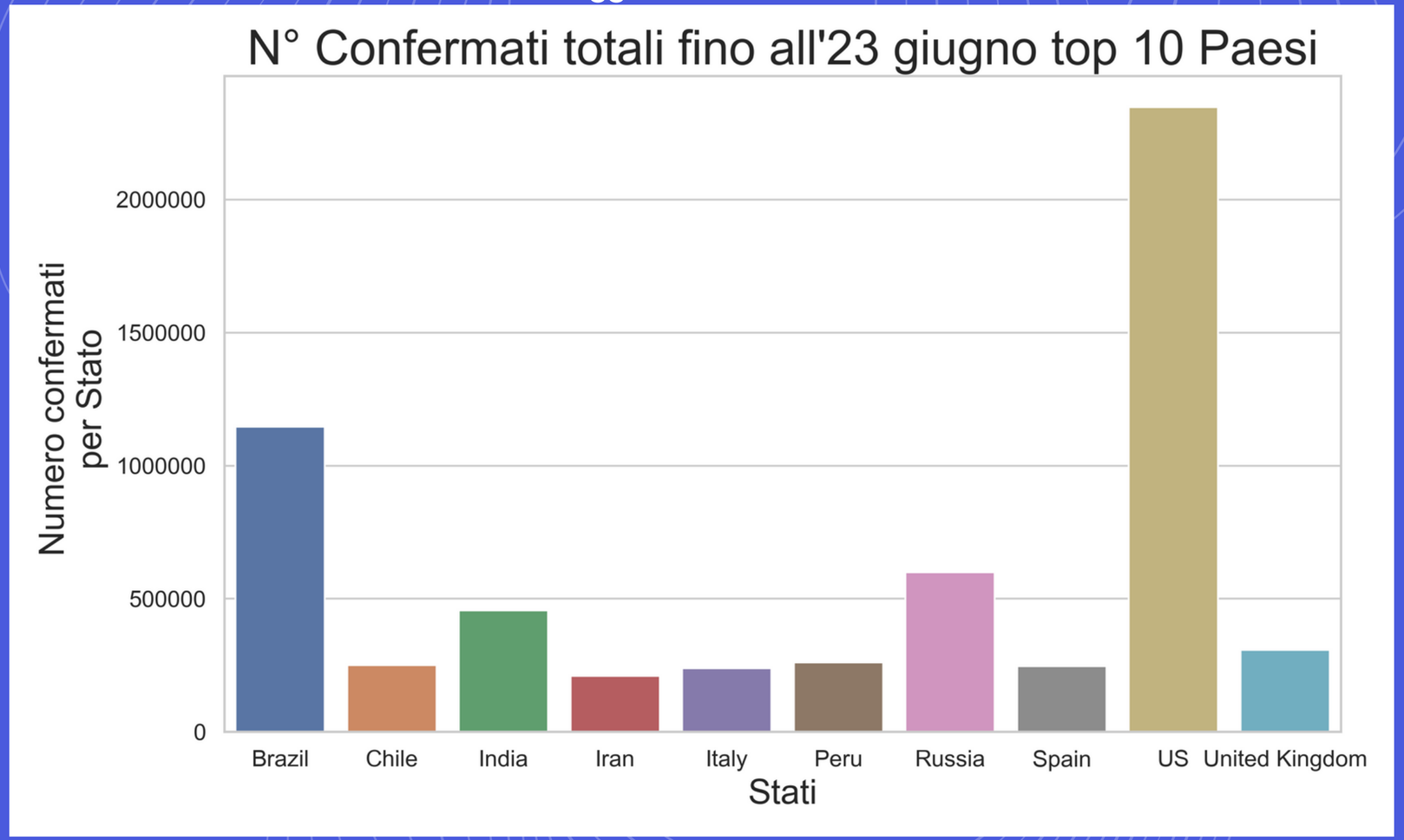






VISUALIZZAZION(III)

Mostriamo ora l'andamento totale dei pasi per quanto riguarda il numero di Confermati raggiunti nel giorno 23-06-2020.
Per definire una migliore visione grafica e dei Dati ci soffermiamo sui 10 Paesi con Maggior numero di casi .



VSUALIZZAZION(IV)

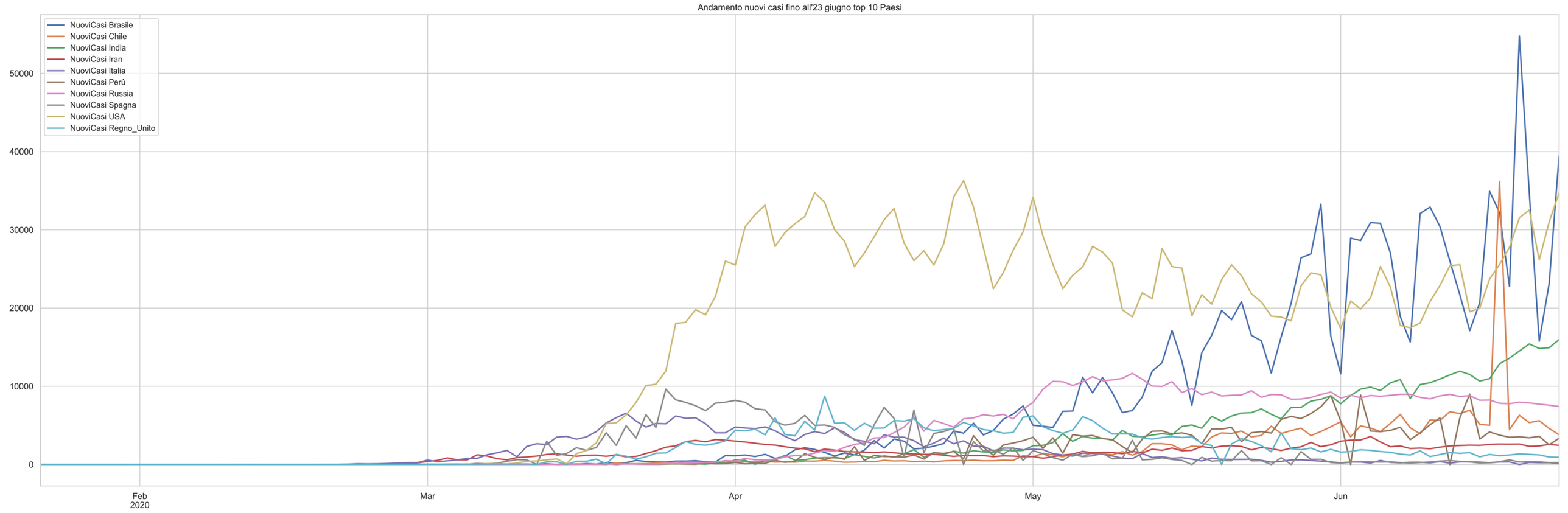
Il DatFrame precedente è riformulato nella seguente versione

	NuoviCasi Brasile	NuoviCasi Chile	NuoviCasi India	NuoviCasi Iran	NuoviCasi Italia	NuoviCasi Perù	NuoviCasi Russia	NuoviCasi Spagna	NuoviCasi USA	NuoviCasi Regno_Unito
2020-01-22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2020-01-25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2020-01-26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0
...	...	...	...	...	...	...	...	...	...	...
2020-06-19	54771.0	6290.0	14516.0	2615.0	0.0	3537.0	7971.0	307.0	31527.0	1350.0
2020-06-20	34666.0	5355.0	15403.0	2322.0	264.0	3413.0	7870.0	363.0	32540.0	1295.0
2020-06-21	15762.0	5607.0	14831.0	2368.0	224.0	3598.0	7717.0	334.0	26171.0	1223.0
2020-06-22	23129.0	4608.0	14933.0	2573.0	221.0	2511.0	7586.0	232.0	31012.0	958.0
2020-06-23	39436.0	3804.0	15968.0	2445.0	113.0	3363.0	7413.0	248.0	34720.0	921.0

154 rows × 10 columns

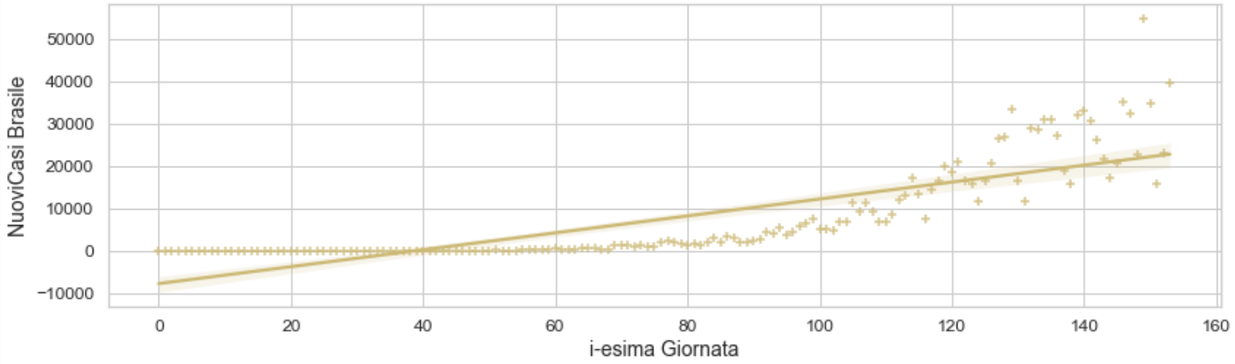
VSUALIZZAZION(V)

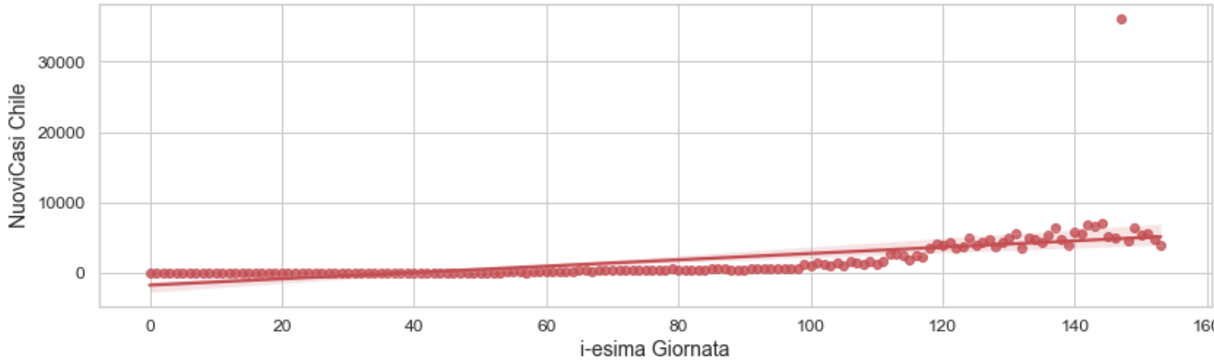
PLOTTIAMO I risultati

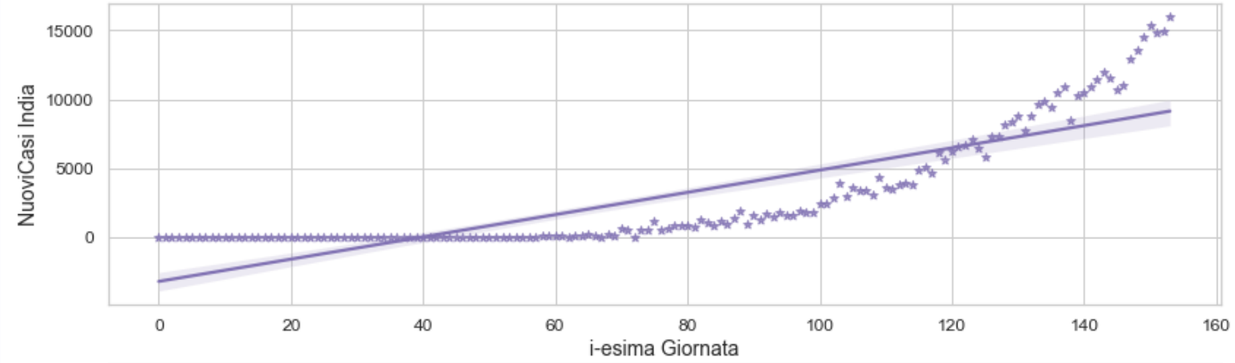


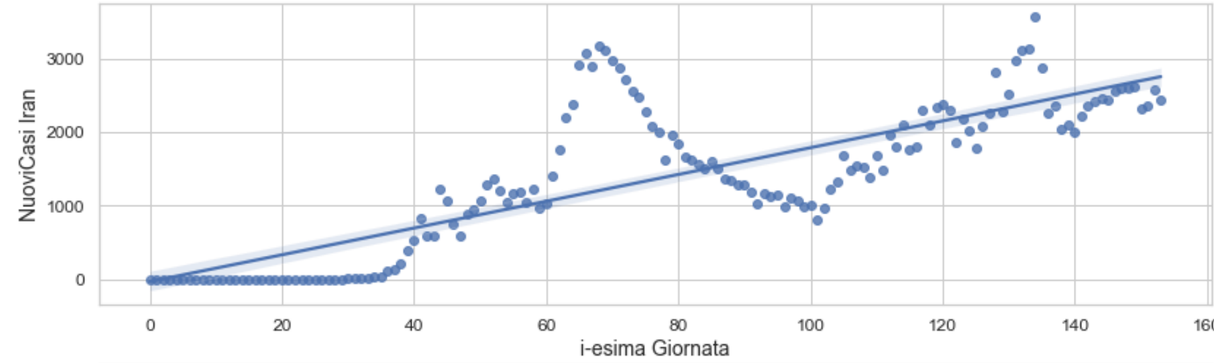
VSUALIZZAZION(VI)

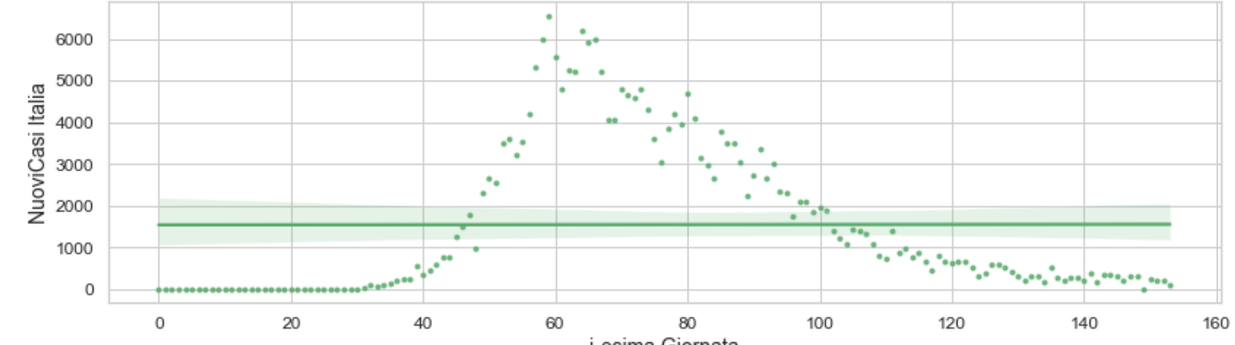
PLOTTIAMO I risultati dei nuovi casi giornaliari per i 10 TOP paesi.

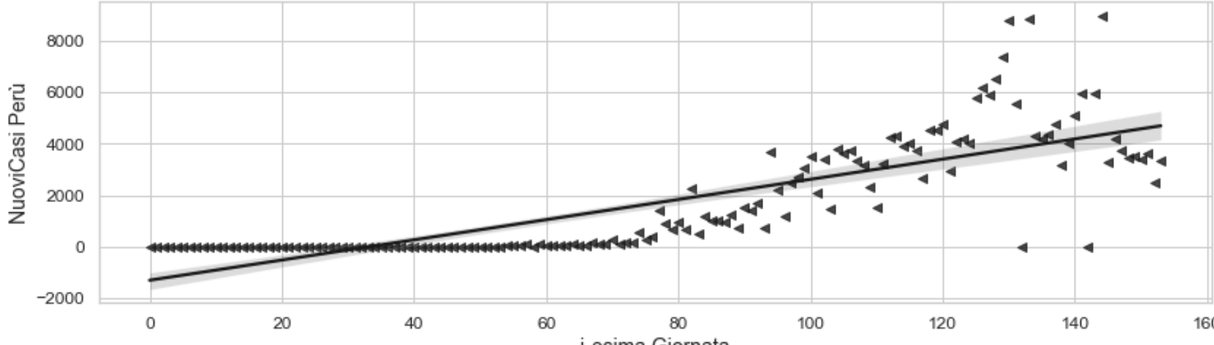




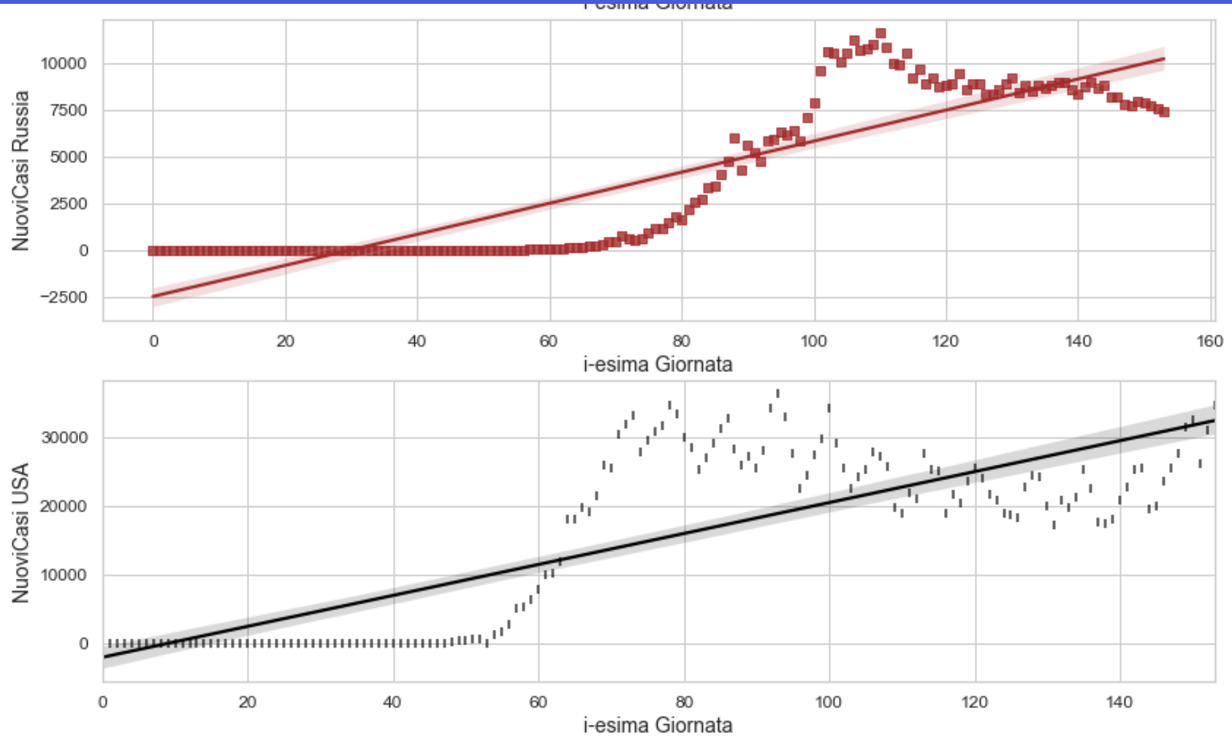


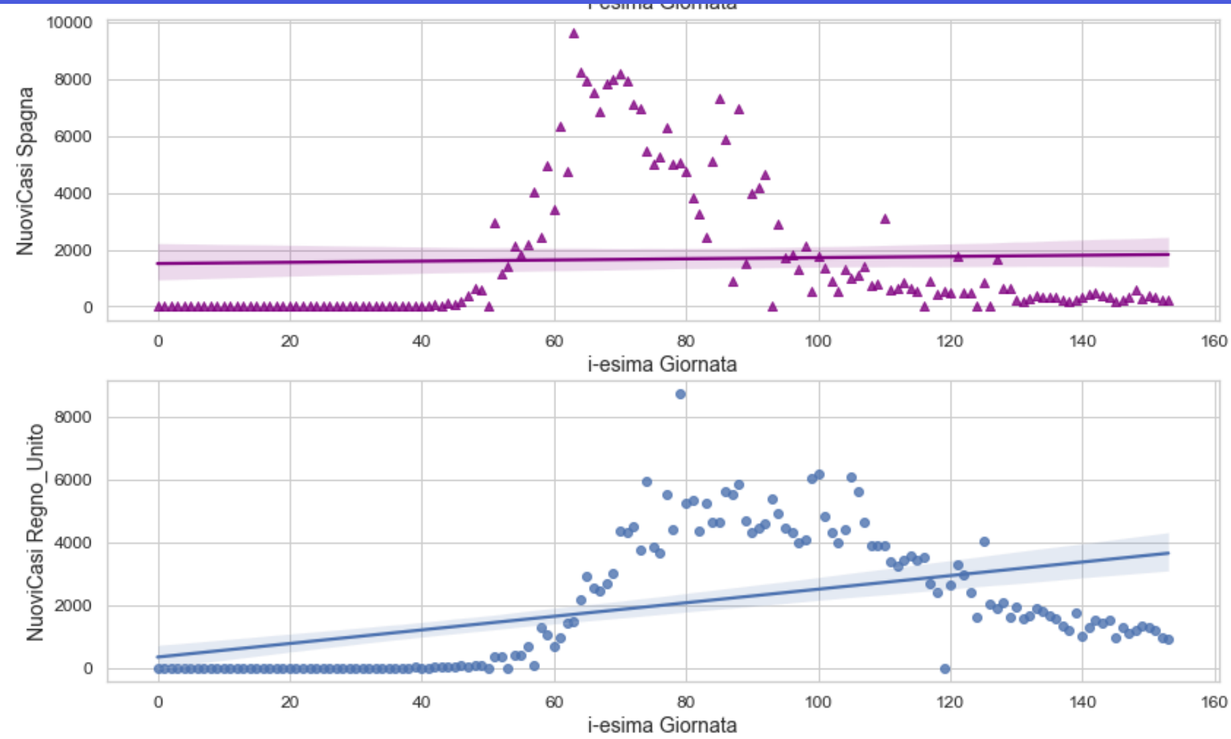






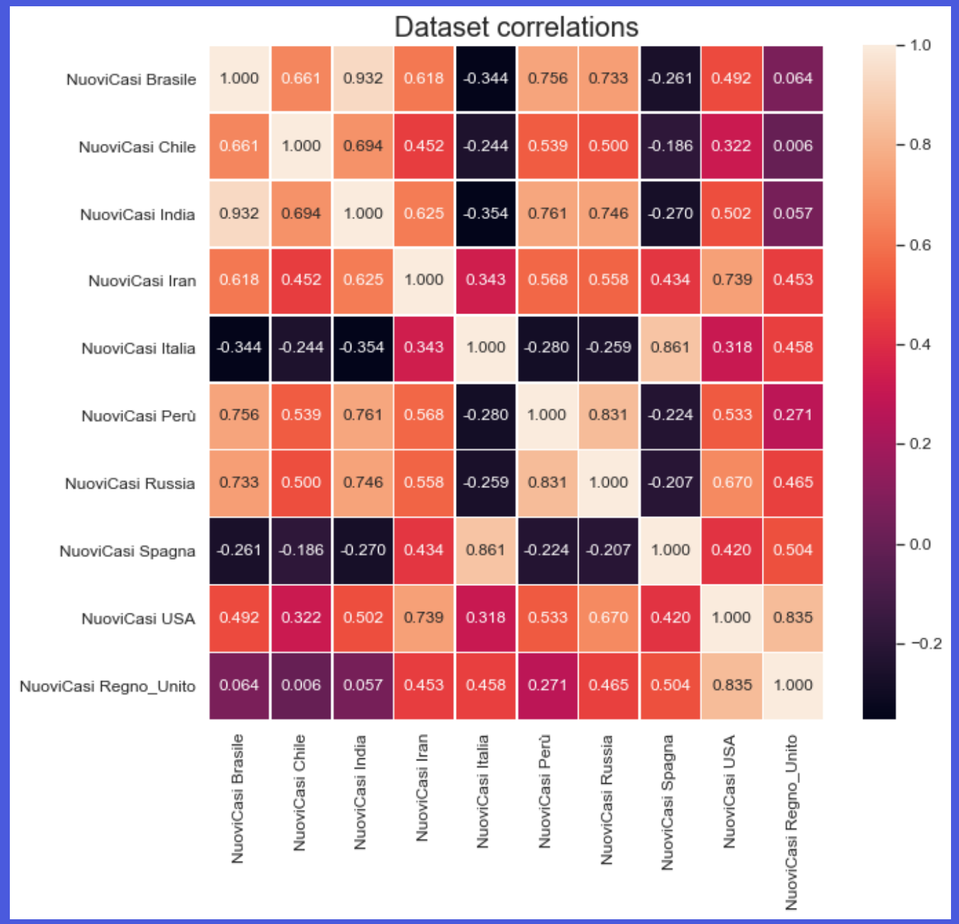
RESTANTI 4 PAESI





MATRICE DI CORRELAZIONE TRA I NUOVI CASI DEI VARI PAESI

ATTRAVERSO LA MATRICE DI CORRELAZIONE VEDIAMO COME SONO CORRELATE LE VARIE NAZIONI ANDANDO A VEDERE IL NUMERO DI NUOVI CASI POSITIVI AL VIRUS GIORNALIERO.

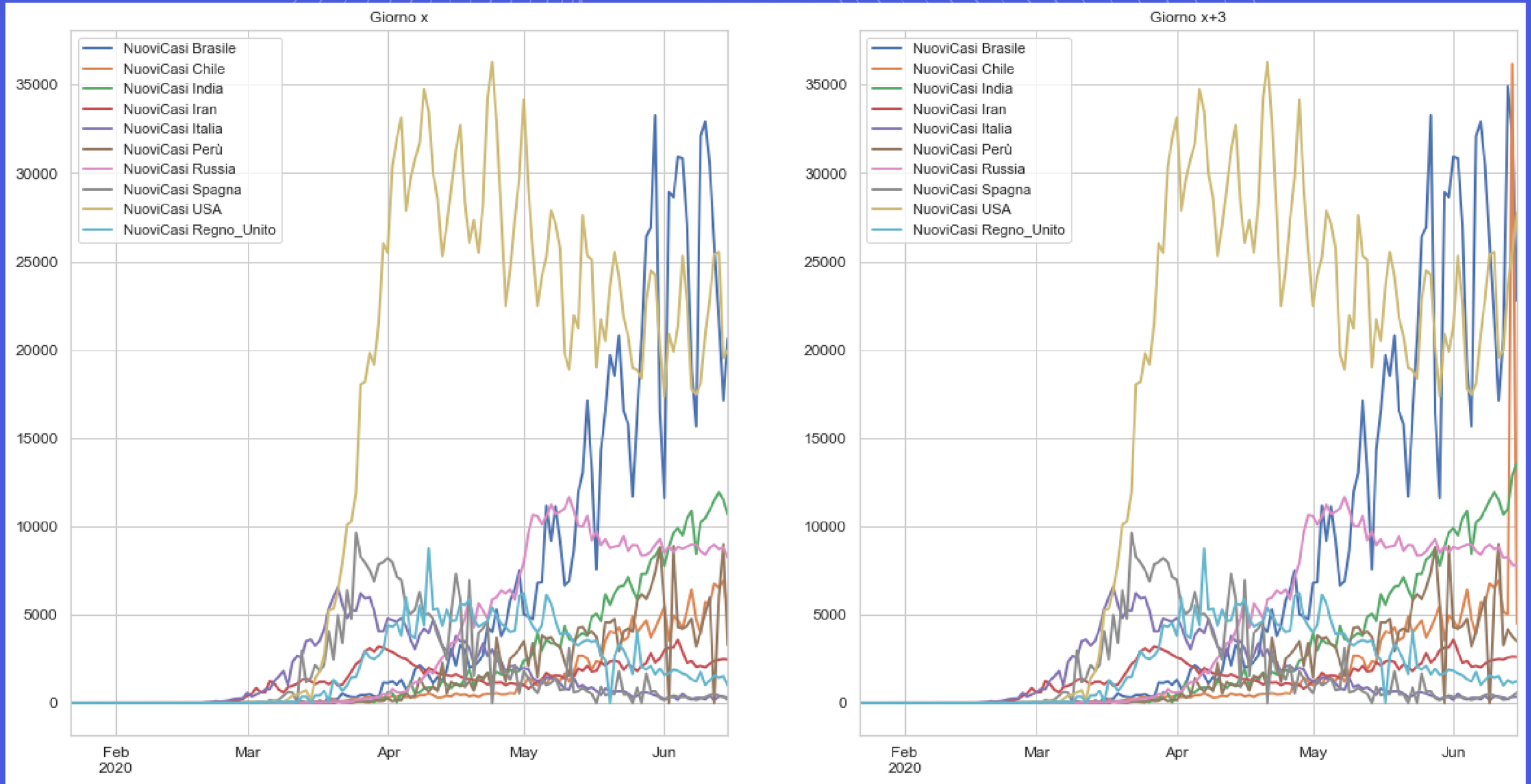


VEDIAMO CHE PER ESEMPIO:

- L'ITALIA HA UNA FORTE CORRELAZIONE CON SPAGNA.
- IL REGNO UNITO CON GLI USA
- BRASILE CON L'INDIA

TRASFORMAZIONI

COSTRUIAMO 8 DATAFRAME, 1 PER OGNI DELTA[I] DOVE AD OGNI RECORD DEL DATAFRAME ORIGINALE IL NUOVO DATAFRAME AVRÀ COME VALORI GLI STESSI MA TRASLATI DI DELT[I] GIORNI. VEDIAMO UN ESMPIO CON GIORNO X E GIORNO X+3.



TRASFORMAZIONI

ORA ANDIAMO A DEFINIRE I NOSTRI TRAIN SET E TEST SET DA POI ANDARE A DARE IN PASTO AGLI ALGORITMI DI REGRESSIONE

INZIALMENTE ANDIAMO A DEFINIRE UN DATAFRAME PER OGNI PAESE NELLA LISTA DEI TOP_10 E SUCCESSIVAMENTE ORGANIZZEREMEO IL NOSTRO METODO DI ADDESTRAMENTO BASANDOCI SU UN RGRESSORE PER OGNI STATO.

	NuoviCasi Italia	Aumento CasiGiorno x+1	Aumento CasiGiorno x+2	Aumento CasiGiorno x+3	Aumento CasiGiorno x+4	Aumento CasiGiorno x+5	Aumento CasiGiorno x+6	Aumento CasiGiorno x+7	Aumento CasiGiorno x+8
2020-03-20	5986.0	571.0	-426.0	-1197.0	-737.0	-776.0	217.0	-77.0	-12.0
2020-03-21	6557.0	-997.0	-1768.0	-1308.0	-1347.0	-354.0	-648.0	-583.0	-1340.0
2020-03-22	5560.0	-771.0	-311.0	-350.0	643.0	349.0	414.0	-343.0	-1510.0
2020-03-23	4789.0	460.0	421.0	1414.0	1120.0	1185.0	428.0	-739.0	-736.0
2020-03-24	5249.0	-39.0	954.0	660.0	725.0	-32.0	-1199.0	-1196.0	-467.0
2020-03-25	5210.0	993.0	699.0	764.0	7.0	-1160.0	-1157.0	-428.0	-542.0
2020-03-26	6203.0	-294.0	-229.0	-986.0	-2153.0	-2150.0	-1421.0	-1535.0	-1618.0
2020-03-27	5909.0	65.0	-692.0	-1859.0	-1856.0	-1127.0	-1241.0	-1324.0	-1104.0
2020-03-28	5974.0	-757.0	-1924.0	-1921.0	-1192.0	-1306.0	-1389.0	-1169.0	-1658.0
2020-03-29	5217.0	-1167.0	-1164.0	-435.0	-549.0	-632.0	-412.0	-901.0	-1618.0
2020-03-30	4050.0	3.0	732.0	618.0	535.0	755.0	266.0	-451.0	-1011.0

REGRESSORI

```
train_set_X1, test_set_X1,train_set_X2, test_set_X2,train_set_X3, test_set_X3,train_set_X4, test_set_X4,train_set_X5, test_set_X
5,train_set_X6, test_set_X6,train_set_X7, test_set_X7,train_set_X8, test_set_X8,train_set_X9, test_set_X9,train_set_X10, test_se
t_X10,train_set_Y1, test_set_Y1,train_set_Y2,test_set_Y2,train_set_Y3, test_set_Y3,train_set_Y4,test_set_Y4,train_set_Y5, test_s
et_Y5,train_set_Y6, test_set_Y6,train_set_Y7, test_set_Y7,train_set_Y8, test_set_Y8= train_test_split(Lista_Stati_Pred[0],Lista_
Stati_Pred[1],Lista_Stati_Pred[2],Lista_Stati_Pred[3],Lista_Stati_Pred[4],Lista_Stati_Pred[5],Lista_Stati_Pred[6],Lista_Stati_Pr
ed[7],Lista_Stati_Pred[8],Lista_Stati_Pred[9],Y_1giorni_da_X,Y_2giorni_da_X,Y_3giorni_da_X,Y_4giorni_da_X,Y_5giorni_da_X,Y_6gior
ni_da_X,Y_7giorni_da_X,Y_8giorni_da_X,test_size=0.3)

lista_Train=[train_set_X1,train_set_X2,train_set_X3,train_set_X4,train_set_X5,train_set_X6,train_set_X7,train_set_X8,train_set_X
9, train_set_X10]
lista_TrainY=[train_set_Y1,train_set_Y2,train_set_Y3,train_set_Y4,train_set_Y5,train_set_Y6,train_set_Y7,train_set_Y8]
lista_Test=[test_set_X1,test_set_X2,test_set_X3,test_set_X4,test_set_X5,test_set_X6,test_set_X7,test_set_X8,test_set_X9,test_set
_X10]
lista_TestY=[test_set_Y1,test_set_Y2,test_set_Y3,test_set_Y4,test_set_Y5,test_set_Y6,test_set_Y7,test_set_Y8]

print(len(train_set_X1), "train +", len(test_set_X1), "test")
```

```
102 train + 44 test
```

- Modelli lineari
  - Linear Regressor, Huber Regressor,SGD Regressor, Bayesian Ridge Regressor.
- Modelli ad albero
  - Decision Tree Regressor, Extra Tree Regressor
- Regressori *instance-based*
  - K Neighbor Regressor
- Modelli *ensemble*
  - Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, Bagging Regressor
- Support Vector Machine
- Neural Network
  - MLP Regressor

Metriche utilizzate

- Variance score

$$variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

- Mean absolute error

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| \cdot y^{(i)}|$$

- Root mean squared error

$$\mathrm{MSE}(y, \hat{y}) = \sqrt{\frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}$$

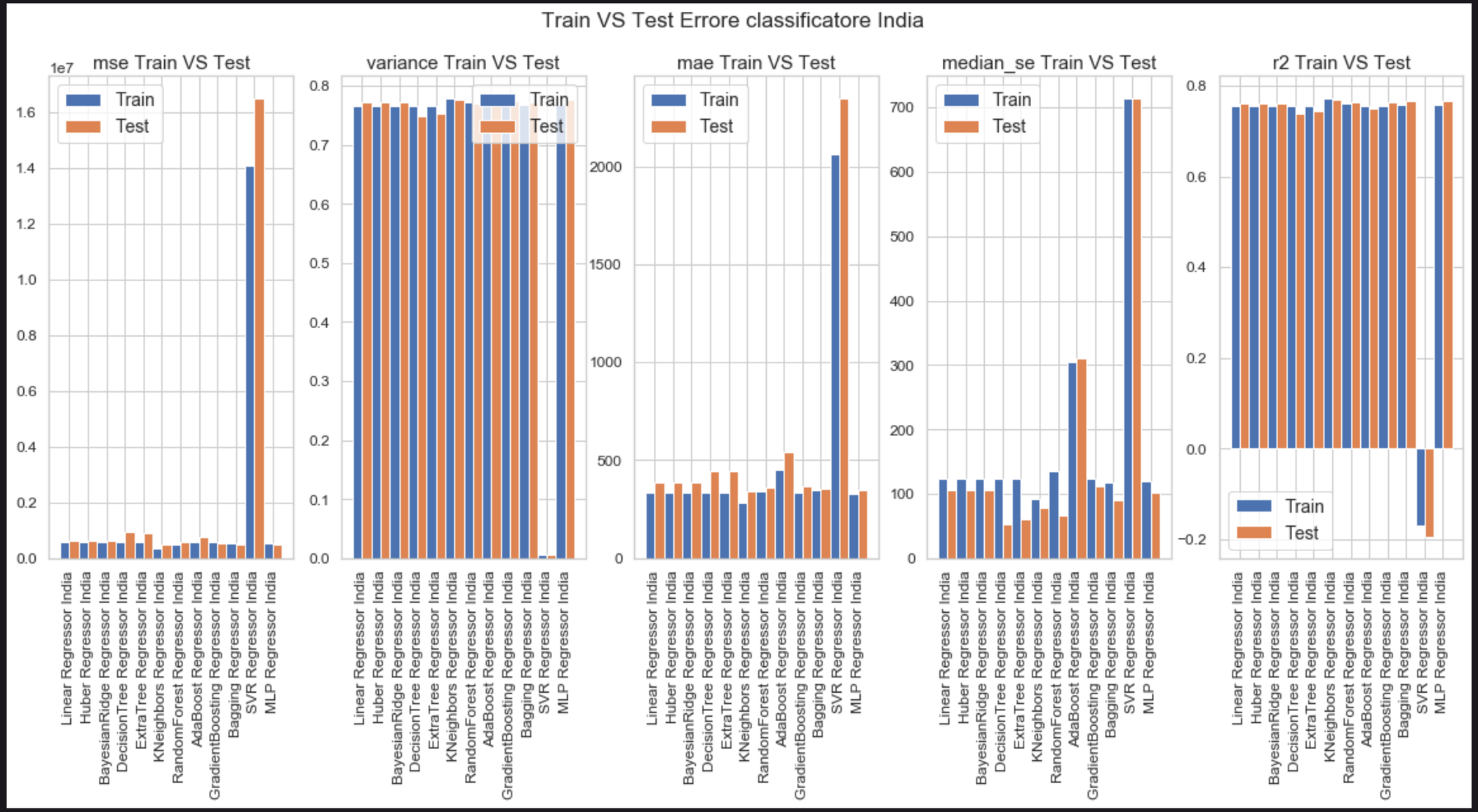
- Median absolute error

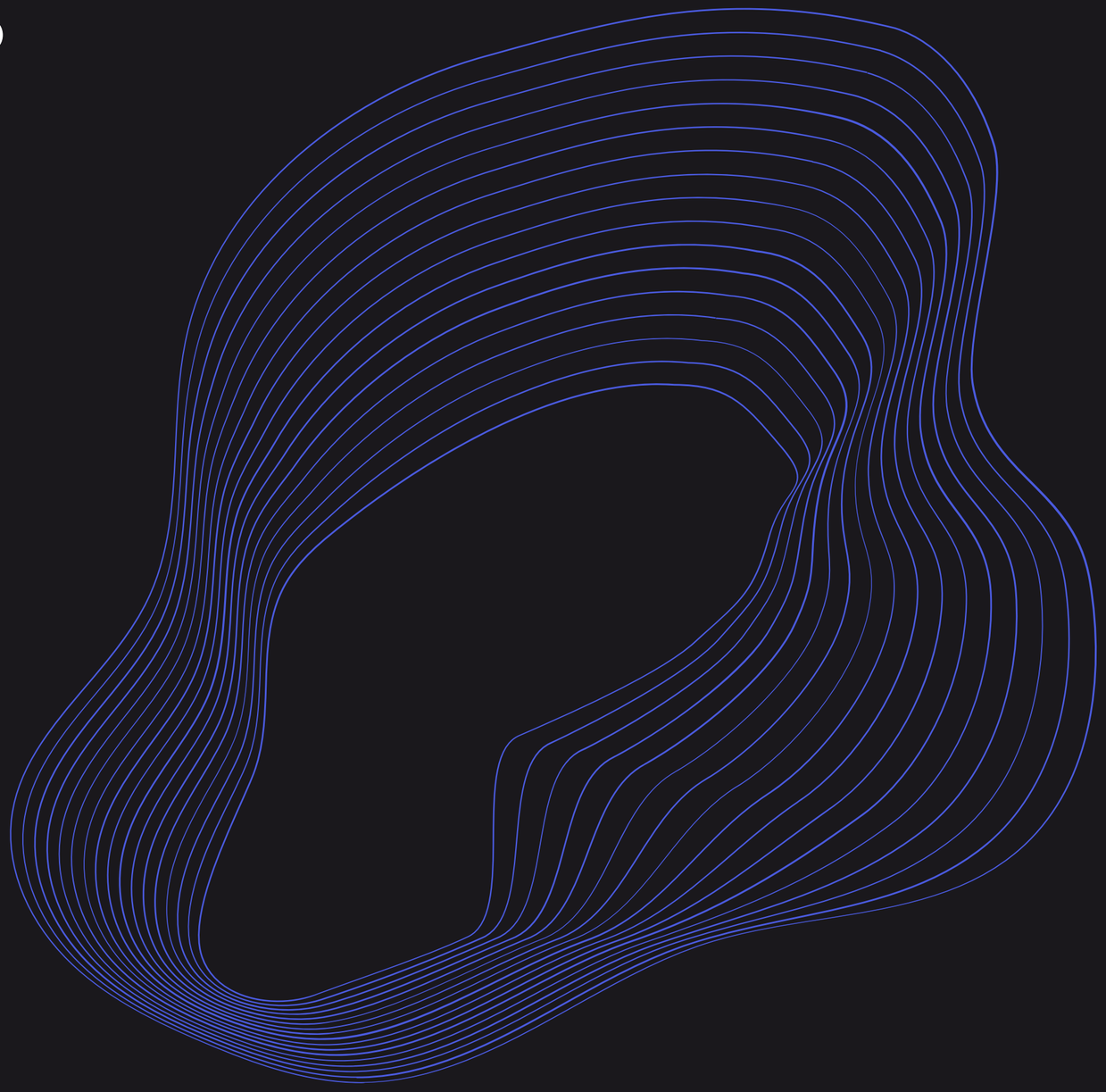
$$\mathrm{MedAE}(y, \hat{y}) = \mathrm{median}(| y_1 - \hat{y}_1 |, \ldots, | y_n - \hat{y}_n |)$$

- R2 score

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Valutazioni Train Vs Test indicativo per l'India





A CAUSA DEI RISULTATI OTTENUTI, PROVIAMO A FARE UN' ALTRA ANALISI UTILIZZANDO PIU' DATI RISPETTO AL SOLO UTILIZZO DEI SINGOLI CASI NELLA FASE DI ADDESTRAMENTO DEI REGRESSORI PER VEDERE SE OTTENIAMO RISULTATI MIGLIORI PER QUANTO RIGUARDA LE METRICHE USATE.

Effetto una trasformazione sul Dataset

Operazioni:

-Binarizzo il valore relativo al LockDown che nel caso dei Paesi Top_10 risulta essere o Totale o Nullo, quindi userò 1 per indicare Lockdown Totale e 0 per indicare Lockdown Nullo.

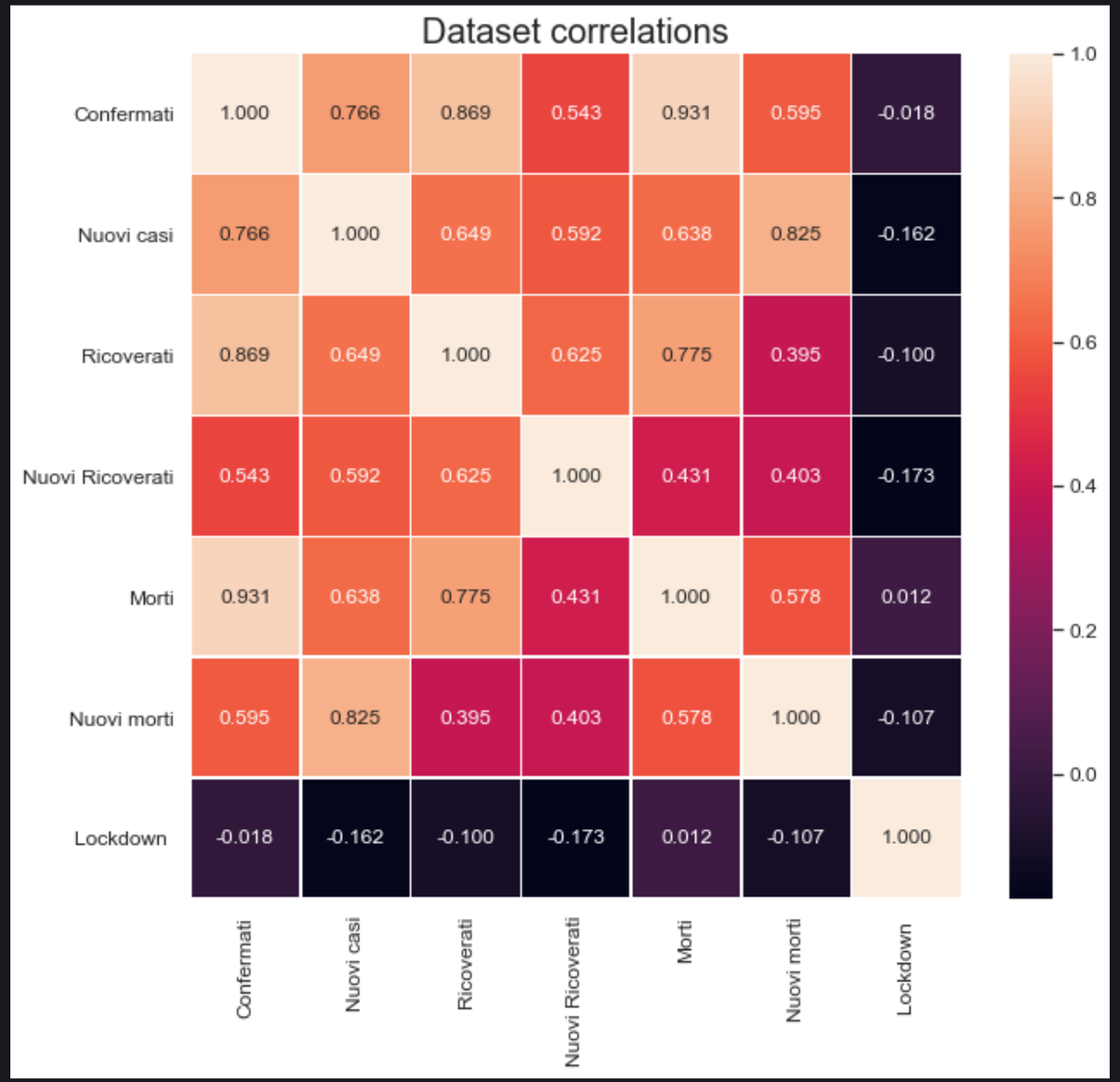
-Elimino il valore relativo ai Paesi successivamente prima di creare un train e un test set da dare poi in pasto agli algoritmi di regressione, poichè comunque rispetta l'ordine alfabetico e i record rimangono in tali posizioni.

	Paese	Confermati	Nuovi casi	Ricoverati	Nuovi Ricoverati	Morti	Nuovi morti	Lockdown
2020-01-22	Brazil	0.0	0.0	0.0	0.0	0.0	0.0	NESSUNO
2020-01-22	Chile	0.0	0.0	0.0	0.0	0.0	0.0	TOTALE
2020-01-22	India	0.0	0.0	0.0	0.0	0.0	0.0	TOTALE
2020-01-22	Iran	0.0	0.0	0.0	0.0	0.0	0.0	TOTALE
2020-01-22	Italy	0.0	0.0	0.0	0.0	0.0	0.0	TOTALE
...	...	...	...	...	...	...	...	...
2020-06-23	Peru	260810.0	3363.0	148437.0	3117.0	8404.0	181.0	TOTALE
2020-06-23	Russia	598878.0	7413.0	355847.0	12000.0	8349.0	153.0	TOTALE
2020-06-23	Spain	246752.0	248.0	150376.0	0.0	28325.0	1.0	TOTALE
2020-06-23	US	2347022.0	34720.0	647548.0	7350.0	121228.0	826.0	TOTALE
2020-06-23	United Kingdom	307682.0	921.0	1330.0	8.0	43011.0	280.0	TOTALE

1540 rows × 8 columns

	Paese	Confermati	Nuovi casi	Ricoverati	Nuovi Ricoverati	Morti	Nuovi morti	Lockdown
2020-01-22	Brazil	0.0	0.0	0.0	0.0	0.0	0.0	0
2020-01-22	Chile	0.0	0.0	0.0	0.0	0.0	0.0	1
2020-01-22	India	0.0	0.0	0.0	0.0	0.0	0.0	1
2020-01-22	Iran	0.0	0.0	0.0	0.0	0.0	0.0	1
2020-01-22	Italy	0.0	0.0	0.0	0.0	0.0	0.0	1
...	...	...	...	...	...	...	...	...
2020-06-23	Peru	260810.0	3363.0	148437.0	3117.0	8404.0	181.0	1
2020-06-23	Russia	598878.0	7413.0	355847.0	12000.0	8349.0	153.0	1
2020-06-23	Spain	246752.0	248.0	150376.0	0.0	28325.0	1.0	1
2020-06-23	US	2347022.0	34720.0	647548.0	7350.0	121228.0	826.0	1
2020-06-23	United Kingdom	307682.0	921.0	1330.0	8.0	43011.0	280.0	1

1540 rows × 8 columns



NOTIAMO DI AVERE UNA FORTISSIMA CORRELAZIONE TRA I DATI.

OVVIAMENTE RISULTA POCA CORRELAZIONE NEL CASO DEL LOCKDOWN POICHÈ COME ABBIAMO GIA OSSERVATO SOPRÀ NON È UN ATTRIBUTO CHE DÀ UNA FORTE IMPRONTA SUI VALORI DELLE ALTRE COLONNE.

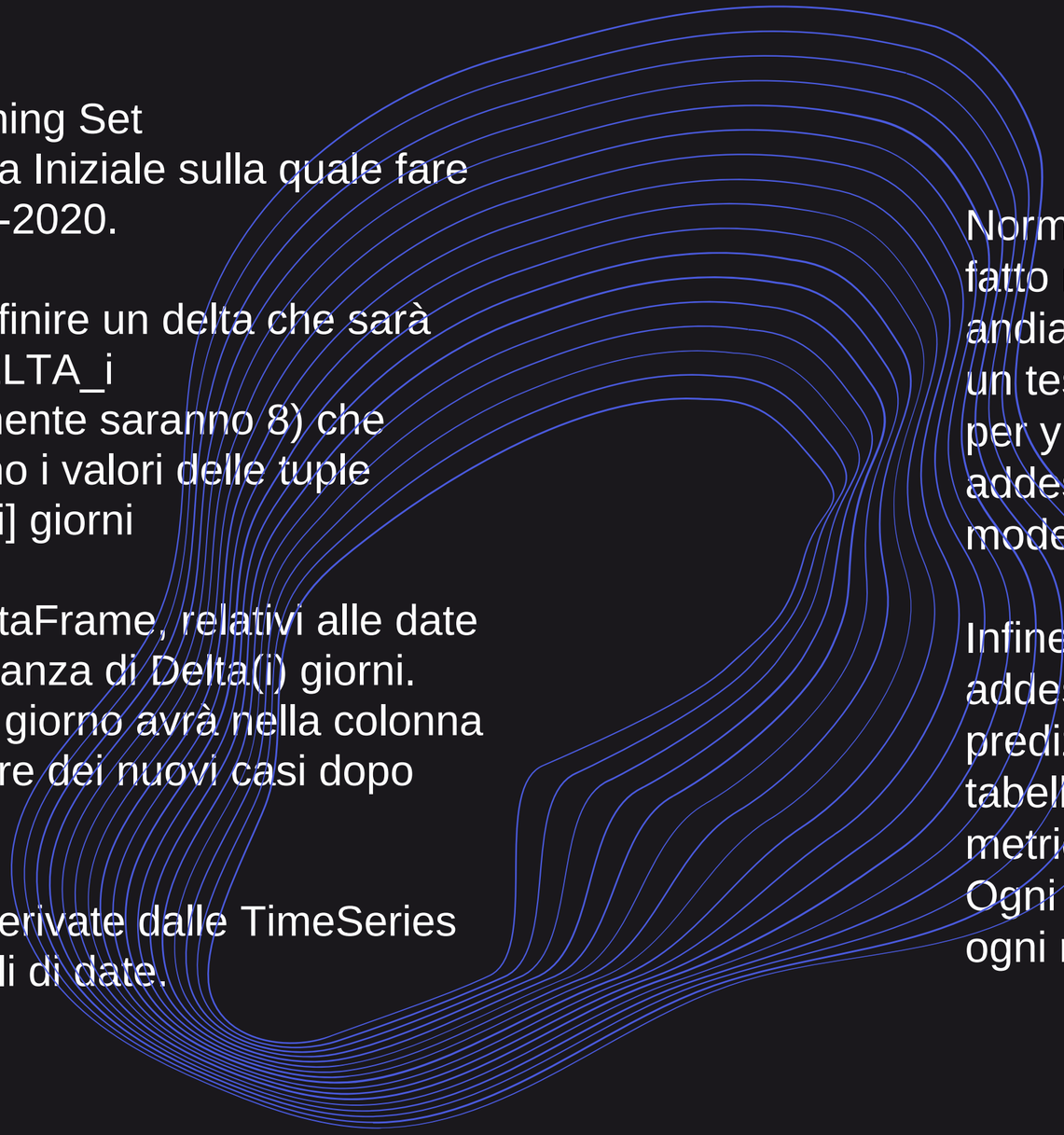
23

Costruzione del Training Set
Prendiamo come data Iniziale sulla quale fare la predizione il 22-01-2020.

Da qui andiamo a definire un delta che sarà utile per costruire DELTA_i DataFrame(precisamente saranno 8) che avranno al loro interno i valori delle tuple traslati a x + DELTA[i] giorni

Ora creo questi 8 DataFrame, relativi alle date per ogni paese a distanza di Delta(i) giorni. Ogni Paese per ogni giorno avrà nella colonna yi il corrispettivo valore dei nuovi casi dopo Delta(i) giorni.

Sfrutto le proprietà Derivate dalle TimeSeries per prendere intervalli di date.



Normalizziamo i nostri dati come fatto nella prova precedente e andiamo a costruire un train e un test set per x e 8 train e 8 test per y che utilizzeremeo per addestrare e predire il nostro modello.

Infine applico uan funzione di addestramento e una funzione di predizione per definire le 8 tabelle con i risultati delle metriche per ogni regressore. Ogni tabella avrà le metriche su ogni regressore a x+i giorni.

	Classifier	mse	variance	mae	median_se	r2
0	Linear Regressor x+1	2.005151e+06	0.956369	636.572119	198.926351	0.956369
1	Huber Regressor x+1	2.147156e+06	0.953524	569.576106	122.631540	0.953279
2	BayesianRidge Regressor x+1	2.015362e+06	0.956147	628.486995	170.360929	0.956147
3	DecisionTree Regressor x+1	3.421213e-01	1.000000	0.089983	0.000000	1.000000
4	ExtraTree Regressor x+1	3.421213e-01	1.000000	0.089983	0.000000	1.000000
5	KNeighbors Regressor x+1	3.133238e+06	0.932234	659.632877	109.900000	0.931822
6	RandomForest Regressor x+1	2.961950e+05	0.993555	213.293572	44.430000	0.993555
7	AdaBoost Regressor x+1	1.797994e+06	0.967334	1024.595931	640.047619	0.960876
8	GradientBoosting Regressor x+1	4.248349e+05	0.990756	348.934456	132.172240	0.990756
9	Bagging Regressor x+1	3.422861e+05	0.992559	228.033093	43.150000	0.992552
10	SVR Regressor x+1	5.324725e+07	0.011207	3546.086468	873.370733	-0.158635
11	MLP Regressor x+1	2.029769e+06	0.956071	611.611277	181.986642	0.955833

	Classifier	mse	variance	mae	median_se	r2
0	Linear Regressor x+8	7.406544e+06	0.852059	1409.634774	434.883622	0.851566
1	Huber Regressor x+8	7.577455e+06	0.851610	1261.517767	329.574018	0.848140
2	BayesianRidge Regressor x+8	7.520094e+06	0.849683	1412.740747	447.075669	0.849290
3	DecisionTree Regressor x+8	7.775873e+06	0.844428	1148.916009	187.500000	0.844164
4	ExtraTree Regressor x+8	6.767120e+06	0.864414	1096.432751	195.000000	0.864380
5	KNeighbors Regressor x+8	1.068142e+07	0.786139	1469.760274	225.900000	0.785934
6	RandomForest Regressor x+8	4.643638e+06	0.907046	952.951657	189.695000	0.906937
7	AdaBoost Regressor x+8	9.950787e+06	0.849407	2511.302941	2524.638462	0.800576
8	GradientBoosting Regressor x+8	4.675148e+06	0.906305	997.847094	214.362546	0.906305
9	Bagging Regressor x+8	4.854198e+06	0.902734	954.454078	164.400000	0.902717
10	SVR Regressor x+8	5.755684e+07	0.007937	3849.265987	1150.014815	-0.153496
11	MLP Regressor x+8	7.587036e+06	0.848666	1304.995404	343.890513	0.847948

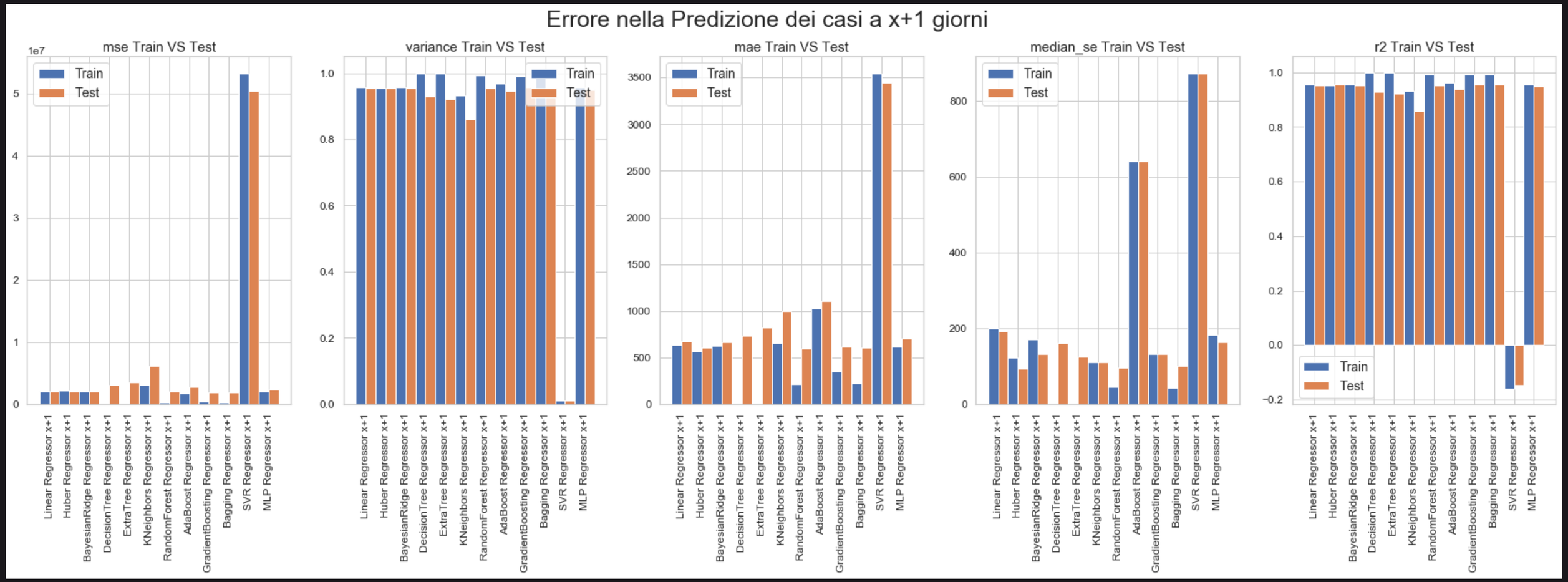
Visualizzazione risultati

Qui sotto vi sono come esempi i risultati nelle tabelle create addestrando:

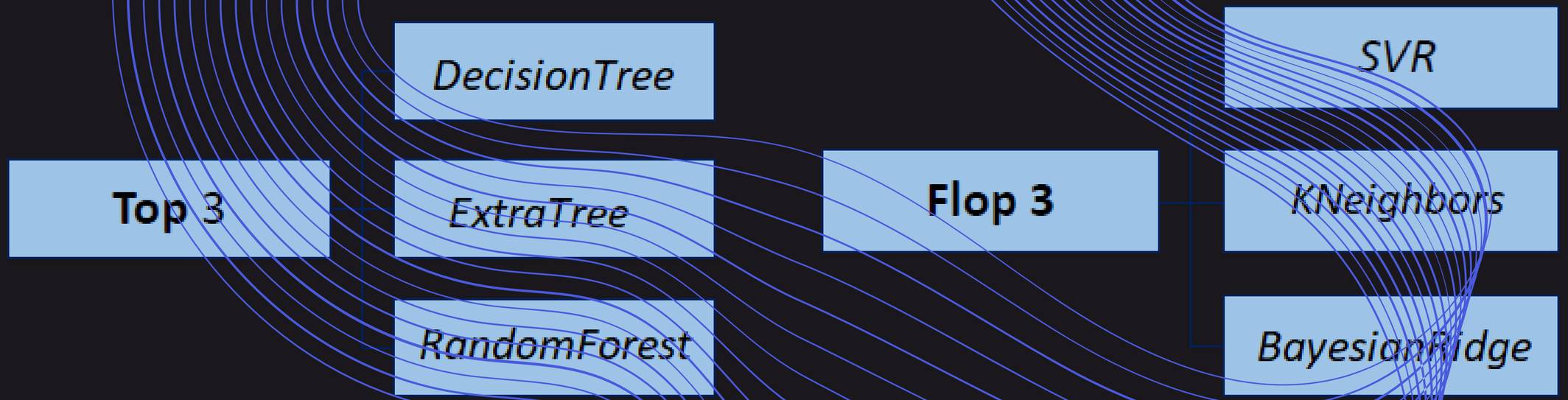
-il train set con gli train_y e facendo predizioni a distanza di x+1 giorni usando sempre il train-set di x e y.

-il train set con gli train_y e facendo predizioni a distanza di x+8 giorni usando sempre il test-set di x e y.

VALUTAZIONI TRAIN VS TEST INDICATIVO PER OGNI PAESE NELLA LISTA DEI TOP_10 IN NUMERO DI CASI A DISTANZA DI X+1 GIORNI



PERFORMANCE: OSSERVAZIONI



CONCLUSIONI

Assenza di **overfitting**

Ottimi scores per il test_set, mettendoli a confonto con quelli ricavati per il train.

Regressori basati su albero e Ensemble Models i migliori.

GRAZIE PER L'ATTENZIONE