Progetto Statistica Numerica

Alessandro Grotti

Indice

1 Introduzione **2**

2 Caricamento del Dataset e Pre-Processing **2**

3 Exploratory Data Analysis (EDA) **3**
3.1 Matrice di Correlazione 3
3.2 Istogrammi delle Variabili Numeriche 3
3.3 Scatter Plot delle Variabili Numeriche 4

4 Splitting del Dataset **5**

5 Regressione Lineare **5**
5.1 3_sound_money e 4_trade 5
5.2 4_trade e 5_regulation 6

6 Addestramento del modello e Hyperparameter Tuning **6**

7 Valutazione della Performance **7**

8 Studio Statistico sui Risultati della Valutazione **7**

1 Introduzione

La **libertà economica** è la capacità degli individui di fare scelte economiche senza eccessive interferenze governative. Include la libertà di spendere, risparmiare, investire, avviare attività, commerciare e proteggere i diritti di proprietà.

L'indice pubblicato in *Economic Freedom of the World* dal Fraser Institute misura il grado in cui le politiche e le istituzioni dei paesi supportano la libertà economica. Ogni anno viene aggiornato con la classifica dell'anno precedente, in ordine dal paese con l'indice di libertà economica migliore a quello con l'indice peggiore.

L'indice è costruito utilizzando 42 indicatori per misurare la libertà economica in cinque aree principali.

1. **Dimensione del Governo**: Maggiore intervento governativo riduce la libertà economica individuale.
2. **Sistema Legale e Diritti di Proprietà**: Protezione efficace delle persone e delle proprietà è fondamentale per la libertà economica.
3. **Moneta Solida**: Stabilità monetaria protegge il valore dei salari e dei risparmi.
4. **Libertà di Commercio Internazionale**: Libertà di scambio con altre nazioni è essenziale.
5. **Regolamentazione**: Regole eccessive limitano la libertà di fare affari e di lavorare.

L'indice viene calcolato tramite questi 5 valori: l'obiettivo del modello di previsione è quello di determinare qual è l'indice a partire dai 5 indicatori. In altre parole, può essere utile per determinare, dati certi valori da input, quale sarebbe il relativo indice di libertà economica.

Link al dataset: *Kaggle Dataset*.

2 Caricamento del Dataset e Pre-Processing

Viene caricato il dataset grazie alla libreria Pandas su Python.

Nel dataset, erano presenti diverse colonne che non sarebbero servite durante il lavoro:

1. **Nome del paese**
2. **Anno della rilevazione**
3. **Posizione nella classifica mondiale**
4. **Quartile**
5. **Sotto-indici diversi dai 5 principali**

Le colonne inutili sono state rimosse e le righe con valori NaN eliminate.

Il target *Economic Freedom Index* è stato discretizzato in classi: questa operazione è stata necessaria perché il modello SVM che verrà utilizzata per fare previsioni funziona con categorie o classi e non con valori numerici continui. Dividendo 'ECONOMIC FREEDOM' in classi, rendiamo più facile per il modello capire e predire a quale categoria appartengono i dati, migliorando così l'accuratezza delle previsioni.

3 Exploratory Data Analysis (EDA)

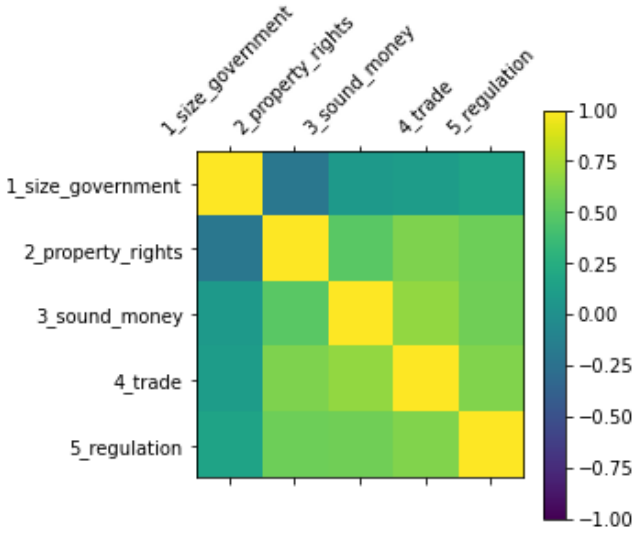
In questa sezione sono stati esplorati diversi aspetti delle variabili numeriche del dataset attraverso strumenti statistici e grafici.

3.1 Matrice di Correlazione

La matrice di correlazione, mostrata nella tabella sottostante, rappresenta i coefficienti di correlazione tra le variabili principali del dataset.

- **var_1**: 1_size_government
- **var_2**: 2_property_rights
- **var_3**: 3_sound_money
- **var_4**: 4_trade
- **var_5**: 5_regulation

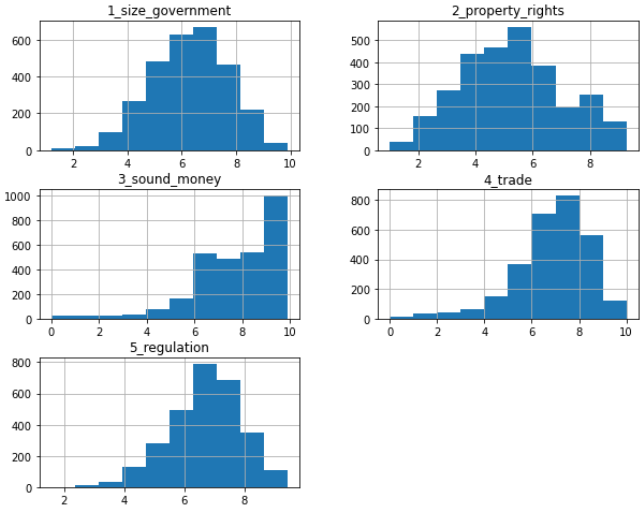
	var_1	var_2	var_3	var_4	var_5
var_1	1.0	-0.2039	0.0851	0.1037	0.1632
var_2	-0.2039	1.0	0.4894	0.6140	0.5588
var_3	0.0851	0.4894	1.0	0.6766	0.5667
var_4	0.1037	0.6140	0.6766	1.0	0.6192
var_5	0.1632	0.5588	0.5667	0.6192	1.0



3.2 Istogrammi delle Variabili Numeriche

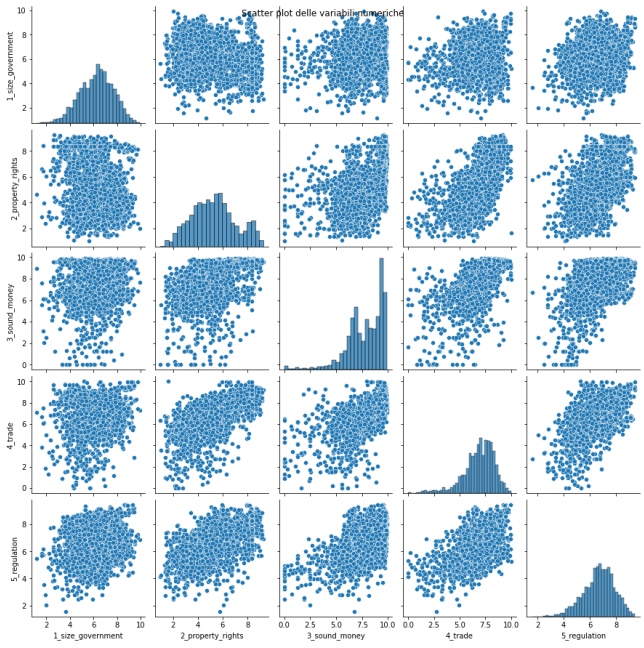
Gli istogrammi delle variabili numeriche, mostrati nella figura sottostante, descrivono la distribuzione univariata di ciascuna variabile nel dataset.

- **1_size_government**: Distribuzione normale con picco tra 6 e 7.
- **2_property_rights**: Distribuzione quasi normale, pochi valori tra 6-8 senno simmetrica.
- **3_sound_money**: Distribuzione sbilanciata verso valori più alti a destra, con picco intorno al valore 10.
- **4_trade**: Distribuzione simile a una distribuzione normale, con picco attorno al valore 7-8, poca coda a destra.
- **5_regulation**: Distribuzione simmetrica con picco intorno al valore 6-7.



3.3 Scatter Plot delle Variabili Numeriche

I grafici a dispersione delle variabili numeriche, rappresentati nella figura qui sotto, mostrano le relazioni bivariate tra le coppie di variabili nel dataset.



- **1_size_government**: In generale poco correlata con le altre variabili, probabilmente perché ha la distribuzione più *normale* tra tutte le variabili.
- **2_property_rights**: Il grafico mostra una parziale correlazione con *4_trade* e *5_regulation*.
- **3_sound_money**: Mostra una correlazione positiva con *4_trade*.
- **4_trade**: Mostra una correlazione positiva con *5_regulation* e, come già detto, una parziale correlazione con *2_property_rights*
- **5_regulation**: Mostra una buona correlazione con *2_property_rights*, *3_sound_money* e *4_trade*.

4 Splitting del Dataset

Il dataset è stato suddiviso in training, validation e test set.

A seguito della pulizia dei dati, le righe nel dataset sono diventate 2891 ed è stata scelta la seguente divisione:

- **data_test** : 20% delle righe totali, 579 righe.
- **data_val** : 20% delle righe totali, 578 righe.
- **data_train** : 60% delle righe totali, 1734 righe.

5 Regressione Lineare

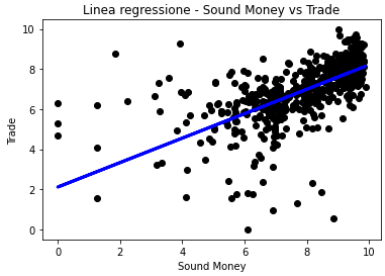
Sono stati eseguiti due modelli di regressione lineare tra due coppie di variabili, ovvero quelle maggiormente correlate tra loro:

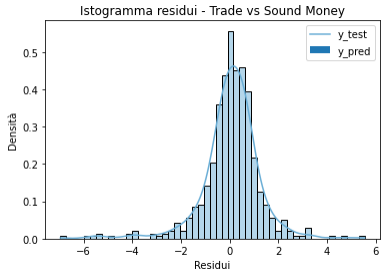
- Regressione tra *3_sound_money* e *4_trade*
- Regressione tra *4_trade* e *5_regulation*

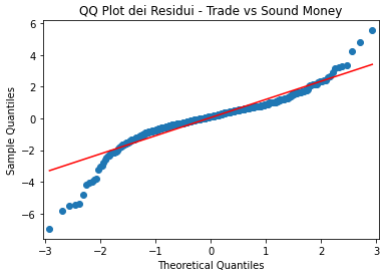
5.1 3_sound_money e 4_trade

Coefficiente di correlazione: **0.676561**
Metriche di regressione:

- R^2: 0.3562093500374176
- MSE: 1.4604609139508336
- Coefficiente: 0.61045335
- Intercetta: 2.11313454





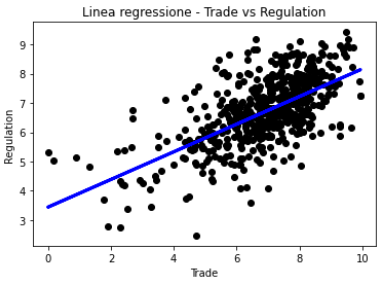


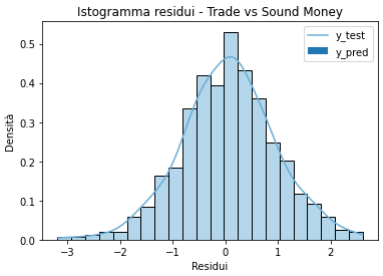
Il valore di R^2, che si aggira intorno al 36%, suggerisce che il modello di regressione lineare spiega il 36% della variabilità nei dati di "Trade" utilizzando "Sound Money". L'alto valore di MSE conferma una correlazione moderata tra le variabili considerate. Il coefficiente positivo indica una relazione positiva tra le due variabili, mentre l'intercetta rappresenta il valore previsto della variabile dipendente quando l'altra è pari a zero. L'istogramma dei residui mostra una distribuzione che ricorda una normale centrata su zero.

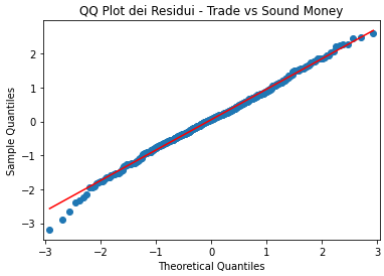
5.2 4_trade e 5_regulation

Coefficiente di correlazione: **0.619235**
Metriche di regressione:

- R^2: 0.3913311722668178
- MSE: 0.7994052446106543
- Coefficiente: 0.4721589
- Intercetta: 3.44482441







Il valore di R^2, che corrisponde a circa il 39.13%, indica che il modello di regressione lineare spiega il 39% della variabilità nei dati di "Trade" utilizzando "Regulation". L'MSE di circa 0.7994 indica un errore quadratico medio relativamente basso. Il coefficiente positivo conferma una correlazione positiva tra "Trade" e "Regulation": un aumento in "Regulation" è associato a un aumento in "Trade". L'intercetta rappresenta il valore stimato di "Trade" quando "Regulation" è zero, mentre l'istogramma dei residui anche in questo caso mostra una distribuzione che ricorda una normale centrata su zero.

6 Addestramento del modello e Hyperparameter Tuning

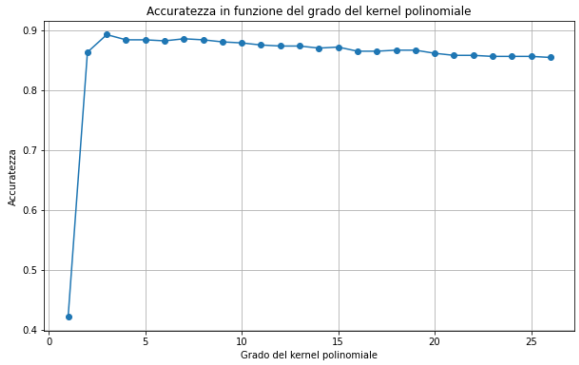
A questo punto è stato necessario cercare il miglior modello di classificazione, con la miglior combinazione di iperparametri. Alla fine è stato scelto il modello di classificazione chiamato SVC (support vector classifier), con kernel di tipo 'linear'. Questa scelta è stata fatta sulla base della miglior accuratezza generata dagli altri.

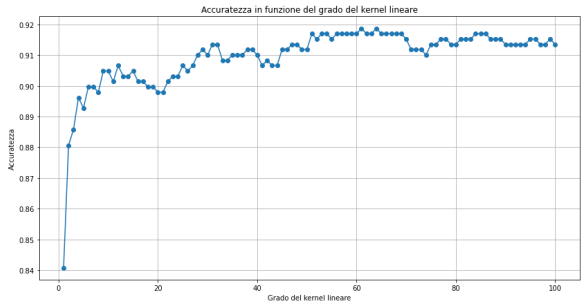
Con random seed 24, questi erano i valori di accuratezza migliore dei 3 possibili modelli:

- linear_model.LogisticRegression() : 0.1228
- svm.SVC(kernel="linear", C = 61) : 0.9186
- svm.SVC(kernel="poly", degree = 3) : 0.8927

Nel caso del modello lineare di regressione logistica, il valore era costante. Dai grafici si evince che il miglior valore di accuratezza è quello prodotto dal modello *svm.SVC(kernel="linear", C = 61)*, con 61 e 64 costi migliori.

Il modello viene addestrato con *data_train* e i parametri vengono scelti valutando il modello con *data_val*.





7 Valutazione della Performance

La performance del modello è stata valutata con *data_set*, per vedere se viene confermata la scelta del modello.

Accuratezza test

- ME : 44.
- MR : 0.07599309153713299.
- Acc : 0.924006908462867.

L'accuratezza del 92.4% conferma l'alta accuratezza ipotizzata e valutata con *data_val*.

8 Studio Statistico sui Risultati della Valutazione

Si ripetono le fasi di splitting, addestramento del modello e valutazione in modo casuale, in modo da ottenere diversi valori di accuratezza del modello su cui fare una valutazione statistica.

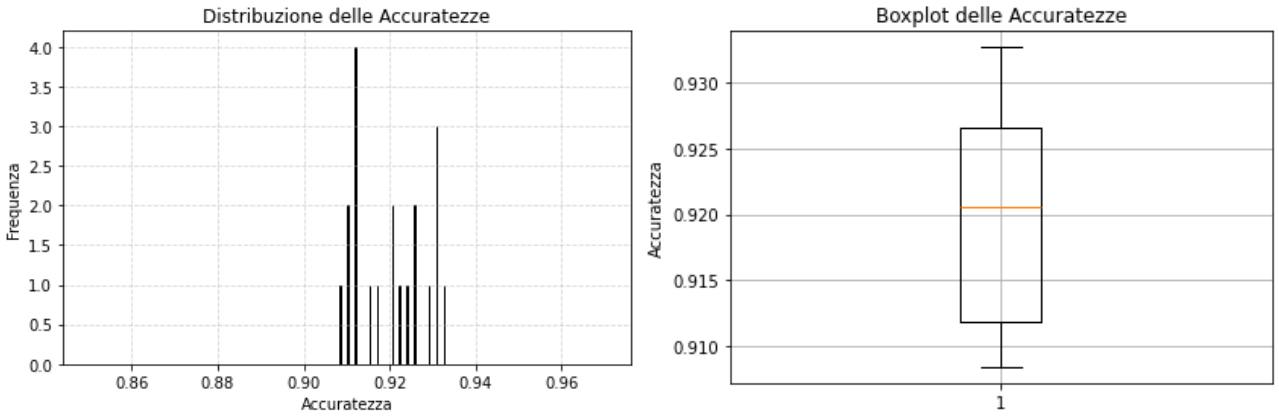
Misure delle caratteristiche dei dati:

- Accuratezza media: 0.9201
- Deviazione standard dell'accuratezza: 0.0081
- Mediana: 0.9206
- Varianza: $6.5289 * 10^{-5}$
- Min: 0.9085
- Max: 0.9326
- 1° Quartile: 0.9119
- 3° Quartile: 0.9266
- Intervallo di confidenza al 95%: [0.9162, 0.9240]

L'analisi dei risultati rivela che l'accuratezza media del modello in questo caso è circa del 92.01%, con una deviazione standard di 0.0081, suggerendo una coerenza accettabile tra le varie misurazioni. La mediana dell'accuratezza, leggermente superiore alla media a 0.9206, indica

una distribuzione tendenzialmente asimmetrica verso valori più elevati. La varianza, pari a 6.5289×10^{-5}, conferma una dispersione minima delle accuratezze rispetto alla media.

Il range interquartile (IQR), compreso tra il primo quartile a 0.9119 e il terzo quartile a 0.9266, evidenzia che la maggior parte delle accuratezze si concentra strettamente intorno alla mediana. I valori minimo e massimo di 0.9085 e 0.9326 rispettivamente mostrano il massimo range di accuratezza raggiunto dal modello in questo caso.



Infine, l'intervallo di confidenza al 95% per l'accuratezza media, calcolato come [0.9162, 0.9240], fornisce un intervallo entro cui si stima con il 95% di sicurezza che il modello mantenga quella specifica accuratezza. Questi risultati suggeriscono che il modello operi con una precisione notevole e consistente.