**PROGETTO ESAME ML E BI**

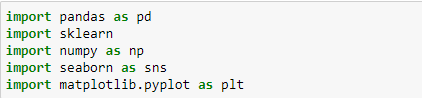
Lo scopo di questo progetto è quello di dare uno strumento utile alle banche al fine di capire e prevedere attraverso la scienza dei dati e il machine learning quando un mutuatario può essere inadempiente sul prestito.

Il dataset è compreso da 34 features qui sotto descritte:

Features

* ID = Customer ID of Applicant
* year = Year of Application
* loan limit = maximum avaliable amount of the loan allowed to be taken
* Gender = sex type
* approv\_in\_adv = Is loan pre-approved or not
* loan\_type = Type of loan
* loan\_purpose = the reason you want to borrow money
* Credit\_Worthiness = is how a lender determines that you will default on your debt obligations, or how worthy you are to receive new credit.
* open\_credit = is a pre-approved loan between a lender and a borrower. It allows the borrower to make repeated withdrawals up to a certain limit.
* business\_or\_commercial = Usage type of the loan amount
* loan\_amount = The exact loan amount
* rate\_of\_interest = is the amount a lender charges a borrower and is a percentage of the principal—the amount loaned.
* Interest\_rate\_spread = the difference between the interest rate a financial institution pays to depositors and the interest rate it receives from loans
* Upfront\_charges = Fee paid to a lender by a borrower as consideration for making a new loan
* term = the loan's repayment period
* Neg\_ammortization = refers to a situation when a loan borrower makes a payment less than the standard installment set by the bank.
* interest\_only = amount of interest only without principles
* lump\_sum\_payment = is an amount of money that is paid in one single payment rather than in installments.
* property\_value = the present worth of future benefits arising from the ownership of the property
* construction\_type = Collateral construction type
* occupancy\_type = classifications refer to categorizing structures based on their usage
* Secured\_by = Type of Collatoral
* total\_units = number of unites
* income = refers to the amount of money, property, and other transfers of value received over a set period of time
* credit\_type = type of credit
* co-applicant\_credit\_type = is an additional person involved in the loan application process. Both applicant and co-applicant apply and sign for the loan
* age = applicant's age
* submission\_of\_application = Ensure the application is complete or not
* LTV = life-time value (LTV) is a prognostication of the net profit
* Region = applicant's place
* Security\_Type = Type of Collatoral
* status = Loan status (Approved/Declined)
* dtir1 = debt-to-income ratio

Questo è lo sviluppo del progetto e la messa a punto del modello predittivo (task di classificazione) per questo dataset.



importiamo le librerie che ci serviranno per manipolare i dati

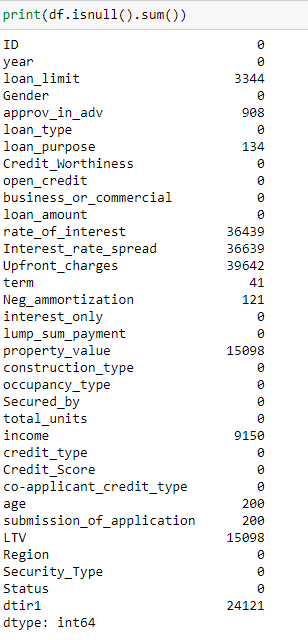


Carichiamo il dataset

Immagine che contiene tavolo

Descrizione generata automaticamente

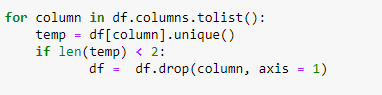
Le nostre informazioni sul dataset



Adesso droppiamo la colonna ID perché è inutile ai fini predittivi



Rimuoviamo inoltre quelle features che hanno solo 1 valore unico e che quindi anch’esse non hanno nessun valore predittivo



Vediamo quante features hanno valori unici sotto i 10 per poi fare il one hot encoding così da poterle usare per creare il nostro modello predittivo in quanto esso ha bisogno di variabili numeriche

Immagine che contiene testo

Descrizione generata automaticamente

Immagine che contiene testo

Descrizione generata automaticamente

Vediamo come nella feature gender ci sia “Sex Not Available” che corrisponde al valore NaN

Quindi lo trasformiamo



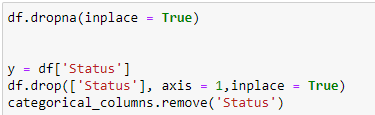
Prendiamo le features che hanno valori numerici che non fanno parte di quelle categoriche e applichiamo la media per ogni variabile che ha un valore NaN

Immagine che contiene testo

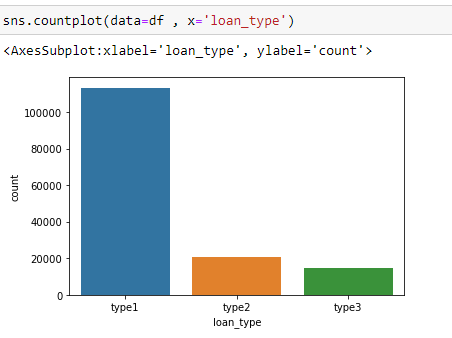
Descrizione generata automaticamente

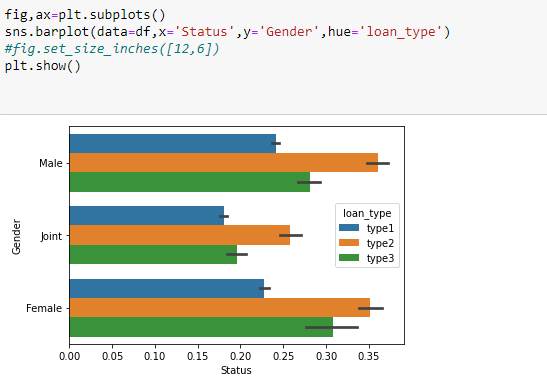
Ora rimuoviamo le restanti variabili che hanno valori Nan.

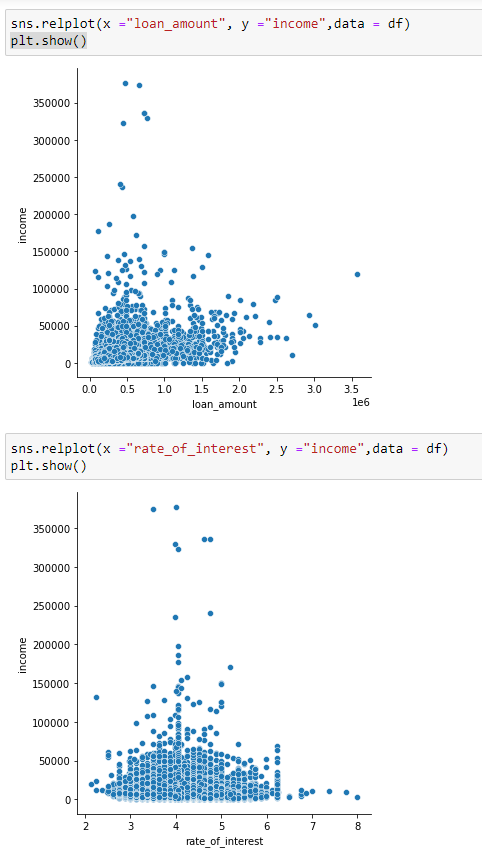
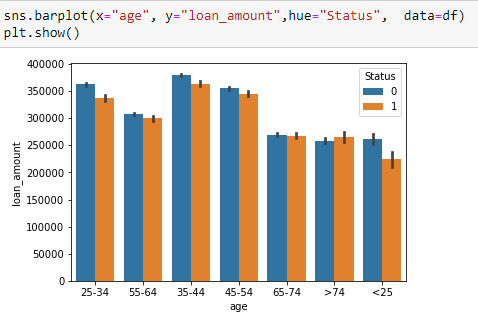
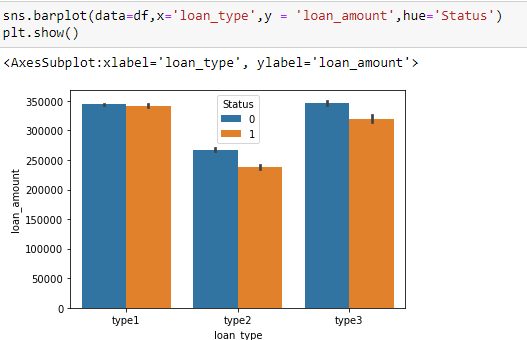
Inoltre droppiamo la feature ‘Status? In quanto nostra variabile target

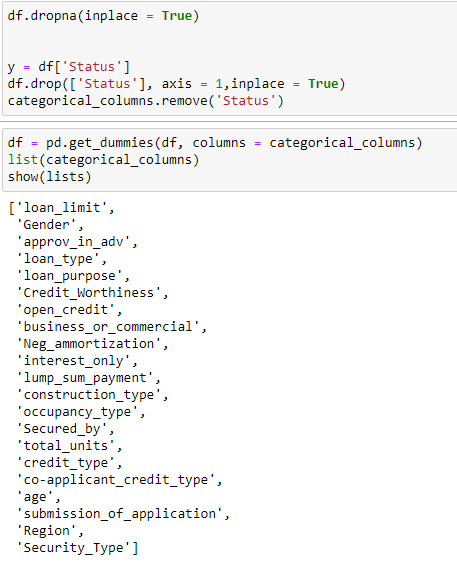


Adesso che il nostro dataset è consistente è tempo della data visualization:









Trasformiamo il dataset sostituendo alle features quelle categoriche

Immagine che contiene testo

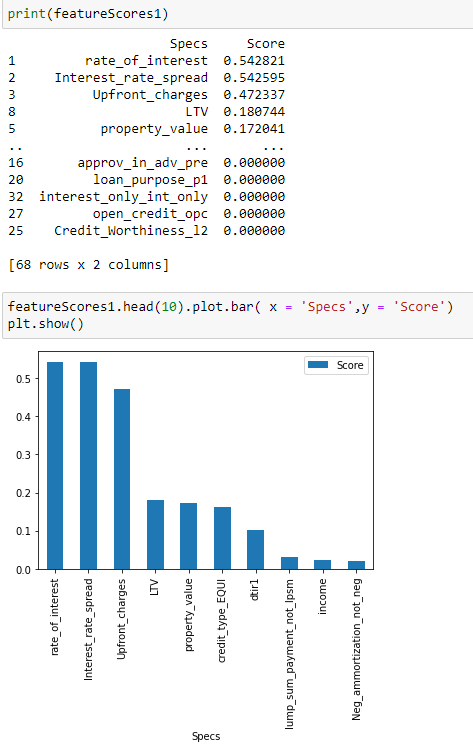
Descrizione generata automaticamente

Facciamo lo scaling ovvero il ridimenzionamento delle vvariabili in modo da permettere al modello di computare in modo più veloce

Selezioniamo le migliori features secondo la mutua informazione

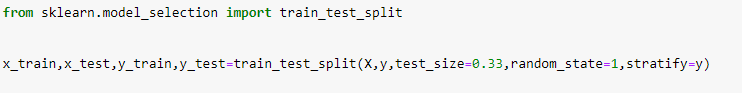
Immagine che contiene testo

Descrizione generata automaticamente





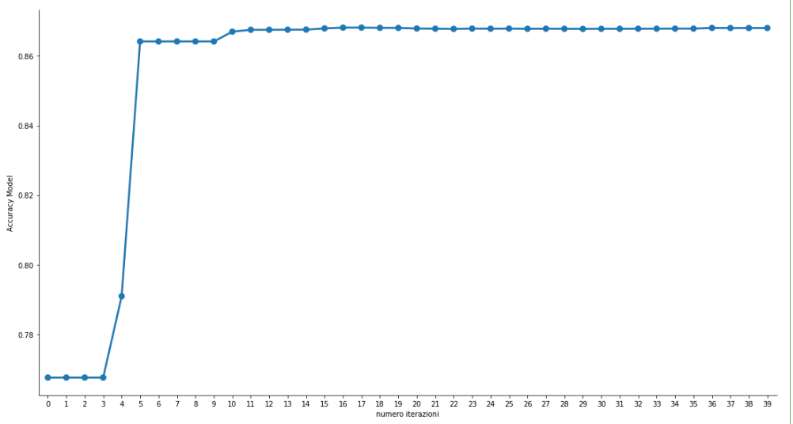
Adesso che ho scalato i valori del dataset e selezionato le features più importanti posso iniziare la fase di classificazione con gli algoritmi di ML e di splitting del dataset in training set per allenare il modello e test set per valutare il modello stesso:



Logistic regression:



Vedo dopo quante iterazioni l’algoritmo converge al valore ottimale di accuracy



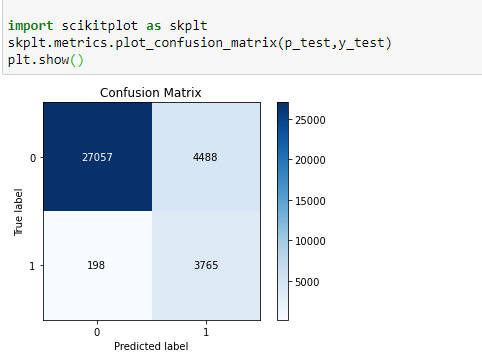
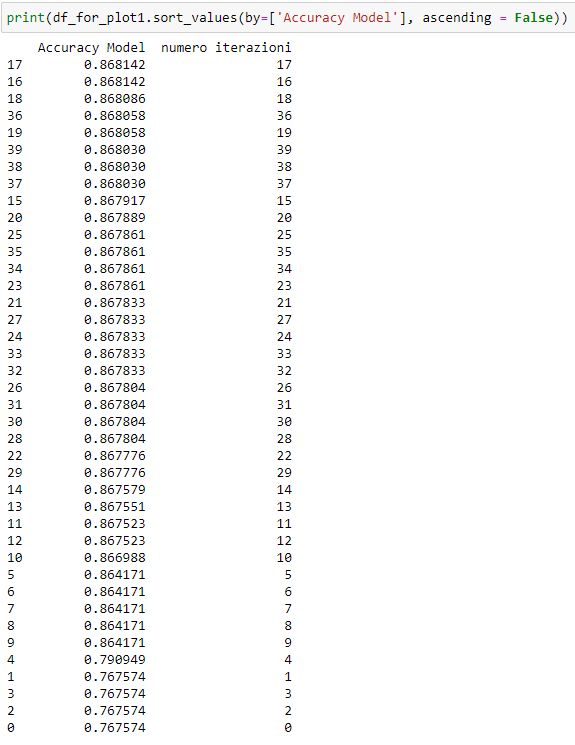


Immagine che contiene tavolo

Descrizione generata automaticamente

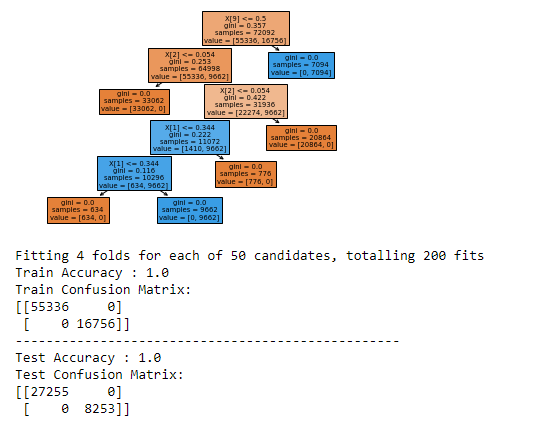
Metriche per valutare il modello

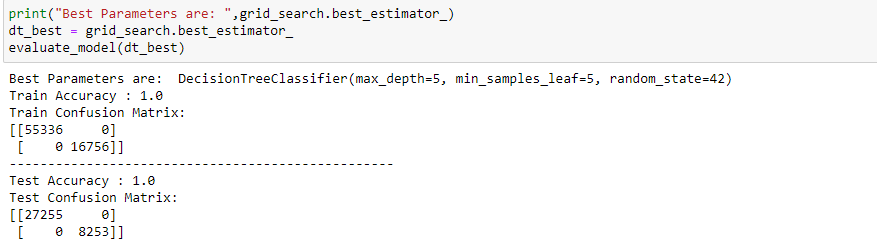


Il valore migliore di Accuracy si ha quando il modello compie 16 iterazioni

Decision Tree Classifier:

Immagine che contiene testo

Descrizione generata automaticamente



Il modello ha una accuracy del 100% classifica tutti i dati in modo corretto

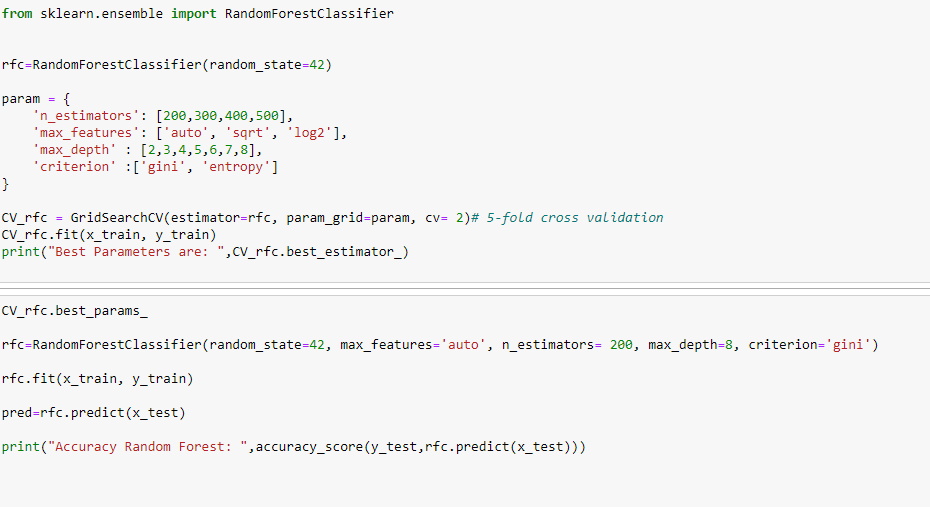
Random Forest Classifier:

Immagine che contiene testo

Descrizione generata automaticamente

Alla fine stampo le 3 accuracy

E noto che il CART e il Random Forest sono più performanti in quanto hanno accuracy 1(cioè 100%)

