



UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

CORSO DI LAUREA IN INFORMATICA

HOUSE PREDICTION AND CLASSIFICATION

GUIDA UTENTE

Caso di studio di

Ingegneria della Conoscenza a.a. 2019–2020

Il progetto è stato realizzato da:

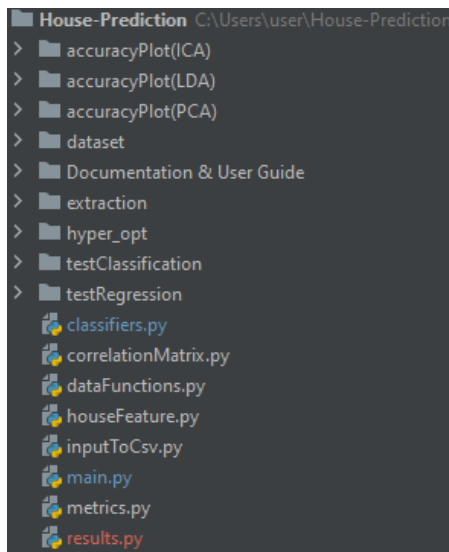
- Leonardo Bellizzi (684712)
e-mail: l.bellizzi@studenti.uniba.it
- Giuseppe Conticchio (675622)
e-mail: g.conticchio2@studenti.uniba.it
- Michelangelo Lopes (639742)
e-mail: m.lopes@studenti.uniba.it

1. AVVIO DEL PROGRAMMA

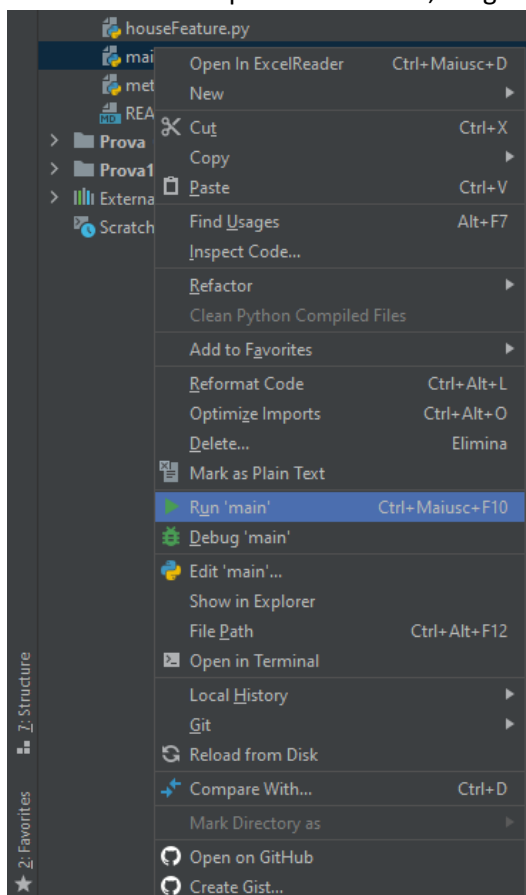
Per avviare il programma, bisogna avviare l'IDE opportuno che supporta il linguaggio Python. In questo caso è stato utilizzato PyCharm della JetBrains.



Una volta aperto PyCharm ed importato il progetto si avrà la seguente directory a disposizione all'interno del Project Explorer:

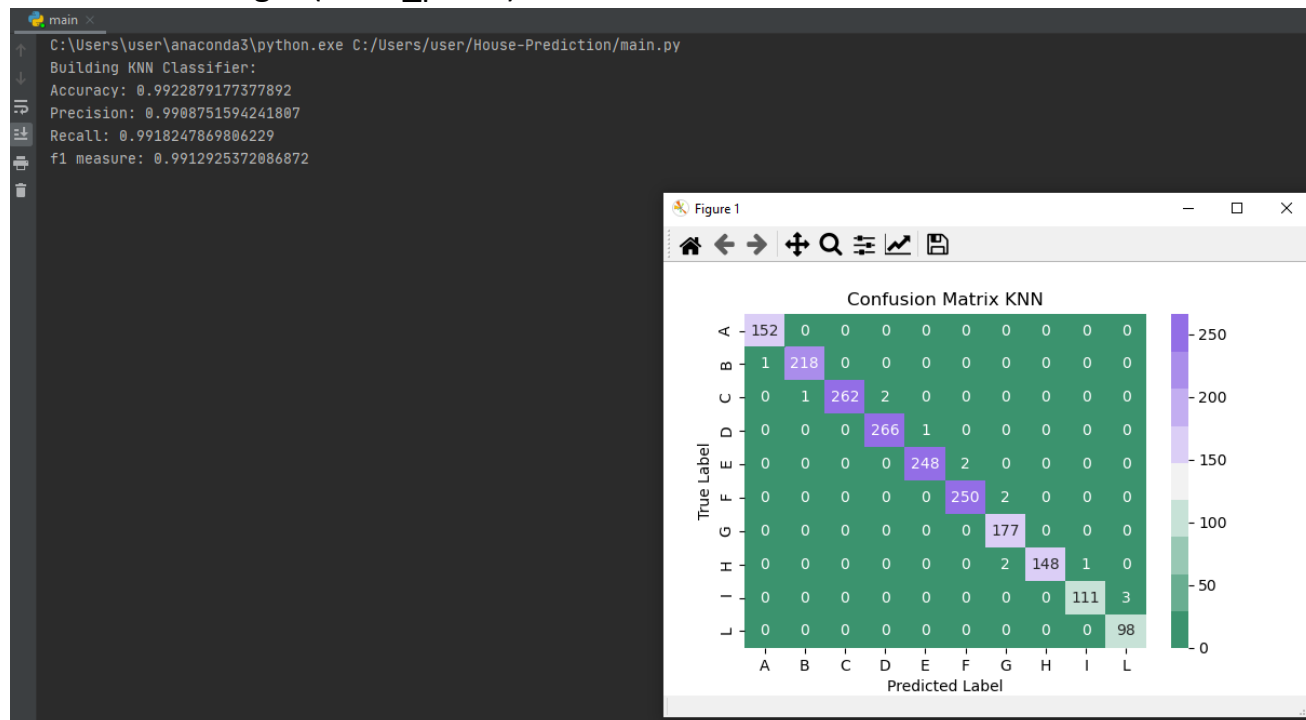


L'utente dovrà far partire il sistema, eseguendo il modulo main.py

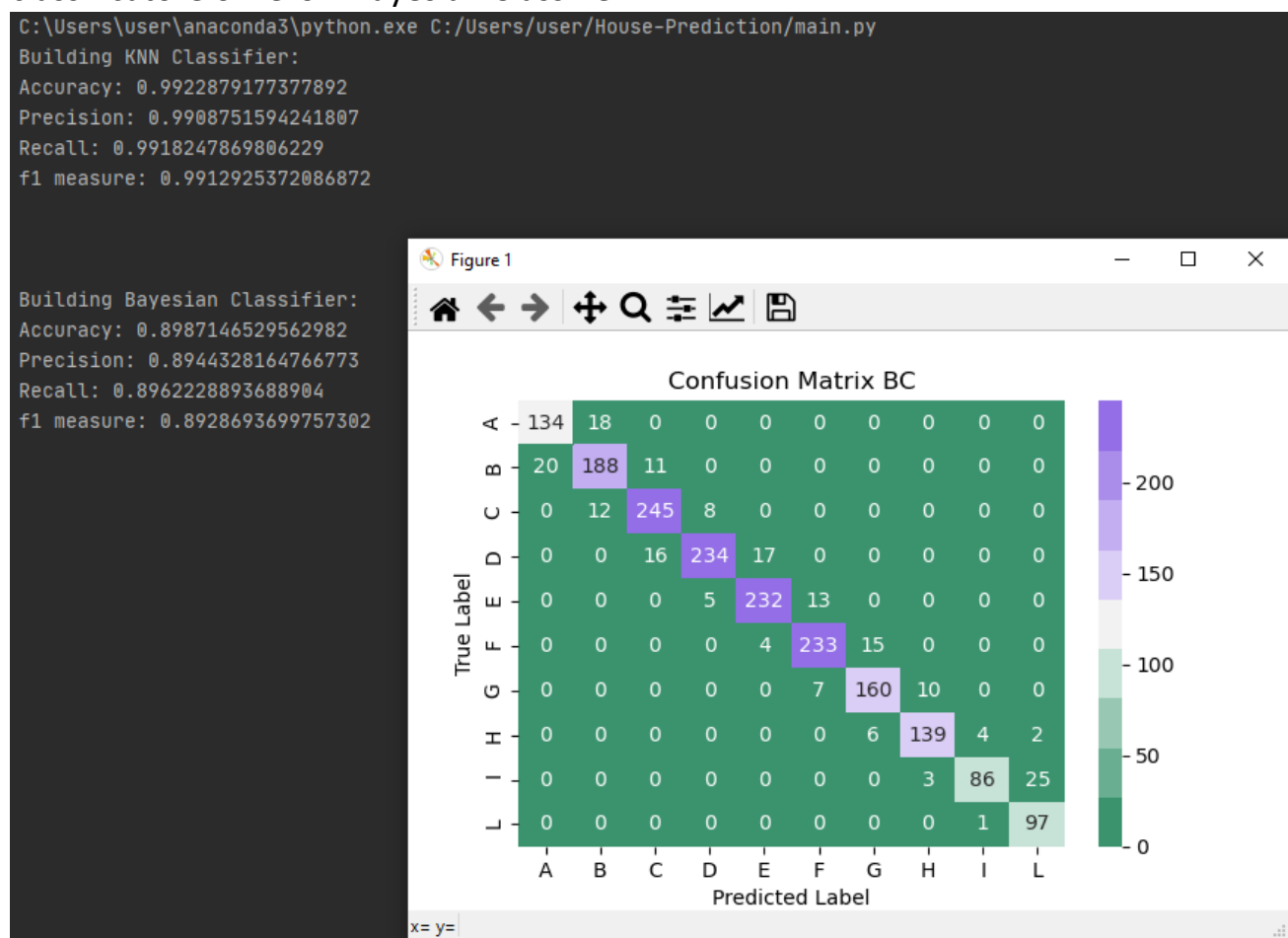


2. COSTRUZIONE DEI CLASSIFICATORI

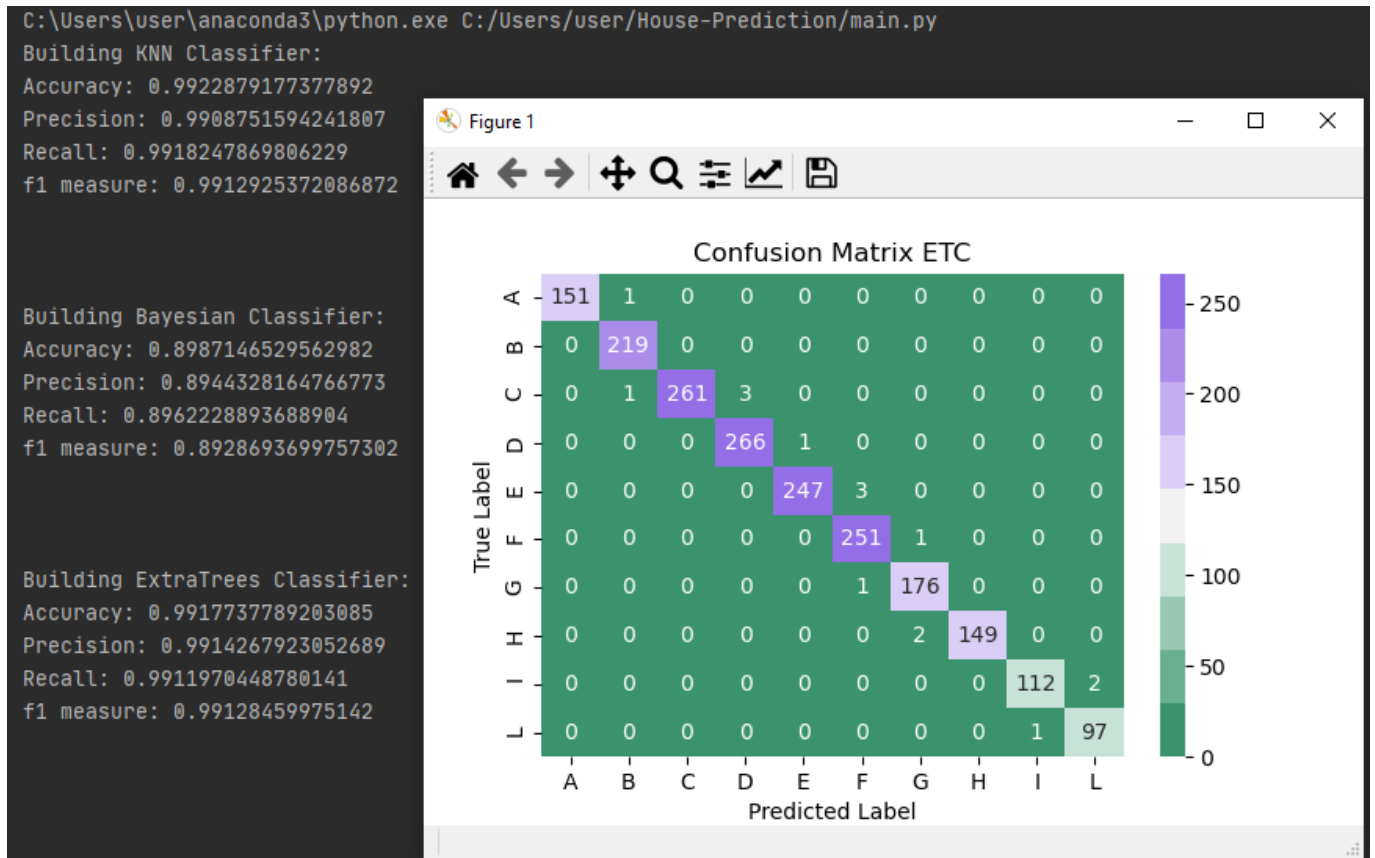
Una volta eseguito il main verrà mostrato a video, sulla console, lo score del primo classificatore ovvero il K-NN composto da Precision, Recall, Accuracy, e F1-Measure e la relativa matrice di confusione importante per capire la distribuzione dei dati appartenenti alla variabile target ('class_price')



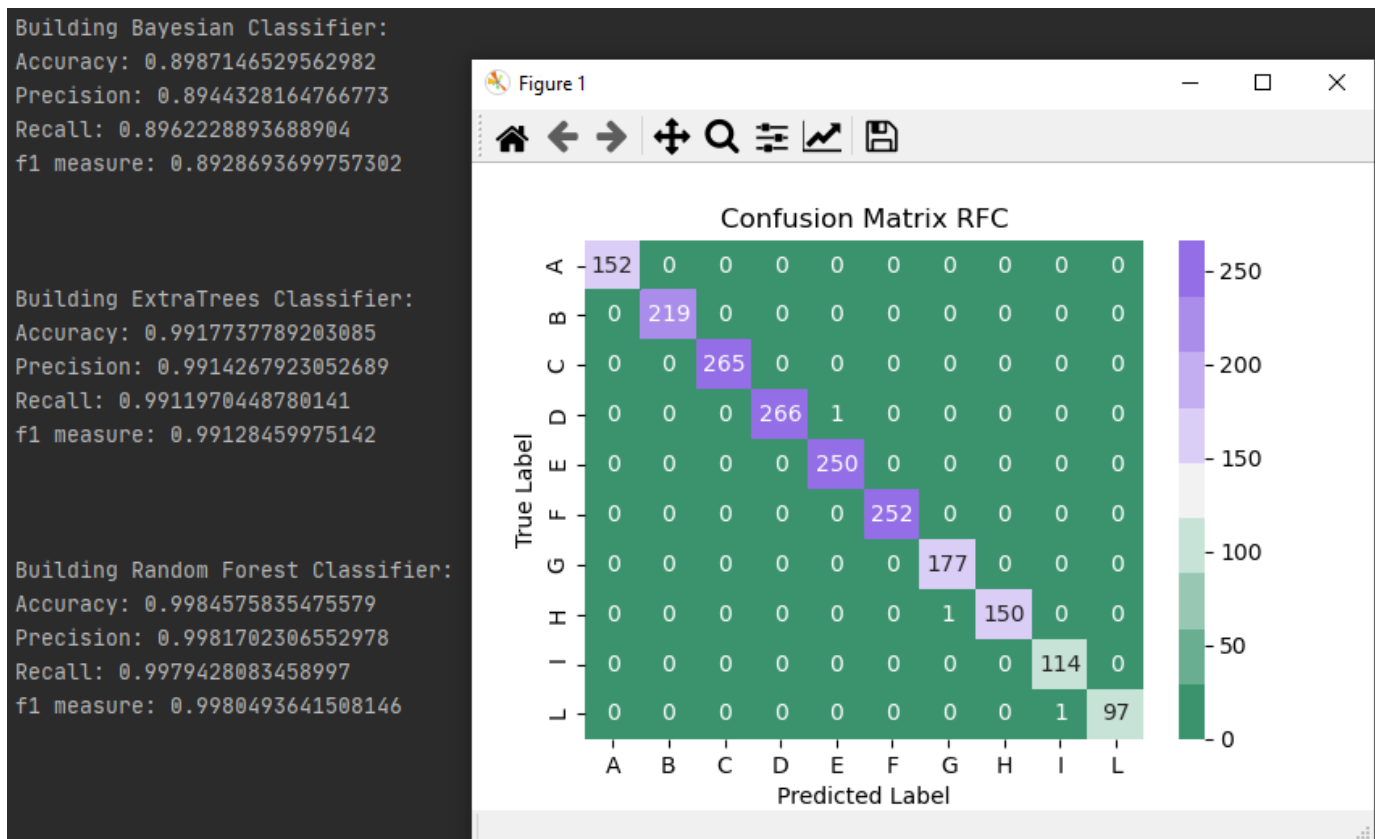
Successivamente verrà mostrato a video lo score e la matrice di correlazione del secondo classificatore ovvero il Bayesian Classifier.



Facendo procedere l'esecuzione verrà mostrato, a video, lo score del terzo classificatore ovvero l'Extra Tree Classifier e la sua corrispondente matrice di correlazione.



Continuando l'esecuzione verrà mostrato, a video, lo score del quarto ed ultimo classificatore previsto ovvero il Random Forest Classifier e la sua corrispondente matrice di correlazione.

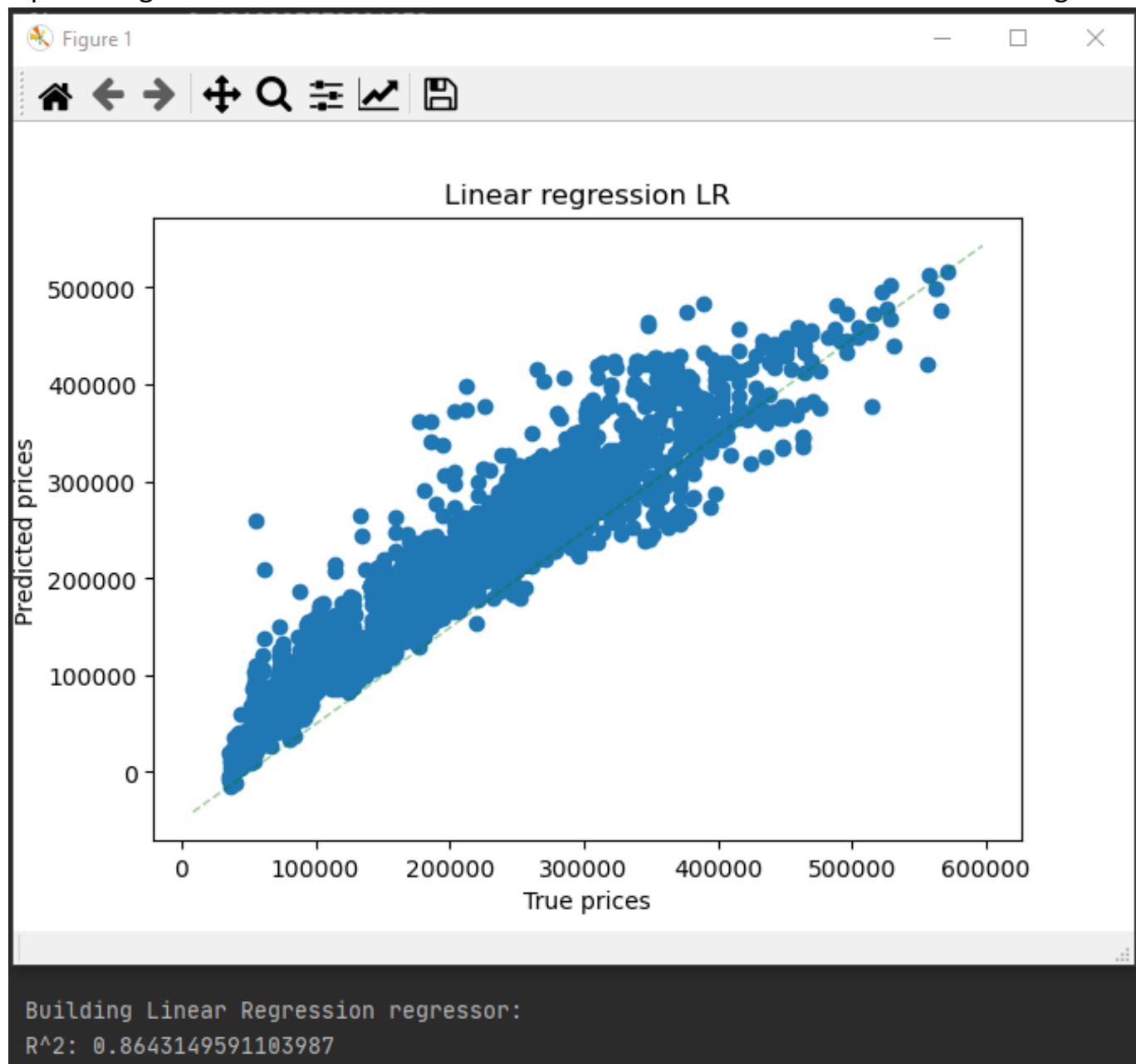


3. COSTRUZIONE DEI REGRESSORI

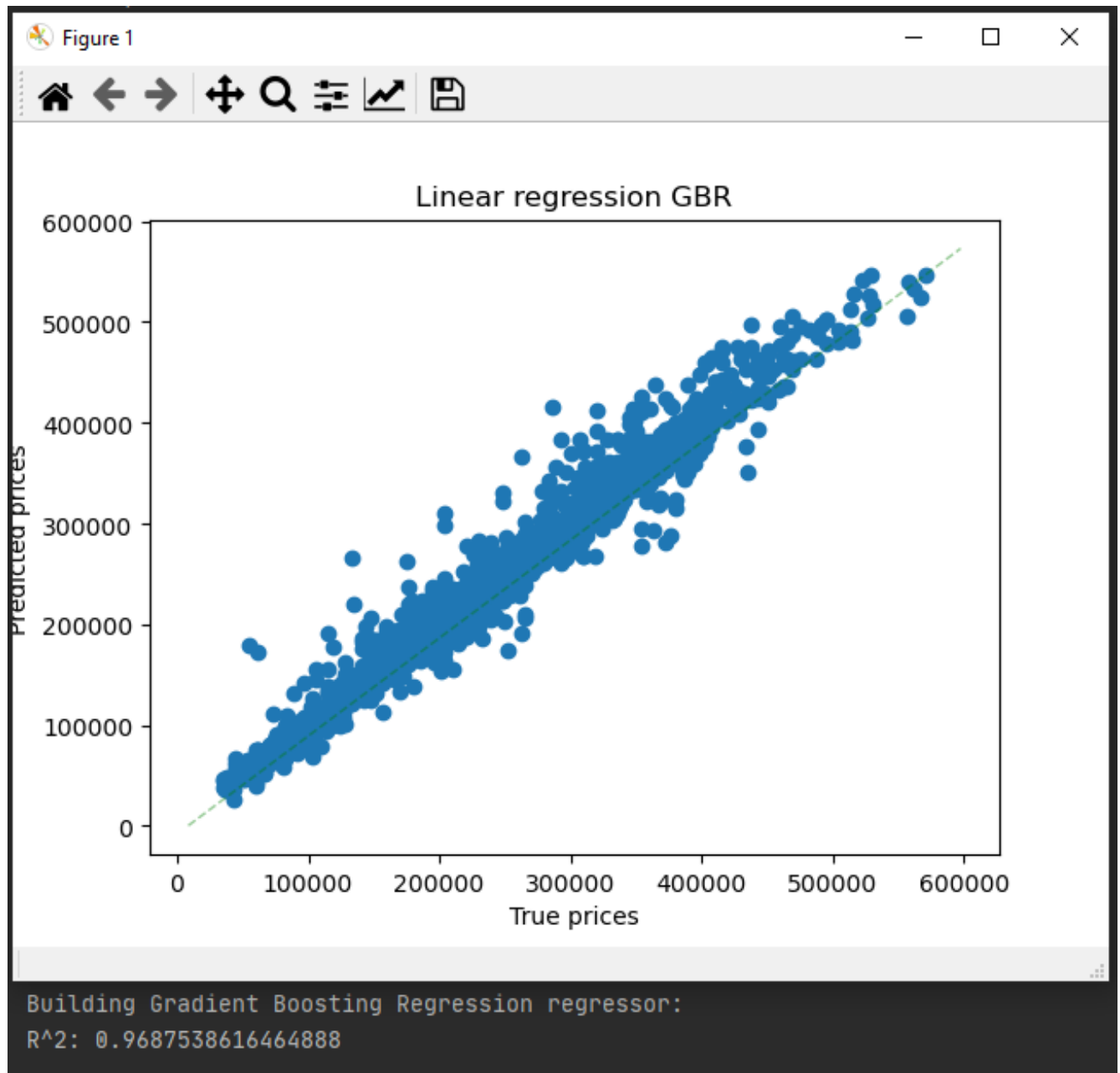
Dopo che tutti i classificatori sono stati costruiti ed eseguiti correttamente dal sistema, sempre a video, verranno mostrati i regressori utilizzati e i loro grafici utili per capire come si distribuisce la relazione tra i prezzi veri e propri presenti nel dataset e i prezzi predetti. Più i “puntini blu” sono vicini alla retta e più quel regressore può essere considerato accurato nelle sue previsioni.

Nel caso dei regressori, lo score è dato dalla metrica R^2 .

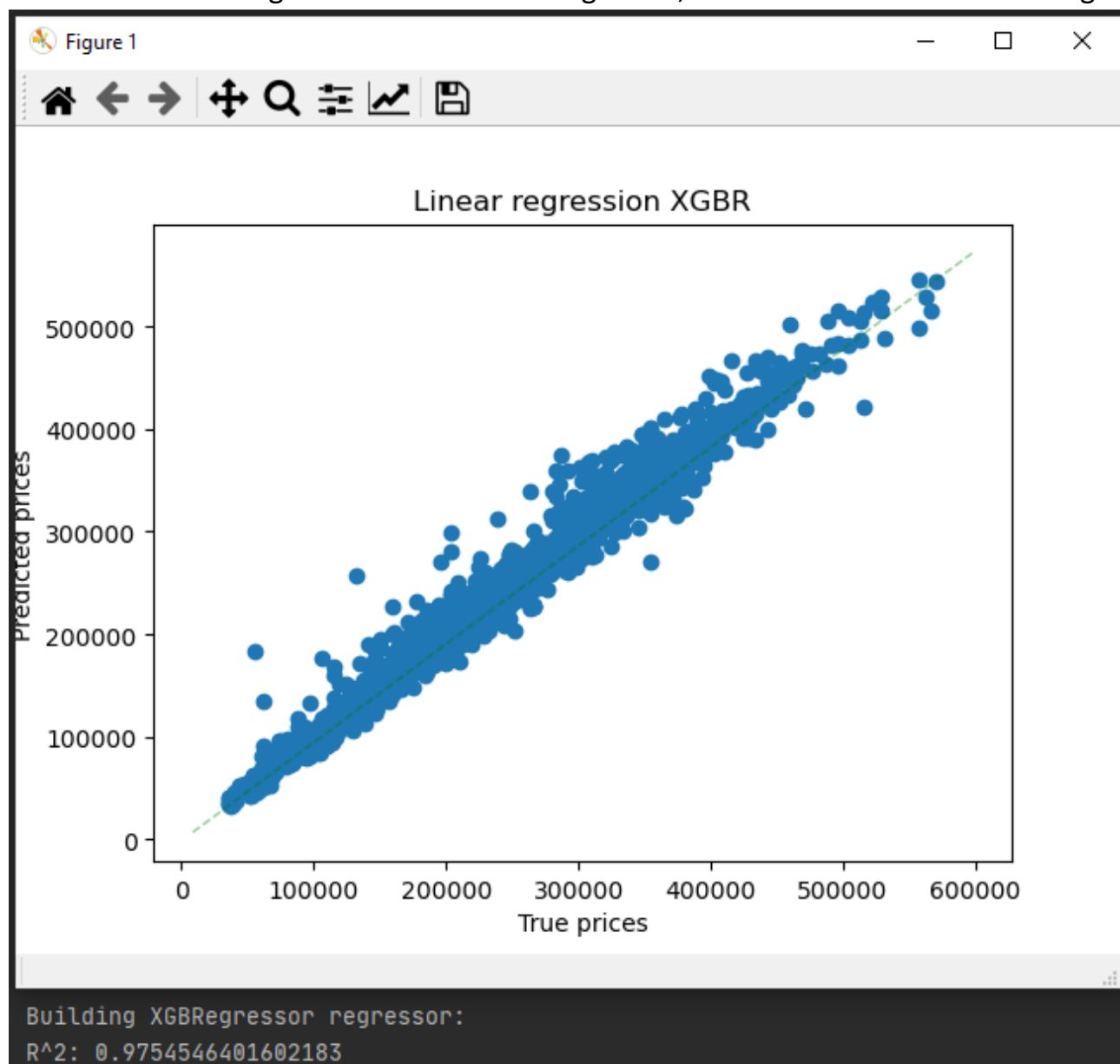
Il primo regressore in ordine di esecuzione ad essere mostrato a video è il Linear Regressor.



Il secondo regressore invece è il Gradient Boosting Regressor



Il terzo ed ultimo regressore è l'XGBoost Regressor, variante del Gradient Boosting



4. LA SCELTA DELL'UTENTE

Dopo che è stata avviata l'esecuzione e che tutti i classificatori ed i regressori sono stati mostrati a video con i loro relativi score ed i loro relativi grafici, sulla console verrà mostrato a video un menù che consente all'utente di scegliere tra diverse opzioni ovvero:

- 1) Classificare due immobili presenti all'interno, rispettivamente, di due csv di prova presenti all'interno della cartella **testClassification** adibiti alla fase di test dei classificatori;
- 2) Predire il prezzo di due immobili presenti all'interno, rispettivamente, di due csv di prova presenti all'interno della cartella **testRegression** adibiti alla fase di test dei regressori;
- 3) Inserire da tastiera i valori delle features. In questo modo l'utente non sarà limitato ai csv di test forniti dal sistema ma potrà verificare in tempo reale l'efficacia dei classificatori utilizzati;
- 4) Inserire da tastiera i valori delle features. In questo modo l'utente non sarà limitato ai csv di test forniti dal sistema ma potrà verificare in tempo reale l'efficacia dei regressori utilizzati;
- 5) Terminare l'esecuzione del sistema.

```
Possible choices:
1) Classification with csv
2) Regression with csv
3) Classification with input values
4) Regression with input values
5) Quit
What would you like to do? Put your choice:
```

L'utente dovrà inserire il numero corrispondente all'azione che vuole intraprendere, seguendo il menù sopra indicato.

4.1.1 CLASSIFICATION WITH CSV

Nel momento in cui l'utente sceglie la prima opzione, a video, comparirà la seguente stringa: *"Insert house's path:"* e l'utente dovrà inserire il percorso corretto per arrivare al csv di test che vuole effettivamente provare (i.e.:C:\Utente\User\...\TestClassification\testClassificatore.csv).

```
What would you like to do? Put your choice: 1
Insert house's path: C:\Users\user\House-Prediction\testClassification\testClassificatore.csv
Loading house...
Csv loaded
```

Una volta inserito il percorso corretto il csv di test verrà caricato dal sistema, il quale passerà le features dell'immobile ai classificatori, i quali dovranno classificarlo nella fascia di prezzo corretta.

```
[278945.0, 2007.0, 2008.0, 2.0, 978.0, 4.0, 75.0, 104.0, 12.5, 12.5, 2.0, 3.0, 5.0, 2.0, 1.0, 1.0, 2.0, 2.0, 0.0, 3.0, 0.0, 2.0, 1.0, 5.0, 6.0, 6.0]

Prediction KNN: ('F', 1.0)
Prediction Bayes: ('F', 0.8034337681562077)
Prediction Extra Tree: ('F', 0.6799555435601865)
Prediction Random Forest: ('F', 1.0)
```

Nell'esempio sopra riportato, l'immobile è della fascia di prezzo F, che è la fascia di prezzo corretta.

Nel momento in cui la classificazione termina comparirà a video la seguente stringa:

```
Have you another csv with example of house to predict class price? {y/n}
```

Se, l'utente vuole terminare l'esecuzione del sistema, dovrà digitare **n** altrimenti dovrà digitare **y** e ricomparirà a video il menù iniziale.

4.1.2 REGRESSION WITH CSV

Nel momento in cui l'utente sceglie la seconda opzione, a video, comparirà come nel caso della classificazione la stringa per l'inserimento del percorso per poter raggiungere il csv di test per i regressori.

```
Possible choices:
1) Classification with csv
2) Regression with csv
3) Classification with input values
4) Regression with input values
5) Quit
What would you like to do? Put your choice: 2
Insert house's path: C:\Users\user\House-Prediction\testRegression\testRegressore.csv
Loading house...
Csv loaded

[2006.0, 2007.0, 8.0, 814.0, 3.0, 111.0, 184.0, 7.5, 12.5, 3.0, 3.0, 0.0, 2.0, 1.0, 1.0, 1.0, 1.0, 0.0, 3.0, 2.0, 2.0, 2.0, 5.0, 6.0, 9.0]

Prediction Linear Regression: (0.8643149591103987, 'Price: 122649.11')
Prediction Gradient Boosting Regression: (0.9682244827915063, 'Price: 141591.17')
Prediction XGBoost Regression: (0.9754546401602183, 'Price: 152371.88')

Have you another csv with example of house to predict class price? {y/n} |
```

Il prezzo predetto per un immobile con le caratteristiche, riportato nell'esempio precedente, è quello dato dall'XGBoost il quale è il regressore più preciso.

4.1.3 CLASSIFICATION WITH INPUT VALUES

Nel momento in cui l'utente sceglie la terza opzione, a video, verrà mostrato il primo *“Insert values”* che riguarderà il prezzo di vendita. Per inserire da tastiera tutti i valori basterà inserire i valori richiesti e premere Invio. Arrivati alla fine, l'immobile “creato” attraverso l'immissione dei dati in tempo reale, verrà classificato.

```
1) Classification with csv
2) Regression with csv
3) Classification with input values
4) Regression with input values
5) Quit
What would you like to do? Put your choice: 3
Insert value for SalePrice: 341598
Insert value for YearBuilt: 2010
Insert value for YrSold: 2012
Insert value for MonthSold: 10
Insert value for Size(sqft): 1120
Insert value for Floor: 5
Insert value for N_Parkinglot(Ground): 10
Insert value for N_Parkinglot(Basement): 116
Insert value for TimeToBusStop: 7.5
Insert value for TimeToSubway: 12.5
Insert value for N_APT: 8
Insert value for N_manager: 3
Insert value for N_elevators: 10
Insert value for N_FacilitiesNearBy(PublicOffice): 2
Insert value for N_FacilitiesNearBy(Hospital): 4
Insert value for N_FacilitiesNearBy(Dpartmentstore): 0
Insert value for N_FacilitiesNearBy(Mall): 1
Insert value for N_FacilitiesNearBy(ETC): 2
Insert value for N_FacilitiesNearBy(Park): 2
Insert value for N_SchoolNearBy(Elementary): 1
Insert value for N_SchoolNearBy(Middle): 0
Insert value for N_SchoolNearBy(High): 0
Insert value for N_SchoolNearBy(University): 1
Insert value for N_FacilitiesInApt: 0
Insert value for N_FacilitiesNearBy(Total): 10
Insert value for N_SchoolNearBy(Total): 4
Csv created

[341598.0, 2010.0, 2012.0, 10.0, 1120.0, 5.0, 98.0, 116.0, 7.5, 12.5, 8.0, 3.0, 10.0, 2.0, 2.0, 0.0, 1.0, 2.0, 2.0, 1.0, 0.0, 2.0, 1.0, 5.0, 15.0, 4.0]

Prediction KNN: ('H', 1.0)
Prediction Bayes: ('H', 0.9716775757684663)
Prediction Extra Tree: ('H', 0.7991765964760836)
Prediction Random Forest: ('H', 1.0)

Have you another csv with example of house to predict class price? {y/n} |
```

In questo esempio, l'immobile inserito, è stato classificato all'interno della fascia di prezzo H. I valori immessi da tastiera verranno salvati in un csv all'interno della cartella TestClassification.

4.1.4 REGRESSION WITH INPUT VALUES

Nel momento in cui l'utente sceglie la quarta opzione, a video, verrà mostrato il primo *"Insert values"* che riguarderà l'anno di vendita e non il prezzo di vendita poiché è quello che deve essere predetto dai regressori. Per inserire da tastiera tutti i valori basterà inserire i valori richiesti e premere Invio. Arrivati alla fine, il prezzo dell'immobile "creato" attraverso l'immissione dei dati in tempo reale, verrà predetto.

```
Possible choices:
1) Classification with csv
2) Regression with csv
3) Classification with input values
4) Regression with input values
5) Quit
What would you like to do? Put your choice: 4
Insert value for YearBuilt: 1998
Insert value for YrSold: 2007
Insert value for MonthSold: 2
Insert value for Size(sqf): 784
Insert value for Floor: 2
Insert value for N_Parkinglot(Ground): 45
Insert value for N_Parkinglot(Basement): 79
Insert value for TimeToBusStop: 12.5
Insert value for TimeToSubway: 12.5
Insert value for N_APT: 4
Insert value for N_manager: 5
Insert value for N_elevators: 7
Insert value for N_FacilitiesNearBy(PublicOffice): 1
Insert value for N_FacilitiesNearBy(Hospital): 1
Insert value for N_FacilitiesNearBy(Dpartmentstore): 2
Insert value for N_FacilitiesNearBy(Mall): 0
Insert value for N_FacilitiesNearBy(ETC): 5
Insert value for N_FacilitiesNearBy(Park): 0
Insert value for N_SchoolNearBy(Elementary): 1
Insert value for N_SchoolNearBy(Middle): 1
Insert value for N_SchoolNearBy(High): 1
Insert value for N_SchoolNearBy(University): 0
Insert value for N_FacilitiesInApt: 1
Insert value for N_FacilitiesNearBy(Total): 7
Insert value for N_SchoolNearBy(Total): 3
Csv loaded

[1998.0, 2007.0, 2.0, 784.0, 2.0, 45.0, 79.0, 12.5, 12.5, 4.0, 5.0, 7.0, 1.0, 1.0, 2.0, 0.0, 5.0, 0.0, 1.0, 1.0, 1.0, 0.0, 3.0, 7.0, 3.0]

Prediction Linear Regression: (0.86443149591103987, 'Price: 221005.80')
Prediction Gradient Boosting Regression: (0.9682244827915063, 'Price: 134968.32')
Prediction XGBoost Regression: (0.9754546401602183, 'Price: 69377.03')
```

Nell'esempio è stato "creato" un immobile di scarsa qualità poiché è obsoleto ed ha pochi servizi sia al suo interno che nelle vicinanze, infatti l'XGBoost prevede un prezzo sotto i 70.000\$ il che è molto plausibile data la precisione del regressore. I valori immessi da tastiera verranno salvati in un csv all'interno della cartella TestRegression.

4.1.5 QUIT

Se l'utente dovesse scegliere la quinta opzione, il sistema terminerà la propria esecuzione.

```
Possible choices:
1) Classification with csv
2) Regression with csv
3) Classification with input values
4) Regression with input values
5) Quit
What would you like to do? Put your choice: 5
Process finished with exit code 0
```