# Valutazione del danno alluvionale mediante strumenti di Statistica Multivariata e Machine Learning


Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

SAPIENZA UNIVERSITÀ DI ROMA

**Corso di Laurea in Statistica Gestionale**

Relatore: **Prof. Maurizio Vichi**

Relatori Esterni: **Dr. Christian Natale Gencarelli**
**Dr. Simone Sterlacchini**

Candidato: **Montanari Simone**

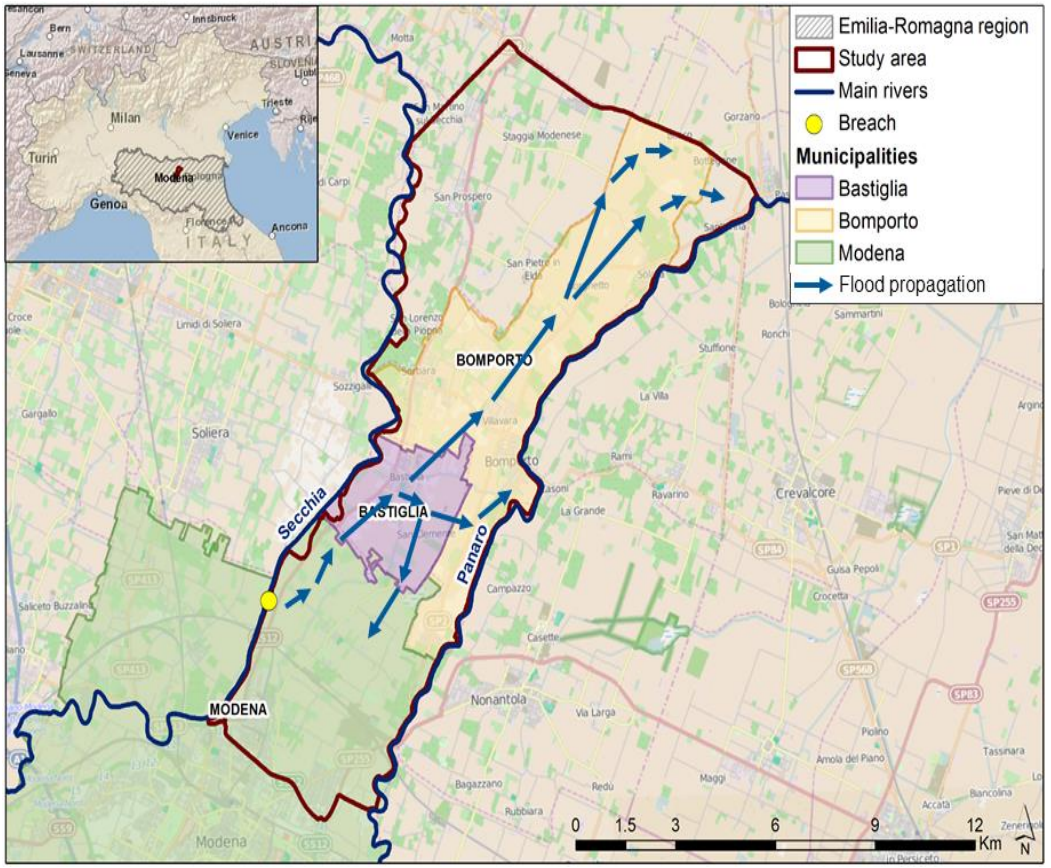

Anno Accademico 2020/2021

Roma, 23 Maggio 2022

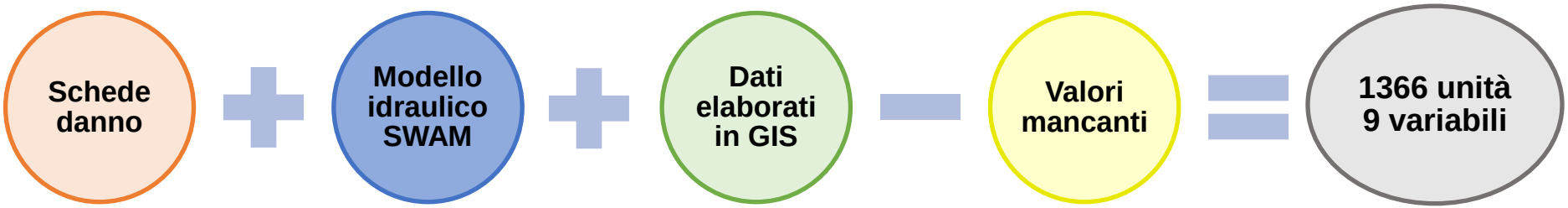
# Contesto

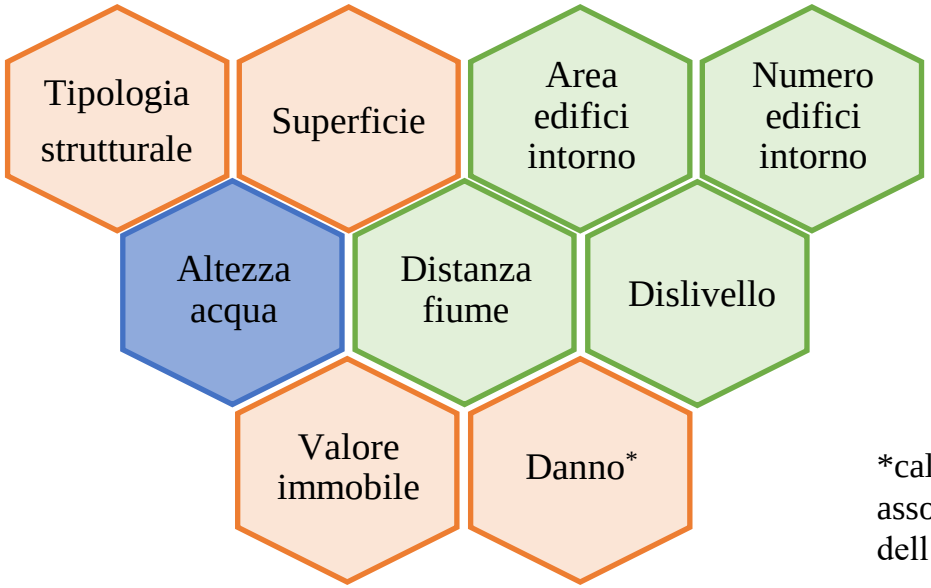
Alluvione del 19 gennaio 2014 nei comuni di Bastiglia e Bomporto (Modena)

- 52 $km^2$ invasi da 37 milioni di $m^3$ d'acqua in meno di 30 ore
- Circa 1500 sfollati
- Circa 500 milioni di euro di danni



# I dati



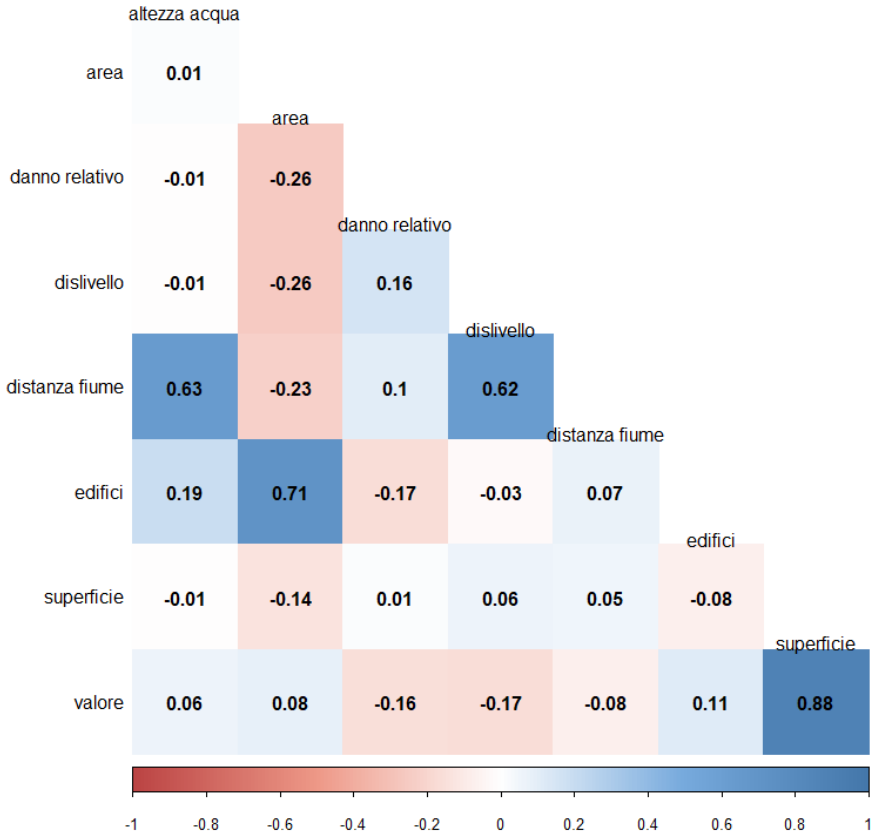




*calcolato come danno assoluto/valore dell'immobile

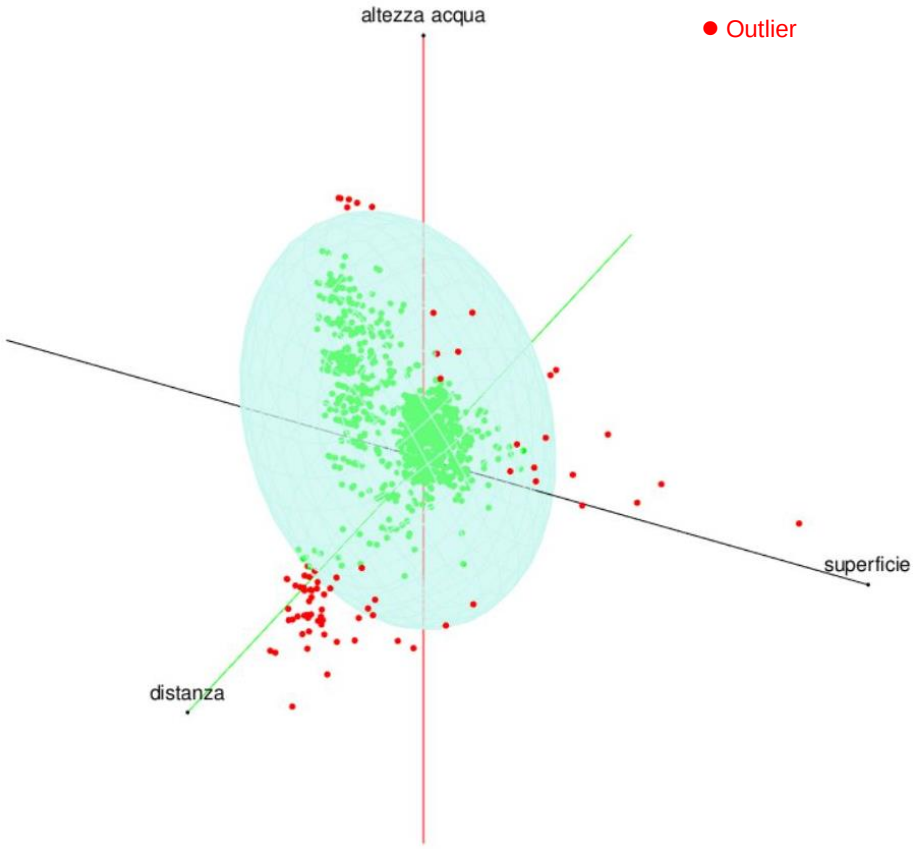
# I dati

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | altezza dell'acqua | distanza dal fiume | dislivello rispetto gli argini | tipologia strutturale | superficie | area edifici | # edifici | valore dell'immobile | danno relativo |
| 2 | 0.39 | 1106.13 | 10.58 | muratura | 58.00 | 2222.00 | 7.00 | 68150.00 | 0.00 |
| 3 | 0.39 | 1106.13 | 10.58 | muratura | 70.00 | 2222.00 | 7.00 | 82250.00 | 0.00 |
| 4 | 0.55 | 1103.58 | 10.99 | muratura | 40.00 | 1155.00 | 6.00 | 47000.00 | 0.23 |
| 5 | 0.44 | 1113.50 | 10.76 | muratura | 30.00 | 2409.00 | 5.00 | 35250.00 | 0.18 |
| 6 | 0.55 | 1109.58 | 10.61 | cemento | 200.00 | 1155.00 | 6.00 | 235000.00 | 0.03 |
| 7 | 0.77 | 900.32 | 10.80 | muratura | 110.00 | 2135.00 | 9.00 | 129250.00 | 0.26 |
| 8 | 0.67 | 979.88 | 11.00 | cemento | 120.00 | 1041.00 | 5.00 | 141000.00 | 0.11 |
| 9 | 0.73 | 901.15 | 10.81 | cemento | 150.00 | 1695.00 | 8.00 | 176250.00 | 0.02 |
| 10 | 0.81 | 992.80 | 11.14 | cemento e muratura | 60.00 | 2771.00 | 7.00 | 70500.00 | 0.00 |
| 11 | 0.59 | 951.41 | 10.80 | muratura | 84.00 | 2098.00 | 10.00 | 98700.00 | 0.00 |
| 12 | 0.59 | 951.41 | 10.80 | cemento | 50.00 | 2098.00 | 10.00 | 58750.00 | 0.01 |
| 13 | 0.59 | 951.41 | 10.80 | muratura | 80.00 | 2098.00 | 10.00 | 94000.00 | 0.01 |
| 14 | 0.59 | 951.41 | 10.80 | cemento e muratura | 80.00 | 2098.00 | 10.00 | 94000.00 | 0.01 |

# Statistica descrittiva

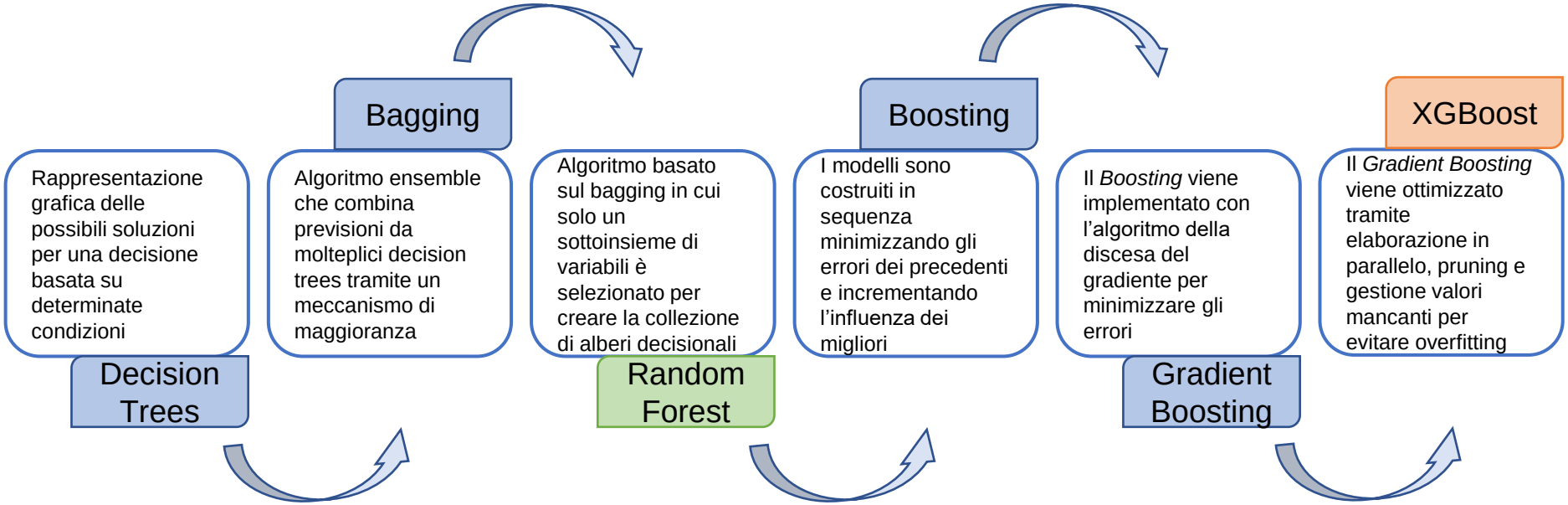



Correlazioni




Valori anomali (288 su 1366)

# Regressione: Alberi decisionali



# Regressione: Modellazione

**Random Forest**

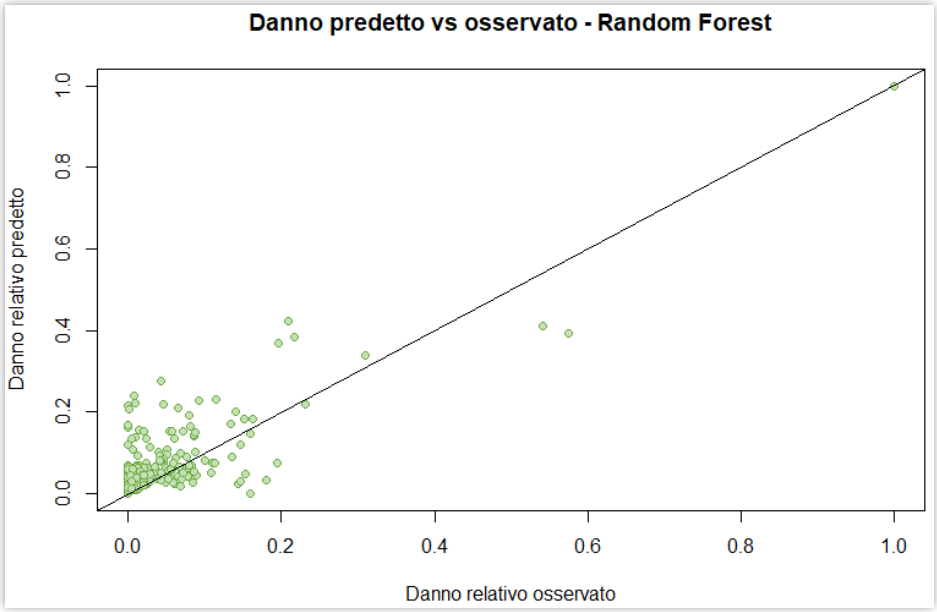
| # | % Train | Outliers | ntree | nodesize | mtry | Folds CV | RMSE | | $R^2$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Train | Test | Train | Test |
| 1 | 80% | ✓ | 500 | 5 | 3 | - | 0,09 | 0,13 | 0,84 | 0,50 |
| 2 | 80% | - | 500 | 5 | 3 | 5 | 0,10 | 0,14 | 0,82 | 0,10 |
| 3 | 70% | ✓ | 300 | 6 | 2 | 5 | 0,11 | 0,14 | 0,77 | 0,42 |
| 4 | 85% | ✓ | 600 | 8 | 2 | 5 | 0,10 | 0,12 | 0,78 | 0,64 |

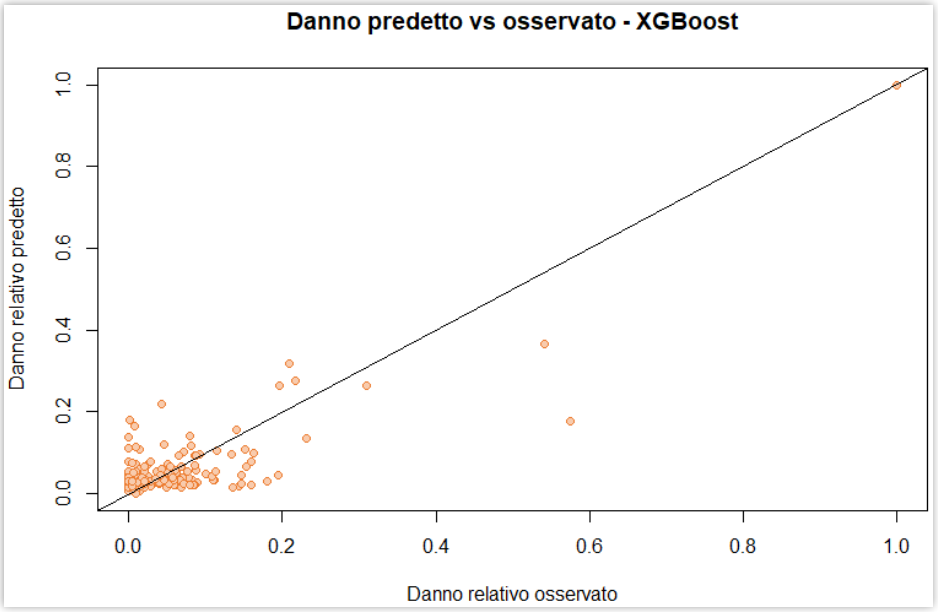
**XGBoost**

| # | % Train | Outliers | eta | gamma | sub sample | max depth | Folds CV | RMSE | | $R^2$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Train | Test | Train | Test |
| 1 | 80% | ✓ | 0,05 | 0,03 | 1 | 6 | - | 0,08 | 0,13 | 0,82 | 0,51 |
| 2 | 80% | - | 0,02 | 0 | 0,7 | 4 | 5 | 0,11 | 0,15 | 0,72 | 0,09 |
| 3 | 70% | ✓ | 0,02 | 0 | 0,7 | 4 | 5 | 0,10 | 0,16 | 0,72 | 0,28 |
| 4 | 85% | ✓ | 0,04 | 0 | 0,7 | 4 | 5 | 0,10 | 0,11 | 0,75 | 0,67 |

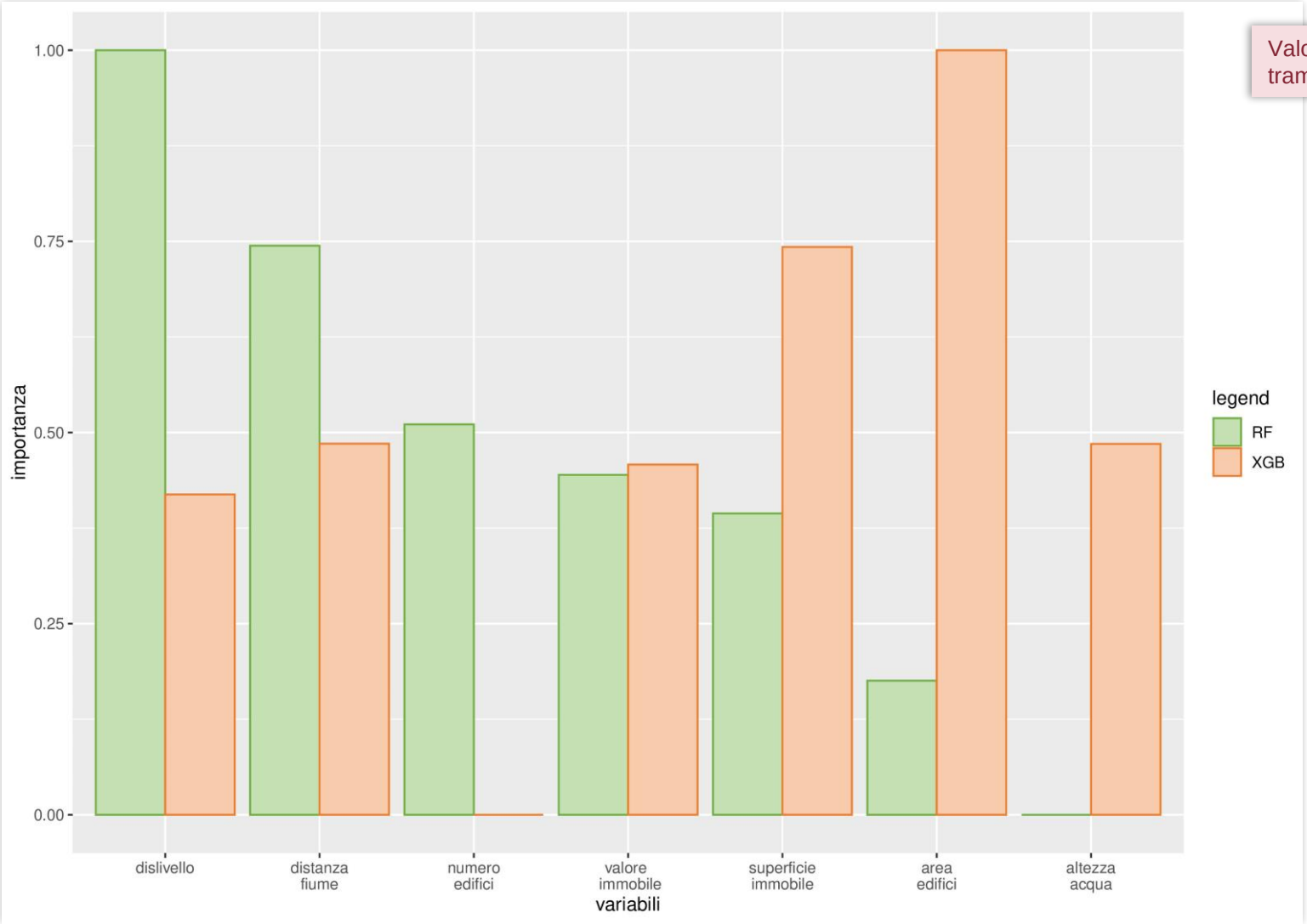
# Regressione: Risultati

| | $R^2$ | Random Forest | XGBoost |
|---|---|---|---|
| 85% | Train set | 0,78 | 0,75 |
| 15% | Test set | 0,64 | 0,67 |



# Regressione: Variabili




Valori scalati tramite min-max

# Regressione: Regole decisionali

Trovate 10.000 regole per il Random Forest, 1.359 per XGBoost

**Random Forest**

Distanza fiume <= 4.360 mt & dislivello <= 11 mt & area > 495 mq & edifici > 1,50 & valore immobile > 53.400 €

Previsione: 0,08

1,05 mt < altezza acqua <= 1,30 mt & dislivello > 11 mt

Previsione: 0,88

**XGBoost**

Superficie <= 220 mq & 30.500 € < valore immobile <= 370.000 €

Previsione: 0,10

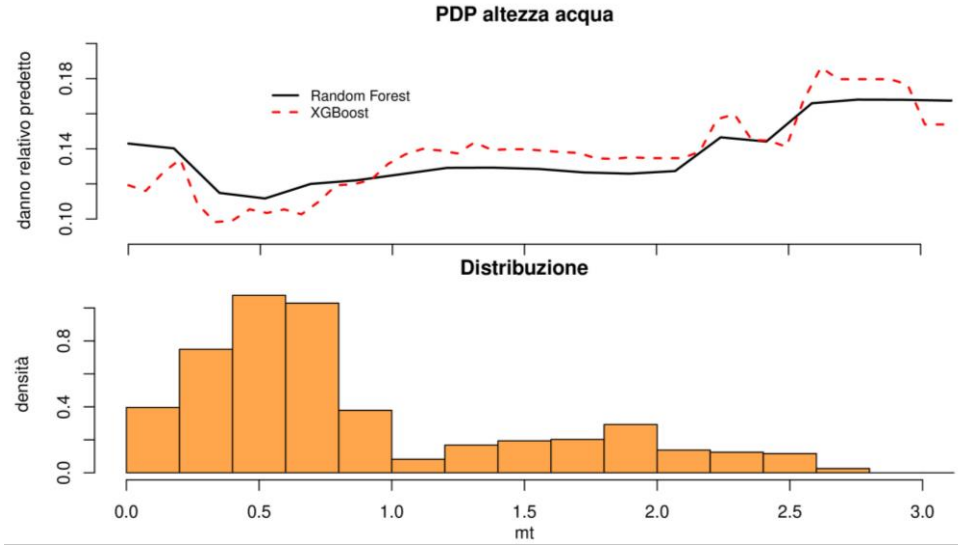
Superficie <= 68,5 mq & altezza acqua > 0,13 mt & area <= 85 mq

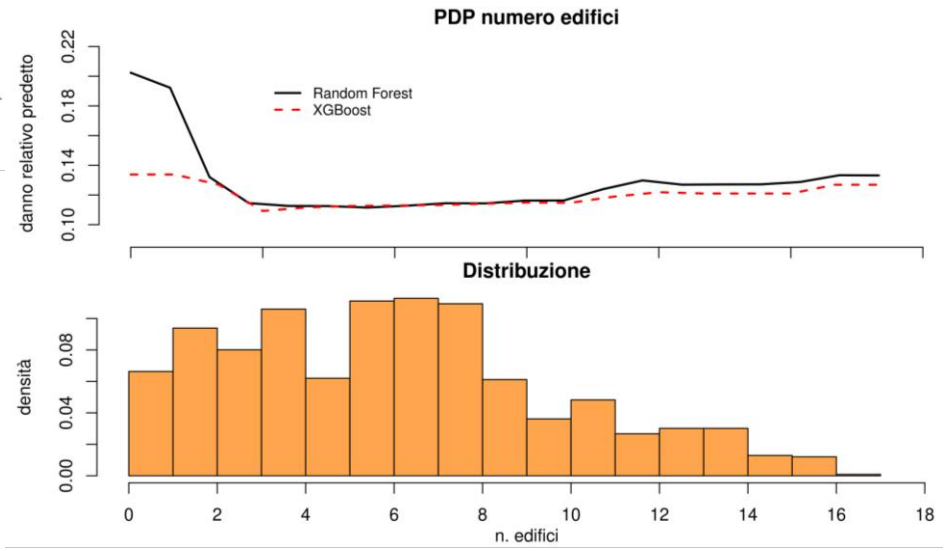
Previsione: 0,86

Esempio di interpretazione:
Gli alberi del modello Random Forest che rispettano la prima condizione restituiscono come output una previsione in media pari a 0,08, ossia un danno pari all' 8% del valore dell'immobile

# Regressione: PDP

Il *Partial Dependence* mostra l'effetto marginale di una variabile sul risultato previsionale medio di un modello di machine learning



# Clustering: Metodo

**Approccio**

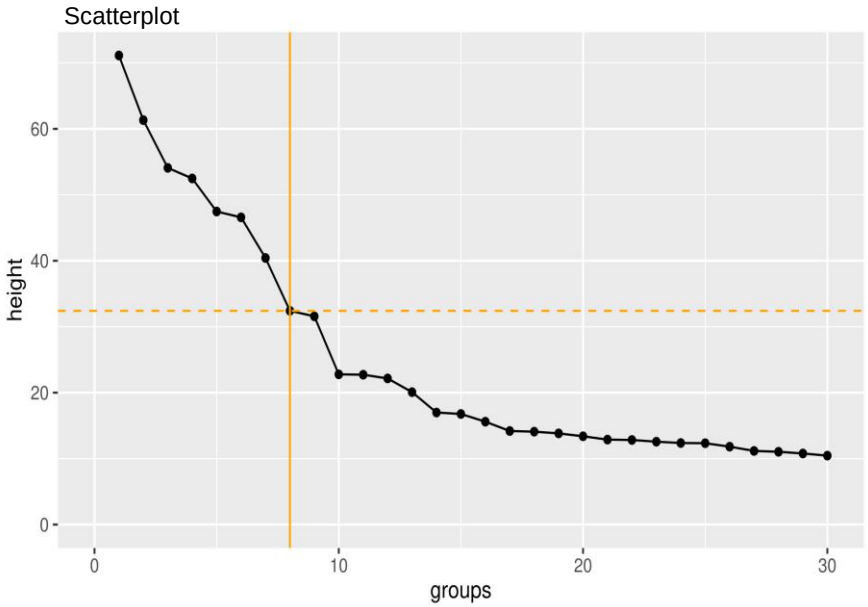
Gerarchico Agglomerativo

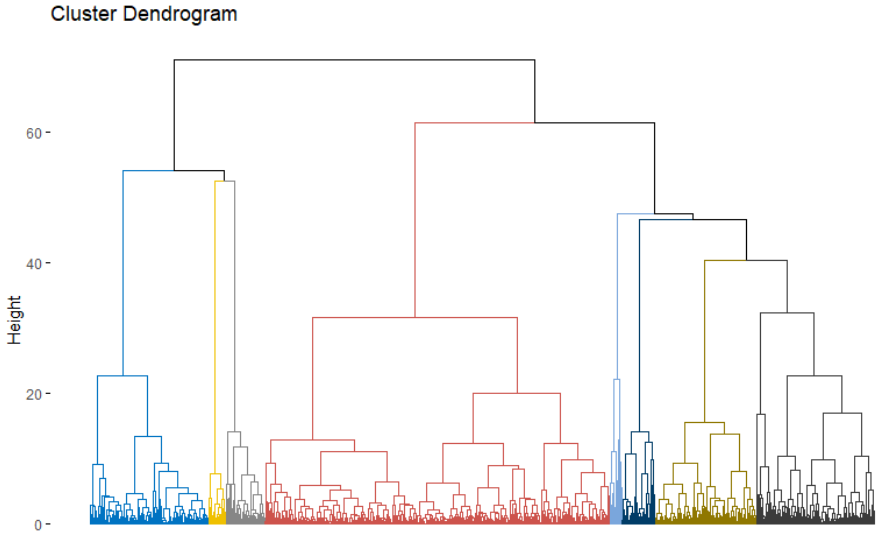
**Metrica**

Metodo di Ward (minimizzazione varianza intra-cluster)

**Risultato**

8 Cluster

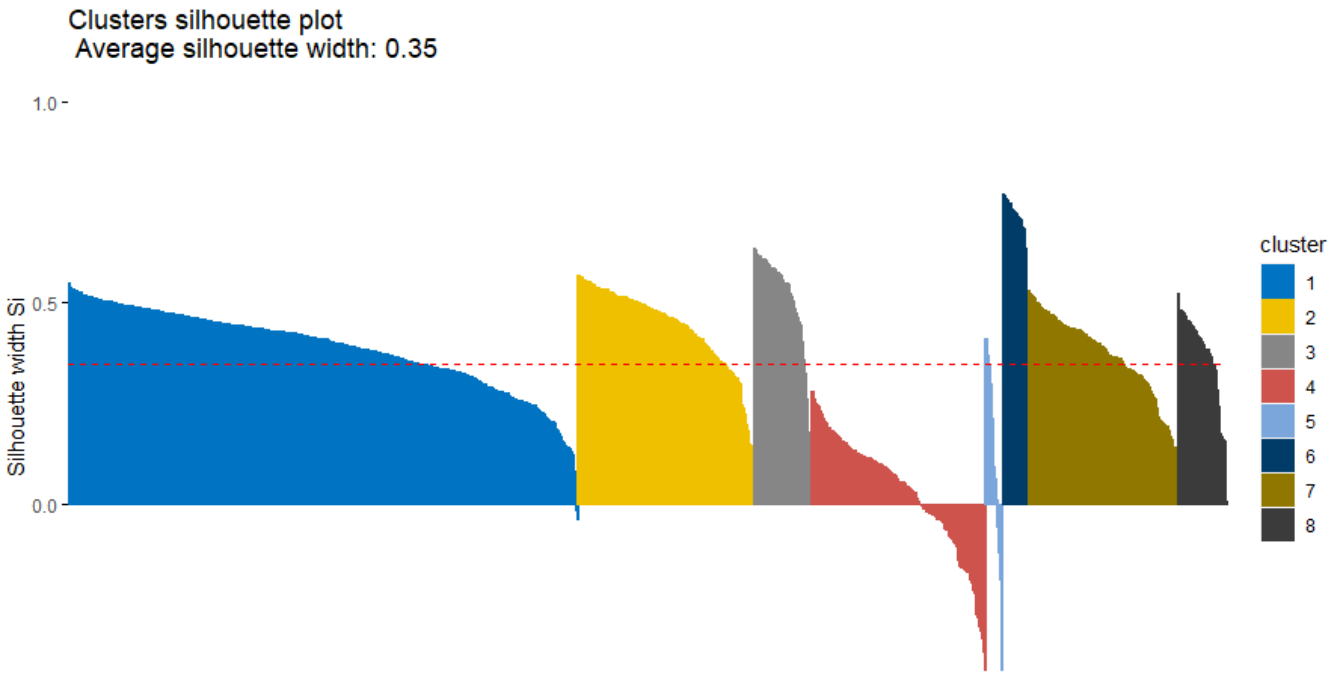


# Clustering: Risultati

| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Numerosità | 617 | 208 | 68 | 129 | 15 | 30 | 238 | 61 |
| % Tipo altro | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 3% |
| % Tipo cemento | 0% | 100% | 0% | 0% | 20% | 0% | 0% | 10% |
| % Tipo cemento e muratura | 0% | 0% | 0% | 0% | 7% | 100% | 0% | 0% |
| % Tipo muratura | 100% | 0% | 0% | 100% | 73% | 0% | 100% | 87% |
| Superificie immobile ($m^2$) | 105 | 90, 70 | 118 | 135 | 560 | 101 | 103 | 140 |
| Altezza dell'acqua (mt) | 0, 52 | 0, 86 | 0, 75 | 1, 40 | 1, 18 | 0, 50 | 1, 87 | 0, 27 |
| Distanza dal fiume (mt) | 1.028 | 1.894 | 1.505 | 3.370 | 2.086 | 1.060 | 3.879 | 5.485 |
| Dislivello (mt) | 10, 40 | 10, 60 | 10, 60 | 11, 30 | 10, 60 | 10, 40 | 10, 90 | 15, 70 |
| Area edifici intorno ($m^2$) | 1.728 | 1.638 | 1.840 | 361 | 1.748 | 1.610 | 1.867 | 741 |
| N. edifici intorno | 7 | 6 | 7 | 2 | 8 | 6 | 9 | 6 |
| Valore immobile (€) | 116.638 | 105.583 | 132.880 | 84.319 | 667.620 | 108.457 | 125.399 | 63.015 |
| Danno relativo (€) | 0, 09 | 0, 06 | 0, 09 | 0, 34 | 0, 02 | 0, 07 | 0, 08 | 0, 19 |
| % Outlier | 7% | 50% | 53% | 26% | 80% | 30% | 9% | 48% |

# Clustering: Silhouette

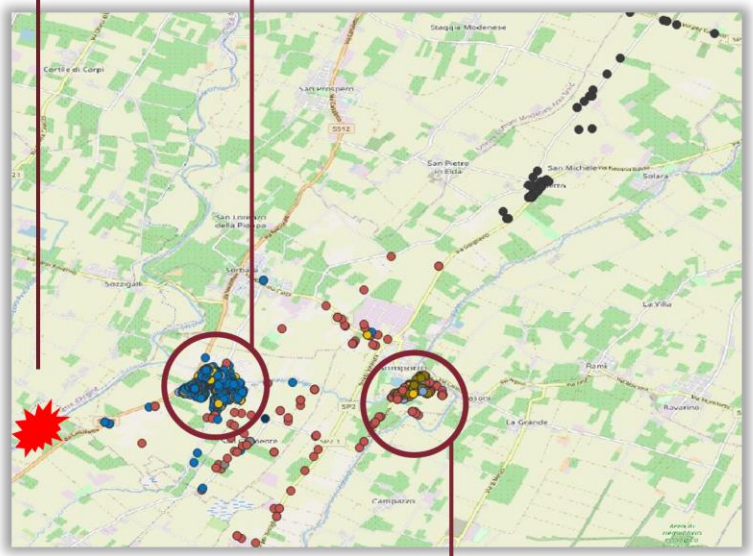


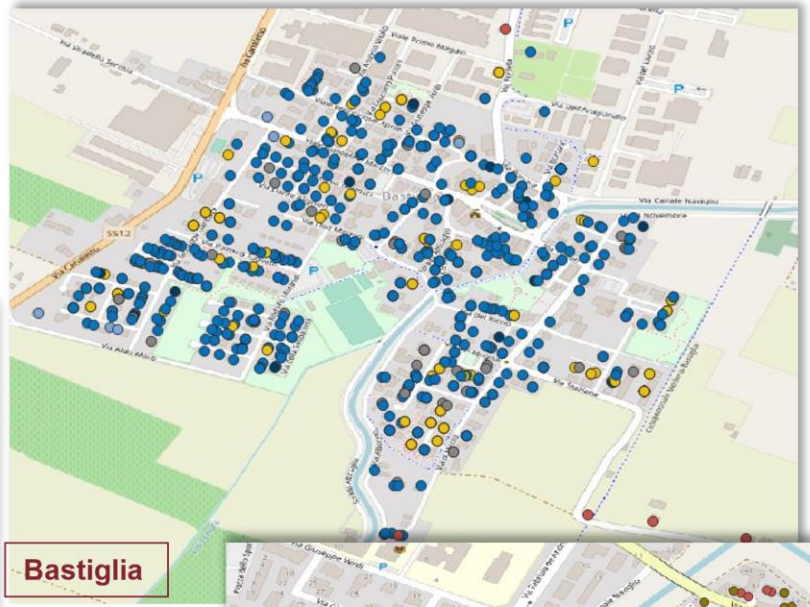

| Cluster | Dimensione | Silhouette media |
|---|---|---|
| 1 | 601 | 0.39 |
| 2 | 207 | 0.44 |
| 3 | 68 | 0.51 |
| 4 | 205 | 0.03 |
| 5 | 21 | 0.10 |
| 6 | 30 | 0.70 |
| 7 | 176 | 0.37 |
| 8 | 58 | 0.37 |

Avendo analizzato i valori della silhouette, sono state ridefinite le appartenenze ai cluster

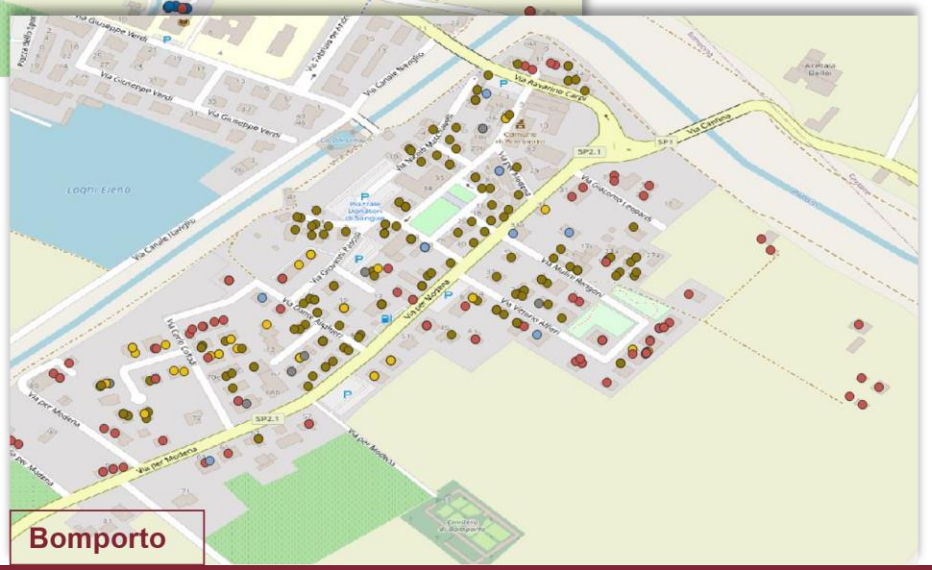
# Clustering: Mappa

Rottura argine
fiume Secchia





Bastiglia



Bomporto

# Clustering: Impatto variabili

| | Varianza nei cluster (WSS) | Varianza tra cluster (GSS) | Pseudo F* |
|---|---|---|---|
| Distanza fiume | 286 | 1079 | 731 |
| Dislivello | 293 | 1072 | 711 |
| Altezza acqua | 548 | 817 | 289 |
| Valore immobile | 726 | 639 | 171 |
| Superficie | 822 | 543 | 128 |
| Area edifici intorno | 964 | 401 | 81 |
| Numero edifici intorno | 1034 | 331 | 62 |
| Danno | 1131 | 234 | 40 |

*$Pseudo\ F = \frac{gss/(k-1)}{wss/(n-k)}$ ,
*k = #cluster,*
*n = #osservazioni*

Variabili legate alla posizione dell'edificio

Variabili legate alle caratteristiche urbane

# Sviluppi Futuri

- Standardizzazione della raccolta e gestione del dato
- Analisi specifica sui valori anomali
- Miglioramento dei modelli ad albero
- Implementazione di ulteriori variabili
- Miglioramento analisi dei grafici PDP
- Approfondimento Pseudo-F

# Grazie dell'attenzione!