UNIVERSITY OF FERRARA
- EX LABORE FRUCTUS -

DE: Department of Engineering Ferrara

Deep Learning

Predizione dei prezzi delle case in vendita a Milano mediante reti neurali sui dati ottenuti con il web scraping di immobiliare.it

Università degli Studi di Ferrara
Andrea Bazerla - 151792

Indice

1. Web Scraping
2. Data Cleaning
3. Data Enrichment
4. Exploratory Data Analysis (EDA)
5. Artificial Neural Network

URL

- Google Colab Notebook
- GitHub
- Datasets
- Immobiliare.it
- TensorBoard

Perché il mercato immobiliare?

- ***Real Estate***: settore molto lento a subire contraccolpi di crisi improvvise come quella del **Covid-19**.
- Congelamento dei prezzi -> Compravendita al ribasso -> Prezzi instabili
- Mercato molto instabile ad ***W*** (*Crisi del mattone*, 1993) anziché a **U** (2008, *Grande recessione*) = imprevedibile.
- **-0.4%** prezzi case in vendita in Italia (Milano **-2.2%**).
  -19,5% compravendite (**-12%** Milano).
- Più attenzione al verde, alla metratura e alla qualità dei servizi
  Dalla città ci si sposta verso l'hinterland.
- **Challenge**: possibile comunque prevedere i prezzi degli immobili in modo accurato?

Web Scraping: definizione

- ***Tecnica per estrarre dati da un sito web in modo automatico via HTTP***
- **Destinatario**: ~~end-user~~, ma software (No strutture dati, protocolli, ecc.)
- **Fasi**: *fetching() -> extracting() -> parsing() -> cleaning() -> storing()*
- **Utilizzi**: monitoring dati meteo, prezzi di prodotti, collezionare informazioni utenti, sentiment analysis, ecc.
- **Scraping** VS **Copy&Paste** VS **API**
- **Difesa**: robot.txt, IP blacklist, CAPTCHA, honeypot
  Contro-attacco: IP routing, Proxy

Web Scraping: librerie utilizzate

- **Pandas**: manipolazione e l'analisi dei dati.
- **Requests**: HTTP requests human-friendly
- **BeautifulSoup**: estrazione e parsing dati da HTML

Web Scraping: immobiliare.it

- **Target**: immobiliare.it, case in vendita, Milano
- **URL**: https://www.immobiliare.it/vendita-case/milano/?criterio=rilevanza
- **Dati annuncio immobiliare**: immagini, planimetria, titolo, prezzo, numero locali, superficie, numero bagni, numero piano, descrizione; area (Milano) > zona, quartiere > [indirizzo]; tipo proprietà, posti auto, caratteristiche, spese condominiali, anno di costruzione, stato immobile, riscaldamento/climatizzazione, efficienza energetica, ecc.

Web Scraping: URL annunci immobiliari

- ***get_last_page()*** -> numero pagine HTML per lo scraping
- ***get_ads_link_list()*** -> lista URL annunci immobiliari da pagina 1 a pagina **get_last_page()**

Web Scraping: dati annunci immobiliari

- ***get_ad_title()*** -> titolo annuncio immobiliare
- ***get_ad_price()*** -> prezzo annuncio immobiliare
- ***get_ad_locations()*** -> area, zona e quartiere annuncio immobiliare
- ...
- ***get_ad()*** -> metodo che in base al tipo di annuncio estrae dati diversi
  - ***get_ad_single()*** -> dati annunci singoli
  - ***get_ad_multi()*** -> dati annunci multipli

Web Scraping: scelte progettuali

- Delay GET requests = 1-2 secondi random tramite *time.sleep().*
- GET requests via Google Colab, IP range = 34.*.*.* - 35.*.*.*.
- Annunci immobiliari di Marzo 2021 = #15’000 = 6h -> Web Scraping eseguito in blocchi

Data cleaning:

- ***Processo di rilevazione e correzione (o rimozione) di record corrotti o imprecisi da un set di dati.***
- **Rimozione** annunci immobiliari multipli, immobili all'asta, immobili privi di un prezzo, immobili privi di un altro attributi fondamentali, colonne superflue, ...
- **Estrazione** dati da stringhe essenziali via regular expression, … https://regex101.com/account/mine#
- **Conversione** prezzi da euro a numeri interi, stringhe in lowercase, liste presenti in attributi in colonne multiple, …

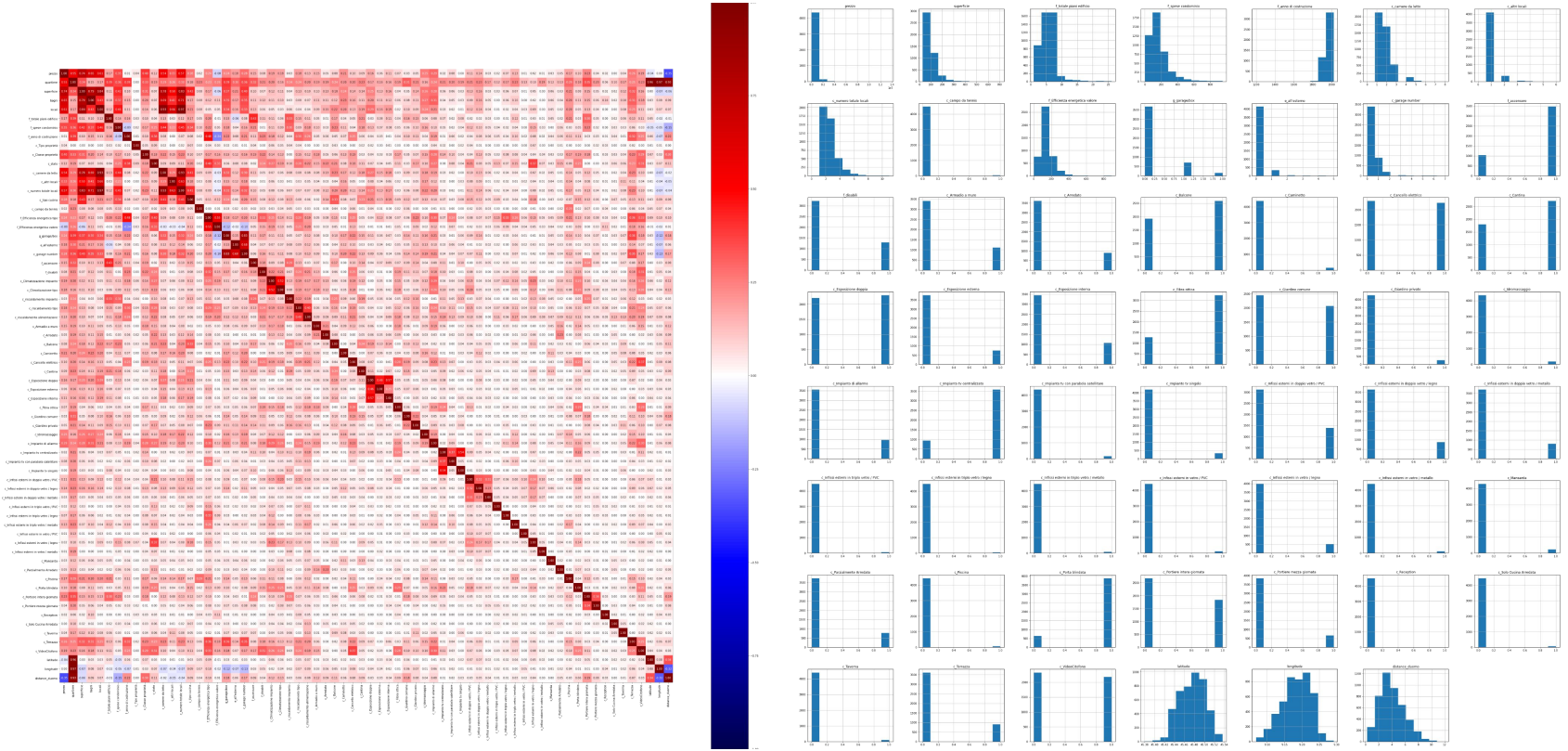
Data enrichment: geographical data

- ***Processo di arricchimento dei dati tramite l'aggiunta di nuove feature da fonti di terze parti.***
- **Geocoder**:
  indirizzo civico -> (Latitude, Longitude) -> Distanza Immobile-Duomo

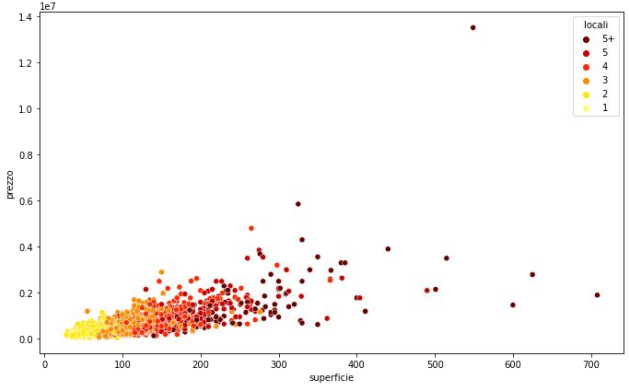
Exploratory Data Analysis (EDA)

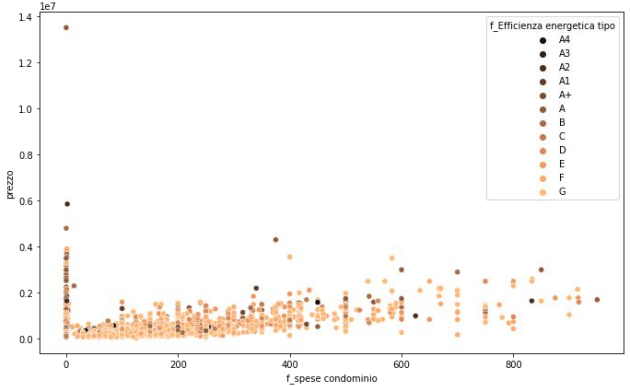
- ***Procedure e tecniche per l'analisi dei dati per riassumere le loro caratteristiche principali, spesso utilizzando grafici statistici e altri metodi di visualizzazione dei dati.***
- Dython: libreria per costruire **matrice di correlazione di Pearson** sia con feature di *valori continui* (prezzi) che con *valori categorici (quartiere)*.

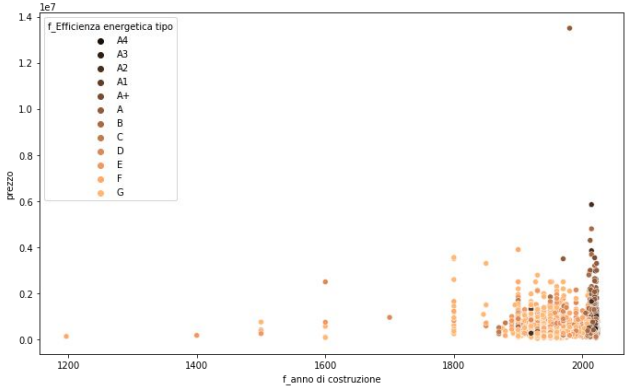
Exploratory Data Analysis (EDA): matrice di Pearson

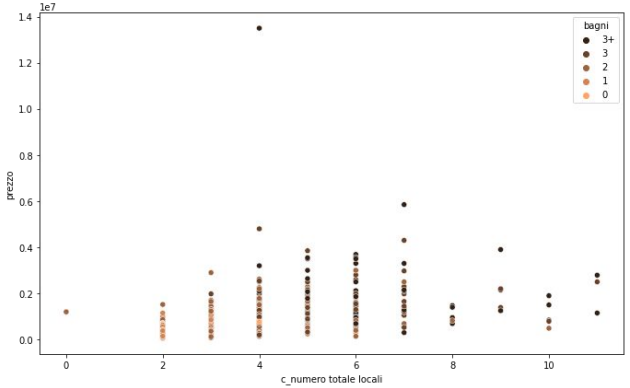


Exploratory Data Analysis (EDA): scatterplots

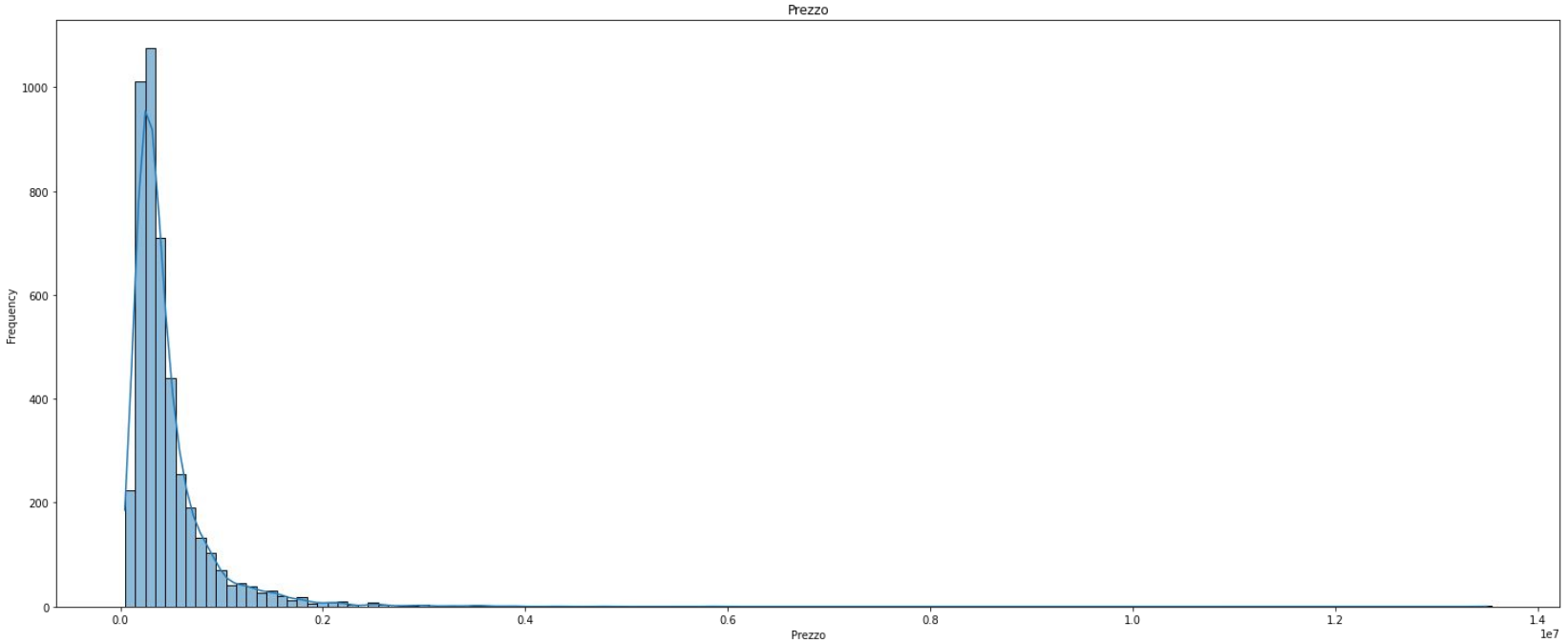






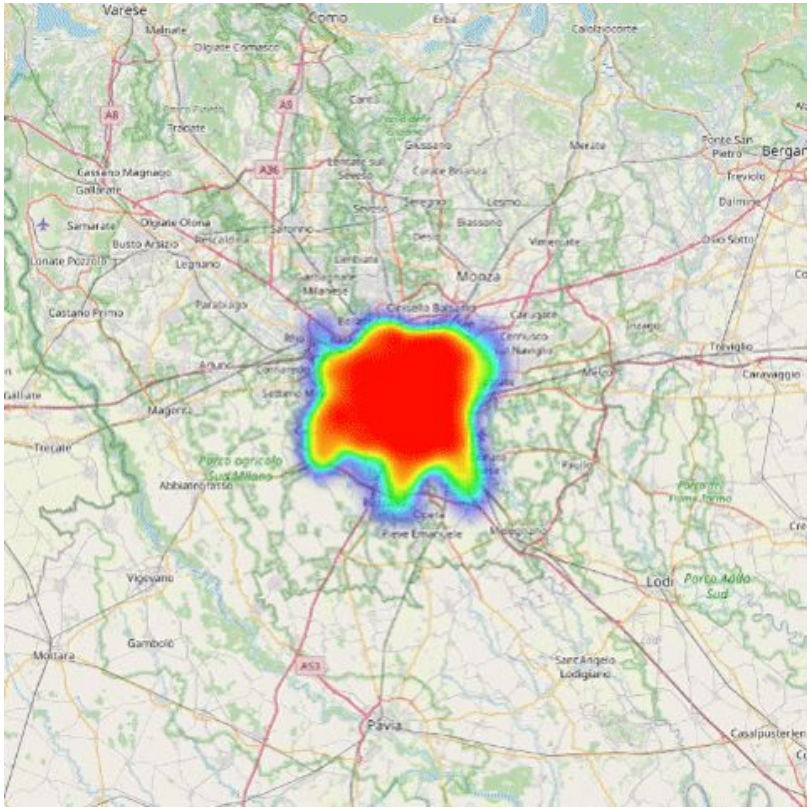


Exploratory Data Analysis (EDA): distribuzioni



Exploratory Data Analysis (EDA): heatmap

- **Heatmap**: densità immobili nello spazio pesata sui prezzi
- **Zone più calde** = alta concentrazione immobili/immobili con prezzi alti



Artificial Neural Network

- **One-Hot Encoding**: quartiere, bagni, locali, ...
- **Standardization**: prezzo, superficie, distanza duomo, ...
- **Batch Normalization**
- **Dropout**: Input Layer = 0-0.2; Hidden Layers = 0-0.5
- **Early Stopping**: patience = 25
- **Keras Tuner**: library for tuning hyperparameters for Keras
- **Cross Validation K-Fold**: K = 10
- **Save/Load modelli NN**: in .h5
- **TensorBoard**

Artificial Neural Network: evaluation

- **M.S.E.** = 0.033
- **Errore medio** = circa €50000
- **% errore medio** = 14%