POLITECNICO DI TORINO

Collegio di Ingegneria Gestionale e della Produzione
Corso di Laurea in Ingegneria Gestionale: Supply Chain Design

Tesi di Laurea di II livello

L'Impatto di Fattori Socioeconomici e Geomorfologici sulla Luminosità Notturna:
Un Approccio di Machine Learning per il Caso Italiano

Relatore:
Prof. Francesco Luigi Milone

Candidato:
Matteo Di Pardo

Anno Accademico 2024-2025

Sommario

Introduzione

Il progresso tecnologico delle immagini notturne satellitari (nighttime Lights, NTL) ha generato un considerevole entusiasmo come potenziale integrazione ai normali dati, contribuendo nello studio di fenomeni socioeconomici e ambientali su diverse scale spaziali e temporali. Questi tipi di dati si sono dimostrati molto utili per la loro capacità di fungere da proxy per spiegare l'attività economica e i fenomeni di urbanizzazione. Le NTL misurano l'intensità della luce catturata passivamente dai satelliti in orbita, fornendo informazioni in modo continuo e a livello globale, a differenza degli altri tipi di dati, che sono spesso caratterizzati da lacune, ritardi e poca affidabilità.

Inizialmente, le luci notturne sono state utilizzate da studi che ne hanno esplorato la correlazione con variabili macroeconomiche (per esempio, il PIL). Tra gli articoli più famosi, Henderson et al. (2017) hanno dimostrato il legame tra le due variabili, aprendo la strada a una ricca letteratura, che impiega l'utilizzo delle luci notturne per dimostrare la correlazione con una vasta gamma di fenomeni.

Le fonti principali a fornire questo tipo di dato sono il Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS), attivo dal 1992 al 2013, e il suo successore, il Visible Infrared Imaging Radiometer Suite (VIIRS), operativo dal 2012 a oggi. Il VIIRS è di gran lunga migliore sotto più aspetti, tra cui la risoluzione e la calibrazione radiometrica. Per questi motivi, è stato deciso di utilizzare il VIIRS come fonte di dati per questa tesi.

Partendo da queste premesse, questa tesi ha lo scopo di esplorare la correlazione tra le NTL e alcune variabili socioeconomiche e morfologiche del suolo italiano, in particolare: la densità di popolazione, il reddito medio, il numero di esercizi ricettivi, la pendenza del suolo, l'altitudine e i laghi. L'obiettivo è quello di confermare la correlazione tra le variabili e di capire la variabile più impattante nella spiegazione della variabilità dell'output del modello. A differenza delle altre ricerche in questo campo, il presente studio adotta un approccio modellistico composto dall'addestramento e la comparazione di tre diversi tipi di modelli: la regressione polinomiale (accompagnata da un'analisi SHAP), la regolarizzazione Lasso (o L1) applicata alla regressione lineare e il Decision Regressor Tree. La prima consente di interpretare i coefficienti stimati e compararli tra loro; la seconda seleziona le feature più importanti e la terza fornisce una rappresentazione visiva della funzione target come una funzione a tratti costante.

Questo studio inizia con la revisione della letteratura dei principali lavori scientifici inerenti al tema della tesi, ovvero a quello delle luci notturne. Nel capitolo successivo sono descritti il modo in cui sono avvenute la gestione e la raccolta dei dataset provenienti da diverse fonti. Il capitolo 3 spiega la metodologia utilizzata, quindi il preprocessing dei dati, il funzionamento dei modelli, la selezione

degli iperparametri e delle metriche di performance. Infine, nel capitolo 4 sono presentati e messi a confronto i risultati dei diversi modelli.

1. Revisione della letteratura

1.1. Luci notturne come proxy di indicatori economici e socioeconomici

Numerosi studi hanno evidenziato come i dati sulle luci notturne (Nighttime Lights o NTL) sono un utile indicatore per monitorare l'attività economica, l'urbanizzazione e i cambiamenti ambientali. Questi dati sono registrati da sensori satellitari, che forniscono informazioni sulle emissioni luminose della superficie terrestre. Esistono due fonti principali: il Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) e il Visible Infrared Imaging Radiometer Suite (VIIRS). Il DMSP-OLS copre il primo periodo di disponibilità di questi dati, ovvero dal 1992 al 2013, mentre il VIIRS ha agito come sostitutivo dal 2013 fino ad oggi. I dati "raw" corrispondono a dei pixel di 0,5 chilometri quadrati e riportano un'intensità luminosa misurata tra 0 e 256.

La letteratura scientifica mostra che i dati VIIRS sono migliori rispetto a quelli DMSP-OLS per più motivi. Ad esempio, la forza predittiva dei dati VIIRS è l'80% più alta di quella dei dati DMSP (Gibson, 2021) e hanno una risoluzione spaziale 45 volte maggiore (Elvidge et al., 2013), senza sfocature (Abrahams et al., 2018) o errori di geolocalizzazione. Un altro problema che accomuna i dati DMSP è il fenomeno del top-coding: il DMSP-OLS codifica la luminosità come numeri interi compresi tra 0 e 63, causando la saturazione del sensore. Per questo motivo, i centri urbani fortemente illuminati ottengono sempre il valore massimo, rendendoli indistinguibili dalle aree meno luminose e causando una sottostima delle aree molto illuminate e una sovrastima di quelle più buie.

Tra i principali articoli che si servono dei dati delle luci notturne c'è Henderson et al. (2012), inerente alla misurazione della crescita economica di un paese attraverso le luci notturne, analizzandone la correlazione sia nel breve che nel lungo termine. Il PIL è la variabile più importante per l'analisi della crescita economica, ma spesso è misurato in modo sbagliato a causa delle complicanze legate al suo calcolo. Tra le principali, c'è la bassa integrazione regionale e una debole infrastruttura statistica. Il calcolo del PIL nominale richiede anche l'implementazione di indici di prezzi affidabili, assenti in molti paesi in via di sviluppo. Nella PWT, database macroeconomico internazionale, i paesi ricevono delle valutazioni in base alla qualità dei loro dati. Tutti i paesi dell'Africa subsahariana hanno un voto di C o D, che corrisponde, secondo Chen e Nordhaus (2011), a margini di errore del 20-30%. Avvalendosi dei dati DMSP, gli autori formulano regressioni con effetti fissi e trend temporali specifici per paese. I risultati mostrano un valore dell'elasticità stimata della crescita del PIL rispetto alla crescita delle luci di circa 0,277, ovvero l'aumento dell'1% delle luci è associato a una crescita del PIL di 0,277%, con un R-quadro di 0,769. Per i paesi più poveri, la stima ottimale della crescita è una composizione tra la crescita misurata convenzionalmente e quella prevista dalle luci, pesate

ugualmente. Le stime differiscono dai dati ufficiali fino a tre punti percentuali all'anno. Questo risultato dimostra come le luci notturne possono essere utilizzate come sostituto (o complemento) dei conti nazionali, nel caso di deboli infrastrutture statistiche. Inoltre, usando le luci notturne, è possibile misurare la crescita per regioni sub- e sovranazionali, non richiedendo più l'uso dei paesi come unità di analisi. Ad esempio, nell'Africa subsahariana è possibile documentare che le aree costiere stanno crescendo più lentamente dell'entroterra, qualcosa che con i conti "tradizionali" non è possibile fare.

In un altro noto articolo di Henderson et al. (2017) sono utilizzate le luci notturne VIIRS per studiare la correlazione tra la luminosità delle luci notturne terrestri osservate dai satelliti (proxy dell'attività economica di un paese) e 24 caratteristiche geografiche del territorio. Gli autori si focalizzano sulle caratteristiche geografiche dette "first-nature", ovvero quelle che non cambiano nel tempo, per esempio l'abitabilità e la produttività del suolo. Nonostante queste caratteristiche siano statiche temporalmente, l'impatto che hanno sull'economia muta a causa del cambiamento tecnologico e delle trasformazioni strutturali. In questo articolo, gli autori si focalizzano sulle due aree di maggiore importanza dove queste caratteristiche sono cambiate di più, ovvero l'idoneità di un paese alla produzione agricola e al commercio internazionale. Negli esperimenti trattati, le covariate si possono dividere in 3 gruppi: quelle che influenzano sia l'agricoltura che il commercio, quelle solo agricole (come la temperatura, le precipitazioni e l'elevazione) e quelle solo commerciali (come l'accesso ai trasporti acquatici). Gli autori scoprono, in primo luogo, che c'è una forte correlazione tra la densità di popolazione e l'illuminazione (R-quadro uguale a 0,53); secondariamente, trovano che le 24 variabili geografiche spiegano il 47% della variazione dell'illuminazione globale. Gli effetti fissi spiegano, da soli, il 35% della variazione, dando un contributo marginale di 11 punti percentuali in più rispetto alle variabili geografiche. Viceversa, i fattori geografici aggiungono 23 punti percentuali oltre gli effetti fissi.

A differenza degli studi precedenti, Mellander (2015) propone un'analisi più precisa della relazione tra le luci notturne e l'attività economica, ma limita l'analisi allo stato svedese. In passato, molti ricercatori hanno studiato la relazione tra le NTL e l'attività economica sfruttando variabili macroeconomiche (es. GDP); l'autore, invece, si pone a un livello di dettaglio maggiore. In particolare, analizza quanto è vicina la relazione tra le due variabili a livello microscopico. Infatti, un satellite può rilevare le luci provenienti dal suolo, ma non può, ad esempio, catturare la luce emessa dagli uffici di un palazzo, oppure distinguere tra un palazzo con degli uffici popolati da sviluppatori di software o da operai tessili. Per comparare le luci notturne con i dati demografici della "Statistics Sweden", è stata generata una griglia composta da celle di 250 x 250 m nelle aree urbane e 1000 x 1000 nelle aree rurali. Per l'analisi, le variabili geocodificate considerate sono relative alla

popolazione, ai redditi da lavoro, al numero di stabilimenti, ai dipendenti e ai salari totali. I risultati, trovati attraverso analisi di correlazione e regressioni geograficamente ponderate (dette GWR), dimostrano una forte correlazione delle luci notturne con la densità di popolazione e degli stabilimenti, mentre una più debole con i salari. Quindi, le luci stimano meglio le variabili demograiche rispetto ai redditi, e le variabili di densità sono più forti delle variabili di conteggio, dimostrando che le NTL sono utili per studi di urbanizzazione e sulla distribuzione spaziale, piuttosto che per stime economiche.

1.2. Approcci empirici utilizzando i dati delle luci notturne per stimare shock esogeni

Nella storia, molti eventi hanno avuto un effetto negativo sull'economia di un paese, come pandemie e guerre. Numerosi studi hanno utilizzato i dati delle luci notturne per studiarne l'impatto.

Ad esempio, Roberts (2021) esamina il potenziale delle luci notturne ad alta frequenza per monitorare quasi in tempo reale l'impatto economico causato dalla pandemia di COVID-19 in Marocco. Il suo contributo scientifico si può riassumere in due punti: la validazione dell'utilizzo delle luci notturne ad alte frequenze temporali e la sua applicazione ad eventi esogeni, come la pandemia COVID-19. Gli autori sfruttano le potenzialità dei dati del satellite VIIRS, disponibili con una frequenza mensile, per monitorare l'impatto della crisi economica in tempistiche vicine a quelle "real-time". Inoltre, la copertura nuvolosa molto bassa del Marocco permette di visualizzare le luci notturne più facilmente, al che si somma la già elevata risoluzione dei dati VIIRS, risultando, nel complesso, in un'ottima fonte di dati. Nell'articolo sono confrontate le applicazioni di più "maschere" ai raster, come l'EOG, utile per rimuovere il rumore di sottofondo dato da luci effimere (come barche e incendi). La variabile dipendente è la somma delle luci (Sum of Lights, SOL), corrispondente alla somma dell'intensità luminosa di una data area (nazionale, subnazionale e urbana), mentre quella indipendente è il PIL. I risultati trovati mostrano una correlazione significativa tra le due variabili: con l'applicazione della maschera EOG, l'elasticità tra il PIL e SOL è di circa 0,3 con un R-quadro uguale a 0,37; ciò corrisponde a una caduta del PIL di 3,3 punti percentuali. A marzo 2020, quando scoppiarono i primi casi di COVID-19, si registrò un forte calo (circa del 10%) delle luci notturne rispetto al periodo prima. È stato anche possibile fare delle stime a livello subnazionale: tra le città che sono state maggiormente colpite ci sono Casablanca (-9,7%) e Safi (-7,72%).

Oltre all'evento esogeno della pandemia, in altri articoli è analizzato l'impatto economico causato dalla guerra, come quello sul conflitto russo-ucraino di L. Buzzacchi et al. (2024). In questo articolo, è esaminato e quantificato l'impatto economico del conflitto vicino al "real-time" a livello

subnazionale (Raion) sfruttando i dati delle luci notturne. Come per Roberts (2021), anche qui è stato necessario essere in possesso di un dato ad alta frequenza e ad alta risoluzione per poter calcolare delle stime nel breve periodo; perciò, sono stati utilizzati i dati del satellite VIIRS dal 2018 al 2023. I dati sul conflitto sono stati ricavati dal dataset VIINA, che traccia in tempo reale qualsiasi tipo di evento militare "violento". Il numero di eventi di difesa aerea, le condizioni meteo e il numero di eventi militari per Raion-mese sono alcune delle variabili di controllo incluse nei modelli, mentre le due variabili dipendenti considerate sono la media e la mediana della luminosità notturna. I risultati mostrano un calo statisticamente significativo della luminosità notturna come conseguenza della guerra (circa l'8% della media e il 3% della mediana da gennaio 2022) ed è dimostrata l'eterogeneità sull'impatto locale del conflitto nell'attività economica dei Raion, dipendentemente dal fatto che l'area sia stata attaccata o meno e con quale intensità. Nel breve termine, l'effetto sulle luci impiega da 3 a 5 mesi a diventare statisticamente significativo, mentre nel medio termine gli effetti negativi sono persistenti. Infine, gli autori calcolano l'elasticità tra luci notturne e GVA (uguale a circa 0,072%), con lo scopo di stimare l'impatto economico della guerra, risultante in un valore compreso tra 130 e 473 milioni di dollari USA.

1.3. Approcci empirici utilizzando i dati delle luci notturne sul suolo italiano

L'Italia rappresenta un caso di studio ideale sulla correlazione tra le luci notturne e il fenomeno dell'urbanizzazione, data la forte eterogeneità delle aree che la compongono (industriali, agricole e turistiche).

Pambuku et al. (2023) hanno utilizzato le luci notturne del satellite VIIRS per analizzare alcune dinamiche di urbanizzazione del territorio pugliese. Gli obiettivi posti sono 3: (1) studiare la variabilità negli ultimi 10 anni, (2) valutare come questa variazione può essere spiegata da variabili spaziali come la distanza tra centri urbani e la prossimità alle coste, (3) monitorare la differenza delle NTL tra le diverse stagioni e come questa può spiegare il cambiamento demografico dovuto all'afflusso turistico. Il test di Kruskal-Wallis e post-hoc Dunn sono i metodi utilizzati per l'analisi temporale delle luci notturne, mentre per l'analisi spaziale è stata modellata una regressione iperbolica, che include variabili come la distanza dalle aree urbane, dalle città e dalla costa. Inoltre, è stato creato un modello esponenziale per stimare la relazione tra la variazione delle luci notturne nelle diverse stagioni e l'incremento di popolazione dovuto al turismo. I risultati mostrano un aumento generale dell'intensità delle NTL in Puglia dal 2014 al 2023, soprattutto lungo la costa (come Bari e Taranto). La stagione estiva si è dimostrata essere quella più luminosa (+15,28% contro +12,38% dell'inverno dal 2014), supportando la correlazione con il turismo (R-quadro di 0,65). Inoltre, come

previsto, il fattore spaziale più influente sulle NTL è la distanza dalle aree urbane, che spiega circa il 40% da solo.

Uno dei fenomeni negativi causati dall'urbanizzazione è l'inquinamento luminoso notturno (ANTL), un tema spesso trascurato ma di grande rilevanza. Marcantonio et al. (2015) analizzano la profondità di questa trasformazione, che ha conseguenze fisiologiche sugli esseri umani e implicazioni ecologiche ed evolutive sulla flora e sulla fauna. Per l'analisi del paesaggio notturno, gli autori introducono un nuovo indice, detto VANI (Vegetation Adjusted NTL Index), che combina i dati ANTL delle luci notturne VIIRS insieme all'Enhanced Vegetation Index (EVI). Ciò ha permesso di studiare gli effetti e l'estensione delle ANTL nel paesaggio notturno di due aree protette in Italia: i Colli Euganei e il Parco Nazionale del Cilento, Vallo di Diano e Alburni. I risultati mostrano che solo il 30% delle aree analizzate rimane "buio" e le aree più idonee si trovano nelle zone più interne dei parchi. Infine, gli autori simulano una decrescita esponenziale di ANTL (4-21% per Colli Euganei, 5-50% per Cilento), dimostrando un possibile recupero fino a circa il 50% di quote di terreno idonee alla biodiversità.

2. Raccolta e gestione dei dati

La prima fase per il raggiungimento dell'obiettivo di questa tesi è stata quella della raccolta e della gestione dei dati. Sono stati raccolti dati economici, socioeconomici, demografici e geografici, tutti su scala nazionale italiana. Le fonti sono state selezionate in base al tipo, alla qualità e alla disponibilità temporale del dato.

In seguito alla raccolta dei dati, è stato realizzato un unico dataset che contenesse, associate ad ogni unità di pixel, le seguenti features: luminosità media, altitudine, pendenza, presenza di laghi, densità di popolazione, reddito medio, numero di esercizi ricettivi, numero di posti letto.

2.1. Dati utilizzati

- Shapefile

Uno shapefile è un formato di archiviazione di dati vettoriali usato per archiviare la posizione, la forma e gli attributi delle feature geografiche. I dati vettoriali utilizzano geometrie (punti, linee o poligoni) per rappresentare caratteristiche discrete, come strade o confini. Gli shapefile dei confini amministrativi di ogni comune italiano aggiornati al 1° gennaio 2025 sono stati ricavati dal sito ufficiale dell'Istat.

I confini delle unità amministrative a fini statistici sono costituiti da tre livelli gerarchici (regioni, province e comuni) e uno statistico (ripartizioni geografiche) a copertura nazionale, a cui si aggiungono negli anni censuari anche le aree speciali (zone in contestazione e isole amministrative). Il formato dei dati è *shapefile* nel sistema di riferimento WGS84; il dettaglio tecnico della proiezione è riportato nel file in formato prj, associato a ciascun file geografico. La scala non è certificabile uniformemente dall'Istat, poiché le basi di acquisizione utilizzate (principalmente foto aeree ed altra cartografia) provengono da fonti e scale differenti che variano tra ambito urbano ed extraurbano. Gli attributi degli shapefile sono codificati (encoding) in UTF-8 come descritto nel relativo file in formato cpg, collegato al dato geografico. Il dataset contiene, per ogni riga, il codice ISTAT di ogni comune e il relativo poligono.

- Luci notturne

Il dataset utilizzato in questa analisi è il Visible Infrared Imaging Radiometer Suite (VIIRS). VIIRS contiene immagini catturate dal satellite Suomi National Polar-orbiting Partnership, messo in orbita congiuntamente dalla NASA e dalla NOAA. Come accennato nel capitolo precedente, il dataset è il successore del Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) per le immagini notturne della Terra. Quest'ultimo copre il primo periodo di disponibilità di

questi dati, dal 1992 al 2013, mentre il VIIRS copre dal 2013 fino ad oggi. Il suo sensore a bassa luminosità cattura immagini con una risposta spettrale di 505-890 nm e una lunghezza d'onda nominale al centro della banda di 705 nm. La risoluzione spaziale delle immagini è di 500 m.

I dati sono stati scaricati da due archivi dell'Earth Observation Group della Colorado School of Mines, accessibili al pubblico. Il primo ricopre il periodo dal 2012 al 2021, il secondo dal 2022 al 2024. All'interno degli archivi, è possibile scaricare i dati VIIRS di frequenza annuale a mensile in formato GeoTIF (raster). I dati sono disponibili in più tipologie: "average" (il valore luminoso medio), "median" (la mediana), "minimum" (il minimo valore), "maximum" (il massimo valore), "raw" e "masked". A quest'ultimo viene applicata una "maschera" al file raster che permette di eliminare i rumori causati da segnali luminosi effimeri come incendi, navi e vulcani. Per questa tesi sono stati considerati solo i dati "average" di tipo "raw" dal 2013 al 2024.

Un raster è un tipo di file costituito da una matrice di celle (o pixel) organizzate in righe e colonne (o griglia) in cui ogni cella contiene un valore che rappresenta un'informazione, come ad esempio la temperatura. L'area (o superficie) rappresentata da ciascuna cella ha la stessa larghezza e altezza ed è una porzione uguale dell'intera superficie rappresentata dal raster. Ad esempio, un raster che rappresenta l'elevazione (ovvero un modello digitale di elevazione) può coprire un'area di 100 chilometri quadrati. Se in questo raster fossero presenti 100 celle, ciascuna di esse rappresenterebbe 1 chilometro quadrato di larghezza e altezza uguali (ovvero 1 km x 1 km). Alcuni raster hanno una singola banda o livello (una misura di una singola caratteristica) di dati, mentre altri hanno più bande. Fondamentalmente, una banda è rappresentata da una singola matrice di valori di cella, e un raster con più bande contiene più matrici di valori di cella spazialmente coincidenti che rappresentano la stessa area spaziale. Un esempio di set di dati raster a banda singola è un modello digitale di elevazione (DEM). Inoltre, GeoTIFF incorpora tag di metadati per i sistemi di riferimento (CRS).

- Densità di popolazione

I dati relativi alla densità di popolazione sono stati ottenuti da uno dei dataset del Global Human Settlement Layer (GHSL), un progetto nato con lo scopo di fornire informazioni spaziali, analisi e conoscenze che descrivono la presenza umana sul pianeta Terra. Il GHSL opera secondo una politica di accessibilità ai dati gratuita e supporta il GEO Human Planet Initiative (HPI), che si impegna a sviluppare una nuova generazione di misurazioni e prodotti informativi che forniscano nuove prove scientifiche e una comprensione completa della presenza umana sul pianeta e che possano supportare i processi politici globali con metriche concordate, attuabili e orientate agli obiettivi. Tra i principali

partner di HPI ci sono la Commissione Europea, la Direzione Generale Centro Comune di Ricerca e il progetto GHSL.

Il raster considerato raffigura la distribuzione e la densità della popolazione del 2015, espressa come numero di persone per cella a una risoluzione di 250 m x 250 m. Le stime della popolazione residenziale, fornite da CIESIN Gridded Population of the World, versione 4.10 (GPWv4.10) a livello di poligono, sono state disaggregate dalle unità censuarie o amministrative alle celle della griglia, basate sulla distribuzione e sulla densità degli insediamenti urbani mappati nel livello globale del Global Human Settlement Layer (GHSL) per epoca corrispondente.

- Reddito medio

I dati sul reddito medio a livello comunale sono stati scaricati dal sito del Dipartimento delle Finanze, disponibili solo dal 2000 al 2023, imponendo la riduzione dell'analisi temporale di un anno.

Il Dipartimento delle Finanze diffonde e promuove i dati statistici sulle dichiarazioni fiscali in formato aperto, in accordo con il Codice dell'Amministrazione Digitale e le linee guida indicate dall'Agenzia per l'Italia digitale. Il dataset fornisce informazioni economiche sui comuni italiani, tra cui il numero di contribuenti, il reddito imponibile, il reddito complessivo e tante altre. Per il calcolo del reddito medio, è stata considerata la media come rapporto tra l'ammontare dell'imponibile del Comune e il numero di contribuenti con reddito uguale o superiore a zero.

- Alberghi

I dati relativi agli alberghi dal 2013 al 2024 sono stati ottenuti dal sito ufficiale dell'Istat. La rilevazione quantifica, a livello di singolo comune, il numero degli esercizi, dei letti, delle camere e dei bagni per le strutture alberghiere ed extra-alberghiere.

L'indagine viene svolta con periodicità annuale in conformità al Regolamento (Ue) n. 692/2011 del Parlamento europeo e del Consiglio relativo alle statistiche europee sul turismo così come modificato dal Regolamento delegato (Ue) n. 2019/1681 della Commissione del 1° agosto 2019, pubblicato nella Gazzetta ufficiale dell'Unione europea del 9 ottobre 2019.

Le unità di analisi sono le strutture ricettive presenti sul territorio nazionale, riferite a ciascun comune italiano, classificate secondo la normativa nazionale e le normative regionali e distinte in:

- Strutture alberghiere: alberghi classificati in cinque categorie, distinte per numero di stelle e residenze turistico-alberghiere;

- Strutture extra-alberghiere: campeggi e aree attrezzate per camper e roulotte, villaggi turistici, forme miste di campeggi e villaggi turistici, alloggi in affitto gestiti in forma imprenditoriale, agriturismi, ostelli per la gioventù, case per ferie, rifugi di montagna, altri esercizi ricettivi non altrove classificati, bed and breakfast e altri alloggi privati.

Ai fini della raccolta dei dati, l'Istat - ai sensi del d.lgs. n. 322/1989 - si avvale degli Uffici di statistica delle Regioni e delle Province autonome, in qualità di organi intermedi. Ogni anno, l'Istat invia agli Uffici di statistica delle Regioni e delle Province autonome, o agli eventuali altri uffici o enti di cui gli Uffici di statistica si avvalgono per la raccolta dei dati a livello regionale o provinciale, una circolare molto dettagliata in cui vengono fornite tutte le indicazioni per la conduzione dell'indagine.

- Altitudine, pendenza e laghi

Le caratteristiche ricavate per lo studio in questione sono state ricavate da 3 file DEM diversi.

Il DEM è un set di misure che registrano l'elevazione della superficie della terra e che contengono anche l'informazione delle relazioni spaziali tra queste misure. Il metodo globalmente più diffuso e consolidato per registrare l'informazione altimetrica è la disposizione delle quote all'interno di una griglia regolare (raster o grid), rappresentabile con una matrice numerica. Solitamente la griglia ha una maglia quadrata, la cui dimensione del lato fornisce la dimensione della cella (cell size o pixel size), che corrisponde, fissata la proiezione, alla risoluzione spaziale del DEM.

I file DEM possono essere suddivisi in 3 categorie:

- DSM (Digital surface model): descrive la superficie terrestre inclusi gli oggetti posti su di essa (vegetazione, edifici, ecc);
- DEM (Digital elevation model): descrive l'altimetria della superficie terrestre (normalmente si riferisce al geoide);
- DTM (Digital terrain model): usato a volte come sinonimo di DEM, più correttamente da usare quando ci si riferisce anche ad informazioni sugli attributi del terreno e non solo alla quota.

In generale i DEM raster presentano i seguenti vantaggi:

- La griglia regolare è una struttura semplice che può facilmente essere ricostruita;
- È più facile derivare parametri relativi alla superficie perché si possono usare algoritmi più semplici;

- Hanno una struttura spaziale uniforme che può generalmente essere definita da un solo parametro, la dimensione della cella;
- Il modello a griglia è maggiormente adatto ai modelli informatici usati nell'analisi delle immagini (image processing).

Le features considerate sono:

- Altitudine: valore in metri;
- Pendenza: valore in gradi compreso tra 0 e 90;
- Laghi: variabile binaria per cui 1 corrisponde a un pixel su un lago e 0 viceversa.

Nonostante provengano da file diversi, questi sono stati ottenuti a loro volta da uno stesso documento DEM chiamato "TINITALY". Il raster fornisce dati sull'elevazione del suolo nudo e può quindi essere definito un DTM (Digital Terrain Model, modello digitale del terreno). Il file è stato ottenuto dalla fusione dei file DEM delle singole regioni amministrative. Il DEM è disponibile come griglia con celle di 10 m (in formato GeoTIFF), nel sistema di riferimento (CRS) UTM WGS 84 zona 32.

Un Sistema di Riferimento delle Coordinate (CRS) è un framework utilizzato nei sistemi informativi geografici (GIS) per definire la posizione dei dati spaziali sulla superficie terrestre. Implicito in qualsiasi set di dati GIS, un CRS stabilisce un riferimento spaziale che può variare da griglie locali arbitrarie (ad esempio, un'area di campionamento di 10 m x 10 m) a sistemi globali legati alla curvatura terrestre. Fondamentalmente, un CRS comprende un sistema di coordinate geografiche (GCS) o un sistema di coordinate proiettate (PCS), che consente la mappatura e l'analisi accurate di elementi quali immagini satellitari o file di forma vettoriale. Comprendere il CRS è fondamentale per attività come il ritaglio di shapefile su dati raster, poiché le discrepanze possono portare a distorsioni spaziali e disallineamenti. Un GCS definisce le posizioni sulla superficie curva della Terra utilizzando misurazioni angolari in gradi di latitudine e longitudine, con riferimento all'equatore e al meridiano fondamentale (tipicamente Greenwich, Inghilterra). La latitudine misura l'angolo a nord o a sud dall'equatore (positivo per il nord, negativo per il sud), mentre la longitudine misura l'angolo a est o a ovest dal meridiano fondamentale (positivo per l'est, negativo per l'ovest).

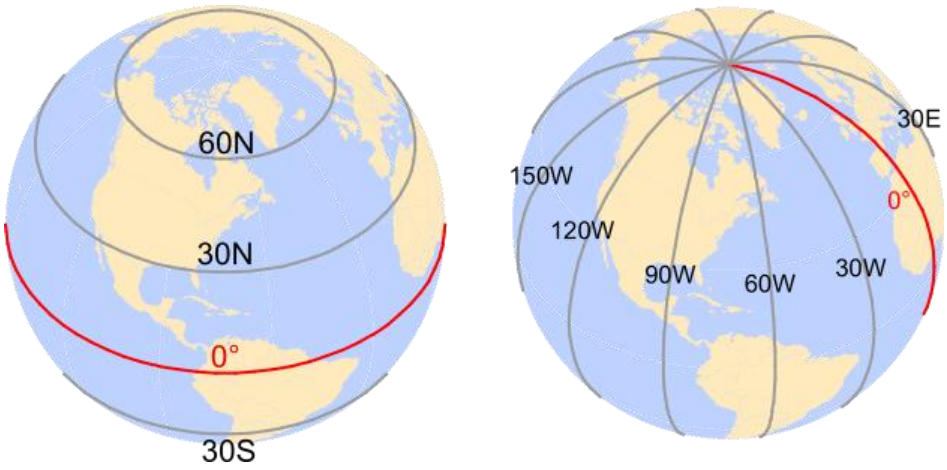

Figura 1: Latitudine e longitudine sulla Terra

Un GCS è costruito su un ellissoide (un'approssimazione della forma della Terra, con semiassi maggiore e minore di circa 6378137 m e 6356752 m, rispettivamente) e un datum, che allinea l'ellissoide al geoide, la vera superficie della Terra. I datum possono essere locali (ad esempio, NAD27 per gli Stati Uniti continentali) o geocentrici (ad esempio, WGS84, ampiamente utilizzato a livello globale).

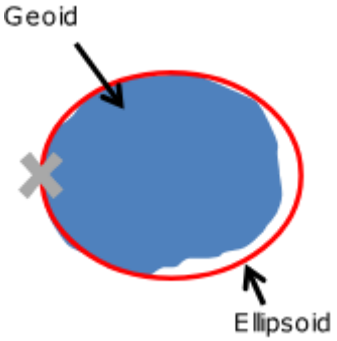

Figura 2: Esempio di rappresentazione di datum locale

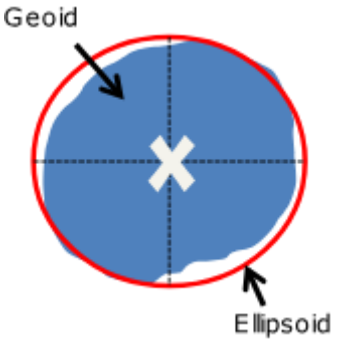

Figura 3: Esempio di rappresentazione di datum geocentrico

Un sistema di coordinate proiettate (PCS) è un sistema di riferimento utilizzato per identificare posizioni e misurare caratteristiche su una superficie piana (mappa). È costituito da linee che si intersecano ad angolo retto, formando una griglia.

L'importanza del CRS è evidente in operazioni come il ritaglio di file shapefile su raster, dove i confini vettoriali (ad esempio, i poligoni comunali italiani) vengono utilizzati per ritagliare le griglie raster (ad esempio, le immagini notturne VIIRS). Ad esempio, senza un CRS coerente, le coordinate di un punto in NAD27 potrebbero apparire sfalsate di centinaia di metri in WGS84, portando a ritagli disallineati. Affinché questo non accada, i CRS dei file raster e degli shapefile devono coincidere. Mentre la conversione del CRS degli shapefile al CRS di un raster è un processo semplice, la trasformazione del CRS di un raster a quello di un altro raster è più complesso. La trasformazione dei CRS dei raster a un unico CRS di riferimento è un passaggio fondamentale, poiché se venisse applicato il ritaglio dei diversi raster per ogni shapefile in modo indipendente, si otterrebbero dei dataset con coordinate sfalsate, e quindi non sarebbe possibile applicare il merge su questi. Il processo è strutturato in più passaggi: per prima cosa, è necessario avere un raster di riferimento (per esempio, quello delle luci notturne), successivamente viene creato un raster virtuale (VRT) per le caratteristiche secondarie (ad esempio, l'altitudine) e, allineandolo alla griglia di riferimento, viene applicato un ritaglio congiunto.

La struttura finale del dataset delle variabili indipendenti è questa:

Tabella 1: Struttura del dataset delle variabili indipendenti

x	y	Codice comune	Codice provincia	Codice regione	Nome comune	Altitudine	Pendenza	Lag hi	Anno	Densità di popolazione	Reddito medio	Numero posti letto
7.7750	45.3875	1001	1	1	Agliè	433.0	8.860	0	2013	0.0	21125.4	60
7.7791	45.3875	1001	1	1	Agliè	444.0	11.160	0	2013	0.0	21125.4	60
7.7666	45.3833	1001	1	1	Agliè	388.8	7.440	0	2013	1.545816	21125.4	60
7.7708	45.3833	1001	1	1	Agliè	410.5	10.805	0	2013	0.67	21125.4	60

Il formato dei dati è:

- Luminosità: Float32. Permette di mantenere una precisione di 6/7 cifre decimali;

- Altitudine: Float16. Il valore massimo in Italia è circa 4800 m, ben al di sotto del massimo rappresentabile da Float16 (circa 65500). Una precisione dell'ordine del millimetro è più che adeguata;
- Pendenza: Float32. Numero decimale compreso tra 0 e 90;
- Coordinate: Float32, poiché i valori arrivano fino a 6 numeri decimali;
- Densità di popolazione: Float32, poiché i valori arrivano fino a 6 numeri decimali;
- Numero posti letto, numero esercizi ricettivi, laghi, anno: Intero. Sono conteggi o valori binari.

2.2. Analisi delle distribuzioni

Prima di procedere con la creazione dei modelli, è stata svolta un'analisi dei dati sulle distribuzioni e sulle caratteristiche principali. Dopo aver controllato che non fossero presenti valori nulli nel dataset e che i valori fossero tutti maggiori o uguali a 0, sono state visualizzate le statistiche principali di tutte le variabili attraverso la funzione "describe()" di Pandas relativa all'anno 2023, nonché a quello più recente. Di seguito è possibile visualizzare tutte le distribuzioni (fino al 98° percentile) e una tabella di riepilogo delle principali statistiche descrittive delle variabili considerate.

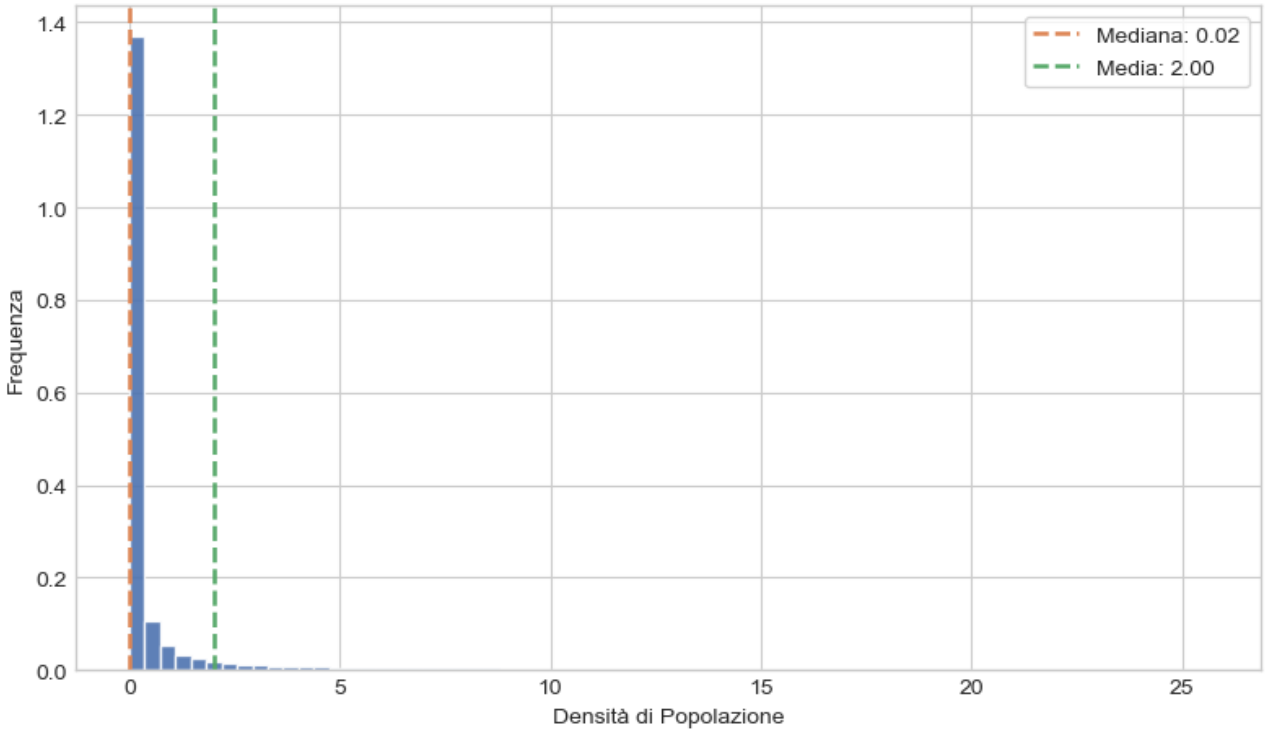

Figura 4: Istogramma della variabile Densità di Popolazione

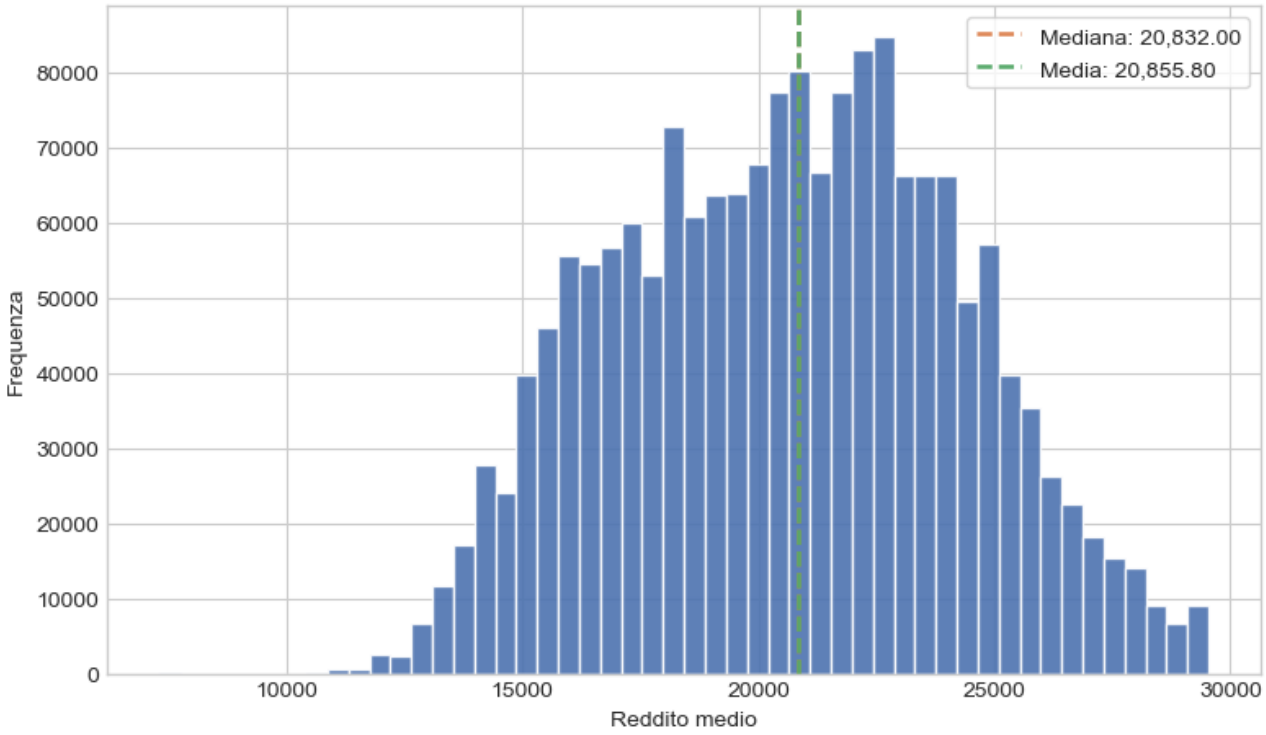

Figura 5: Istogramma della variabile Reddito medio

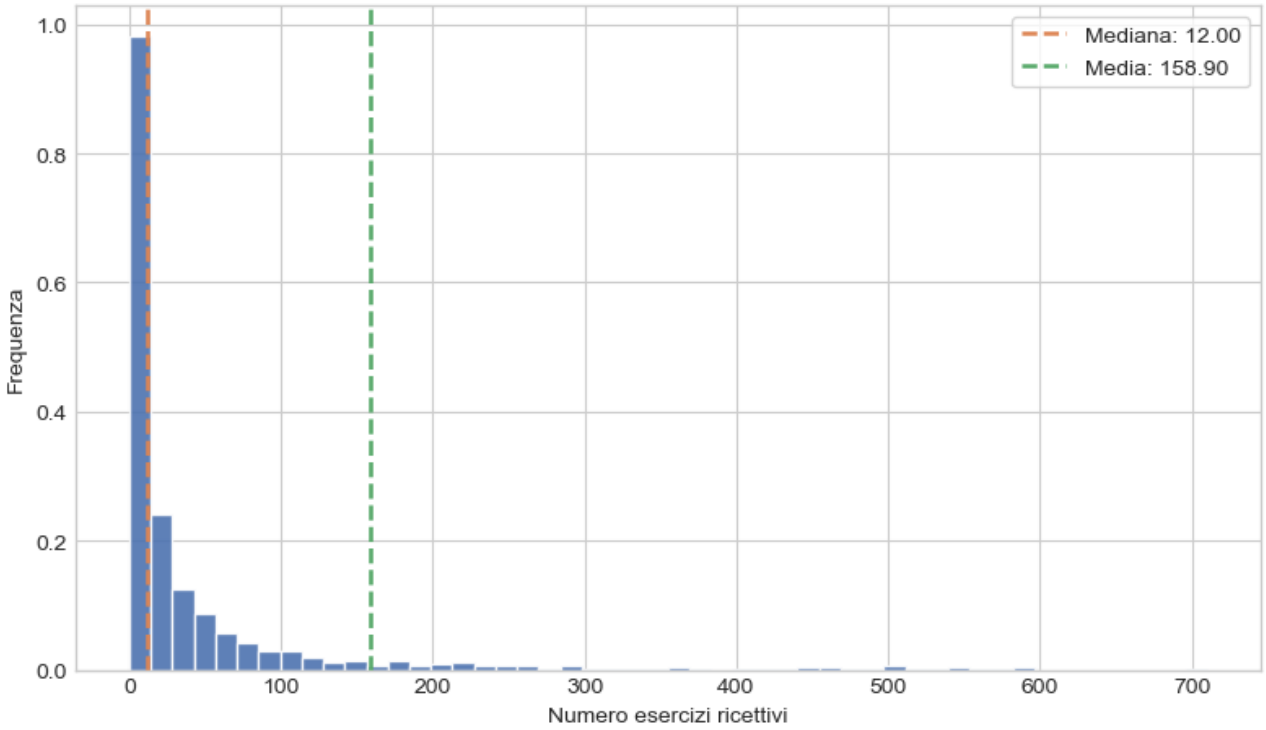

Figura 6: Istogramma della variabile Numero di esercizi ricettivi

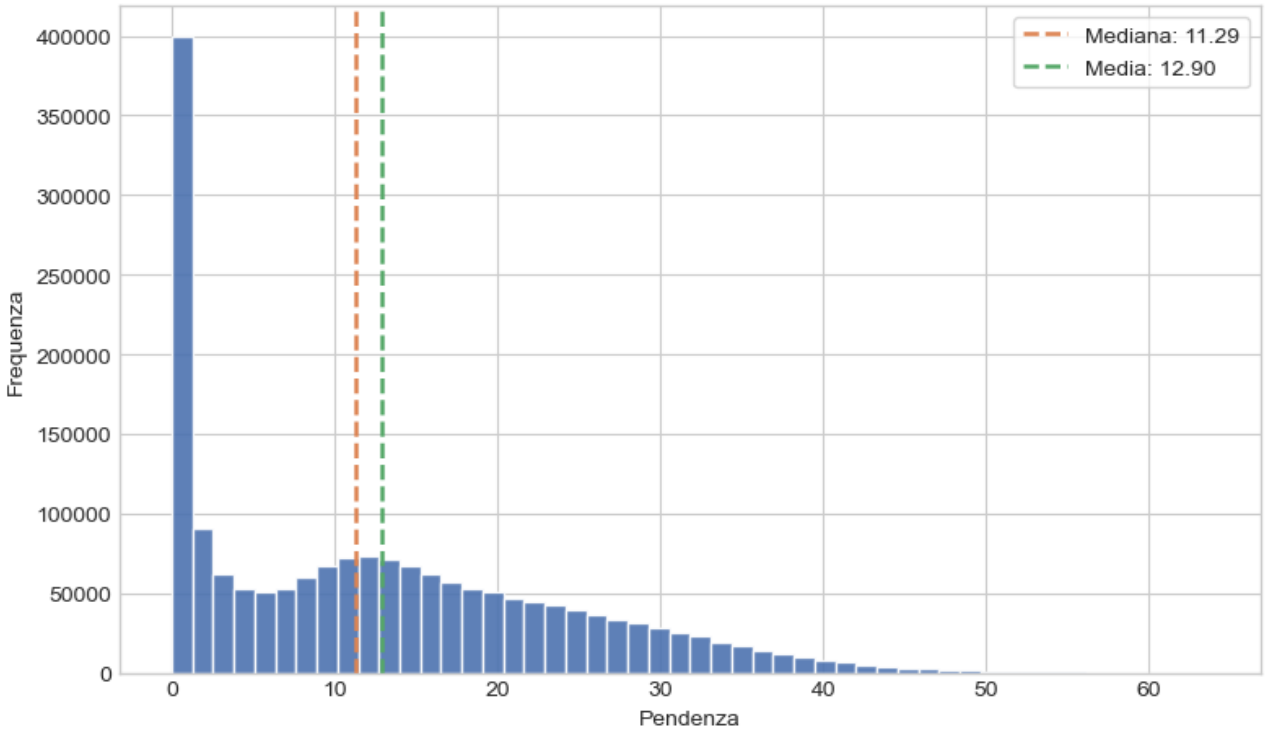

Figura 7: Istogramma della variabile Pendenza

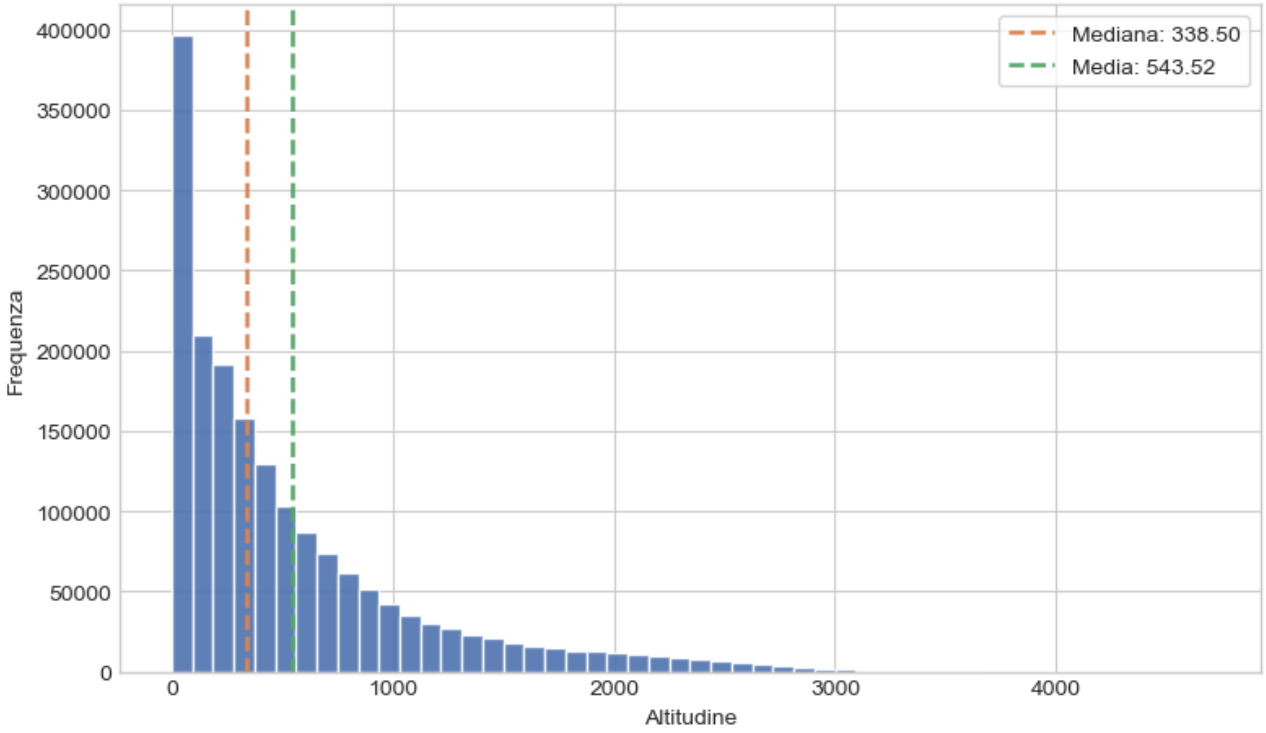

Figura 8: Istogramma della variabile Altitudine

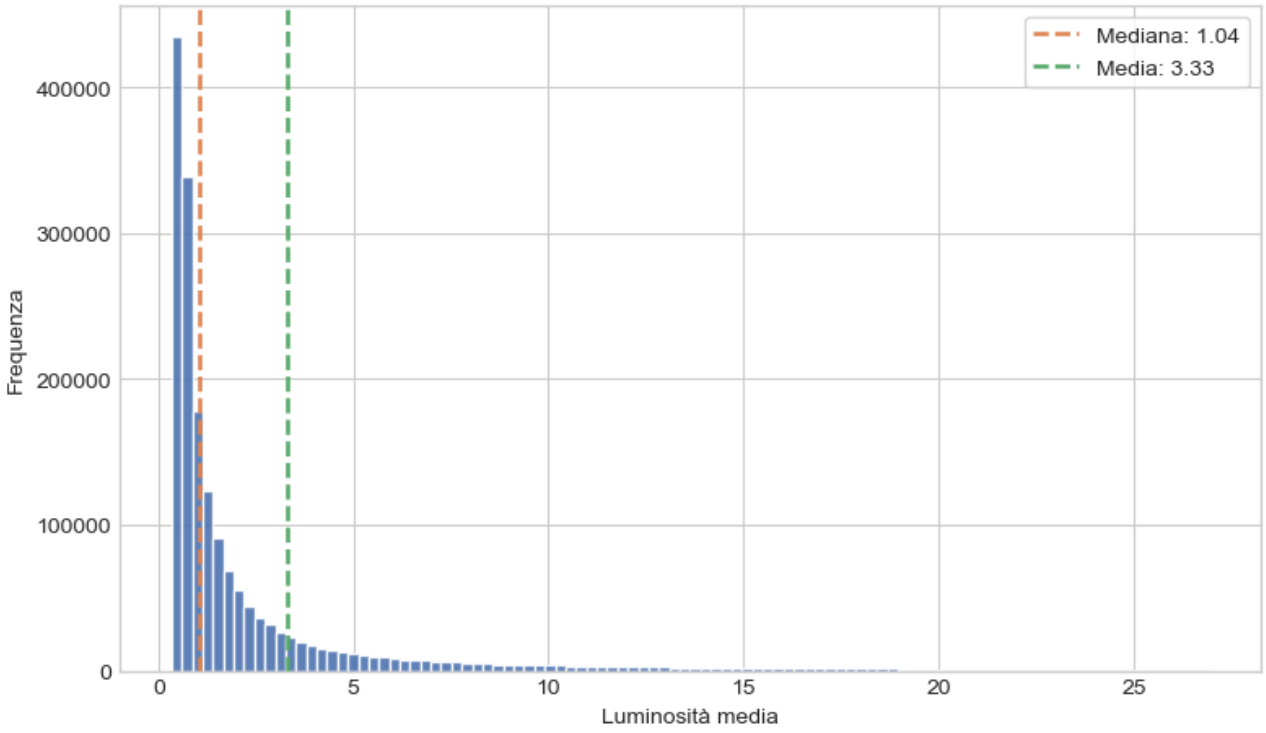

Figura 9: Istogramma della variabile Luminosità media

Tabella 2: Descrizioni statistiche delle features del modello

Statistica	Altitudine	Pendenza	Laghi	Densità di popolazione	Reddito medio	Numero posti letto	Numero esercizi ricettivi	Luminosità media
count	1793259	1793259	1793259	1793259	1793259	1793259	1793259	1793259
mean	543,522	12,9024	0,00704	1,998097	20855,8	2724,16	158,9047	3.325161
std	597,62	11,277	0,08362	10,18073	3996,95	16309,2	1183,09	7,34223
min	0	0	0	0	7318.8	0	0	0.3327
25%	113,7	1,87	0	0	17914.6	47	4	0,60571
50%	338,5	11,29	0	0,017019	20832	203	12	1,037493
75%	746	20,53	0	0,318336	23477,6	1002	42	2,540511
max	4690	63,66	1	463,7168	94505,4	226585	16034	329,9934

- Altitudine: presenta una forma fortemente asimmetrica a destra, infatti la mediana (338 m) è molto inferiore alla media (543,5 m). Ciò dimostra che la maggior parte delle osservazioni sono a bassa quota e una piccola parte alza la media (fino a 4690 m);
- Pendenza: leggera asimmetria a destra, dovuta a una buona parte delle osservazioni vicino allo 0;
- Laghi: variabile altamente sbilanciata, dimostrata dall'assenza (valore 0) di superfici idriche nella gran parte del suolo italiano;
- Densità di popolazione: estremamente asimmetrica a destra, con la mediana molto vicina allo 0 e media poco più alta;
- Reddito medio: è la distribuzione più simmetrica rispetto alle altre, con media e mediana molto vicine tra loro e una variabilità contenuta;
- Numero di posti letto: altamente sbilanciata. Nonostante i molti luoghi turistici in Italia, pochi di questi possiedono la maggior parte della capacità ricettiva;
- Numero esercizi ricettivi: fortemente asimmetrica a destra, per lo stesso motivo del numero di posti letto;
- Luminosità media: asimmetrica a destra, a causa di poche aree fortemente illuminate.

3. Metodologia

3.1. Classificazione delle variabili

Dopo aver analizzato la distribuzione delle variabili del dataset, è importante capire nel dettaglio quali variabili includere per la costruzione del modello e come eseguire il preprocessing.

Le variabili selezionate e le rispettive categorie sono:

- Coordinate: x e y;
- Variabili categoriche: Codice comune, Codice provincia, Codice regione;
- Variabili binarie: Laghi;
- Variabili continue: Densità di popolazione, Reddito medio, Numero esercizi ricettivi, Numero posti letto, Altitudine;
- Variabili continue con range noto: Pendenza (valore compreso tra 0 e 90);
- Variabile target: Luminosità media.

Le luci notturne artificiali catturate dai sensori satellitari provengono principalmente da città e attività umane; quindi, ci si aspetta che la densità di popolazione, il reddito medio e le variabili del turismo siano positivamente correlate con la luminosità notturna. Al contrario, si prevede che la pendenza e l'altitudine siano negativamente correlate con la variabile dipendente, poiché, in genere, alti valori di queste variabili sono associati a un basso livello di urbanizzazione. Allo stesso modo per i laghi, che non sono edificabili.

Non sorprende il fatto che una parte della varianza potrebbe essere spiegata da effetti fissi spaziali, come un set di dummy a livello di comune. Tuttavia, i codici ID sono stati esclusi dal modello: se si dovesse applicare il One Hot Encoder, la misura del dataset aumenterebbe di circa 7900 colonne per riga a causa del gran numero di chiavi uniche comunali, superando la RAM a disposizione e ostacolando il completamento dell'addestramento del modello. Lo stesso accade a livello provinciale e regionale, anche se in misura minore.

L'encoding one-hot è una tecnica utilizzata in Machine Learning per trasformare le caratteristiche categoriali in un formato che può essere facilmente compreso dagli algoritmi. In sostanza, questo metodo trasforma ogni categoria di una variabile in una nuova colonna binaria, dove ogni colonna corrisponde a una categoria e contiene un 1 o un 0 che indica la presenza o l'assenza di quella categoria nei dati.

Oltre alle variabili categoriche, anche le coordinate e il numero di posti letto sono stati esclusi dal modello. Quest'ultima variabile è fortemente correlata con il numero di esercizi ricettivi; per questo

sono state confrontate con i risultati ottenuti ed è stata esclusa la variabile che spiega meno la variabilità delle luci notturne.

3.2. Pre-processing e scelta dei modelli

I modelli selezionati per questo studio sono la regressione polinomiale, la regolarizzazione L1 (o Lasso) e il Decision Regressor Tree.

- Regressione polinomiale

La regressione lineare è una tecnica di modellazione statistica utilizzata per descrivere una variabile di risposta continua in funzione di una o più variabili (predittori). Il modello descrive la relazione tra una variabile dipendente Y (in questo caso la luminosità notturna media) in funzione di una o più variabili indipendenti X (in questo caso le suddette caratteristiche). L'equazione generale per un modello di regressione lineare è la seguente:

Equazione 1: Equazione generale per un modello di regressione lineare

$$Y = \beta_0 + \sum \beta_k X_k + \varepsilon_i$$

dove β rappresenta le stime per i coefficienti lineari da calcolare e ε rappresenta i termini di errore. L'obiettivo è quello di determinare la retta che meglio descrive i punti osservati, ovvero di determinare i coefficienti che minimizzano la somma dei quadrati degli scarti tra i valori stimati e quelli osservati, detti anche residui della regressione. Questo criterio è detto Metodo dei Minimi Quadrati.

La **regressione polinomiale**, invece, utilizza lo stesso metodo della regressione lineare, ma assume che la funzione che meglio descrive l'andamento dei dati sia un polinomio, non più una retta. Quindi è adatta quando lo scatterplot di una relazione bivariata, ad esempio, mostra una forma diversa da quella della retta, ad esempio, una curva. In questa tesi, è stata creata una versione "lineare" e una "quadratica" di regressione, aggiungendo i termini quadratici di tutte le variabili incluse nel modello, ad eccezione della variabile binaria Laghi.

I principali vantaggi legati all'utilizzo della regressione polinomiale per questa analisi sono:

- Interpretabilità: i coefficienti e i p-value sono facilmente leggibili, per capire quali delle caratteristiche risultano più impattanti sulla spiegazione della variabilità delle luci notturne;
- Basso costo computazionale;
- Cattura della non "linearità" dei rapporti tra le variabili indipendenti e quella dipendente.

Affinché la relazione tra le X e la Y sia più lineare, si riduca l'eteroschedasticità e si attenui l'influenza di eventuali outlier, è fondamentale applicare una funzione che riduca l'asimmetria delle variabili, necessaria come fase di pre-processing prima dell'addestramento del modello. Secondo l'analisi nel capitolo precedente, alcune variabili presentano una distribuzione fortemente asimmetrica; per questo è stata applicata la **funzione logaritmo** per ridurne l'asimmetria e renderla più simile a una distribuzione normale. Questa trasformazione riguarda le variabili della Densità di popolazione, Reddito medio, Altitudine, Numero di esercizi ricettivi.

La funzione logaritmica di una variabile quantitativa e non negativa X si ottiene mediante:

Equation 2: Funzione logaritmica

$$X^* = log(X)$$

Dove $log(.)$ è il logaritmo in qualsiasi base. Qualunque sia la base, l'effetto della trasformazione logaritmica è lo stesso, a meno di un coefficiente di proporzionalità. L'effetto della funzione logaritmica riduce la distanza tra la modalità minima e quella massima, il che trasforma la distribuzione in una forma più simmetrica avvicinando i valori estremi a quelli centrali.

Escluso il decision regressor tree, dopo l'applicazione della funzione logaritmo alle features, è stata anche applicata la standardizzazione. Le features standardizzate sono: Densità di popolazione, Reddito medio, Numero esercizi ricettivi, Numero posti letto, Altitudine. La variabile Pendenza è stata normalizzata attraverso il Min-Max Scaling, dividendo ogni valore per il suo massimo, ovvero 90.

La **standardizzazione** (in inglese, Z-score normalization o standardization) è un procedimento statistico di manipolazione dei dati che, nel caso di un dataset, modifica i valori di una o più features affinché abbiano le proprietà di una distribuzione Gaussiana con $\mu=0$ e $\sigma=1$ (media uguale a 0 e deviazione standard uguale a 1). La deviazione standard, riferita ad una variabile casuale, ne indica la dispersione attorno ad un indice di posizione quale, ad esempio, la media aritmetica. Questa applicazione permette di confrontare due valori che hanno magari unità di misura diverse, affinché gli ordini di grandezza non rischino di interferire sulla bontà del modello, andando ad alterare i rapporti tra le variabili indipendenti e la variabile dipendente.

La **normalizzazione Min-Max scaling** è una tecnica di normalizzazione che consente di trasformare i valori delle feature in un intervallo definito, in genere tra 0 e 1. Il valore trasformato della feature si ottiene sottraendo il suo valore minimo e dividendolo per la differenza tra il suo valore massimo e il suo valore minimo. In questo modo, i valori della feature vengono compressi in un intervallo

uniforme, rendendo più facile la comparazione tra le feature e garantendo che abbiano un impatto equo sui risultati del modello.

- Regolarizzazione Lasso (o L1)

Lasso (o Least Absolute Shrinkage and Selection Operator) è una regolarizzazione spesso utilizzata nel campo del Machine Learning per gestire dati ad alta dimensione in quanto seleziona automaticamente le features di maggiore importanza con la sua applicazione. Ciò avviene aggiungendo un termine di penalità alla somma residua dei quadrati (RSS), che viene poi moltiplicato per il parametro di regolarizzazione (lambda o λ), che misura la quantità di regolarizzazione applicata. Più viene aumentato lambda, più i coefficienti sono spinti verso lo zero, il che a sua volta riduce l'importanza di alcune features del modello (o, in alcuni casi, le elimina del tutto), con la loro conseguente selezione automatica. Al contrario, più diminuisce il valore di lambda, più si riduce l'effetto della penalità, mantenendo più features all'interno del modello.

Questa regolarizzazione permette di evitare problemi di multicollinearità e di overfitting all'interno dei dataset, promuovendo la scarsità all'interno del modello. La multicollinearità è un fenomeno che avviene quando più variabili sono correlate tra loro, creando difficoltà per la modellazione. Aumentando il valore di lambda, alcuni coefficienti vengono portati a 0, in modo da eliminare le variabili indipendenti del modello poco utili, e quindi identificando le variabili di maggiore impatto sulla spiegazione della variabilità del target.

Esistono due tipi di penalizzazione:

- L1 (**absolute size**) penalizza il valore assoluto dei coefficienti del modello
- L2 (**squared size**) penalizza il quadrato del valore dei coefficienti del modello.

La regressione Lasso usa la penalità L1. Il rischio qui è che un valore molto alto porterà il modello all'underfitting, cioè non catturerà i pattern presenti nei nostri dati.

In questa tesi, la regolarizzazione Lasso porta diversi vantaggi:

- Seleziona le feature di maggiore importanza in presenza di collinearità: molte feature sono correlate tra di loro (come la densità di popolazione e il numero di esercizi ricettivi), per questo la regolarizzazione porta a 0 i coefficienti ridondanti, lasciando solo quelli principali per la spiegazione dell'output del modello;
- Facile lettura con il Lasso plot: è possibile capire l'importanza relativa in base al momento dell'"entrata" della variabile dovuta al decremento dell'intensità della penalizzazione (alpha).

- Decision Regressor Tree

Gli **alberi decisionali** sono un metodo di apprendimento supervisionato non parametrico utilizzato per la classificazione e la regressione. L'obiettivo è creare un modello che preveda il valore di una variabile target apprendendo semplici regole decisionali dedotte dalle caratteristiche dei dati. Un albero può essere visto come un'approssimazione costante a tratti.

Alcuni vantaggi degli alberi decisionali sono:

- Semplicità di comprensione e interpretazione: gli alberi possono essere visualizzati;
- Il costo dell'utilizzo dell'albero (ovvero la previsione dei dati) è logaritmico rispetto al numero di punti dati utilizzati per addestrare l'albero;
- È in grado di gestire sia dati numerici che categorici;
- In grado di gestire problemi multi-output;
- Utilizza un modello white box, ovvero un modello di cui è osservabile la logica interna ed è spiegabile come una certa previsione venga prodotta. Al contrario, in un modello black box (ad esempio, in una rete neurale artificiale), i risultati possono essere più difficili da interpretare;
- È possibile validare un modello utilizzando test statistici. Ciò consente di valutare l'affidabilità del modello;
- Funziona bene anche se le sue ipotesi sono in qualche modo violate dal modello reale da cui sono stati generati i dati.

Gli svantaggi degli alberi decisionali includono:

- Gli algoritmi di apprendimento degli alberi decisionali possono creare alberi eccessivamente complessi che non generalizzano bene i dati, causando il fenomeno dell'overfitting. Meccanismi come il pruning, l'impostazione del numero minimo di campioni richiesti in un nodo foglia o l'impostazione della profondità massima dell'albero sono necessari per evitare questo problema;
- Gli alberi decisionali possono essere instabili perché piccole variazioni nei dati potrebbero generare un albero completamente diverso;
- Le previsioni degli alberi decisionali non sono né regolari né continue, ma approssimazioni costanti a tratti;

- Gli algoritmi di addestramento degli alberi decisionali si basano su algoritmi euristici, in cui vengono prese decisioni localmente ottimali in ciascun nodo. Tali algoritmi non possono garantire di restituire l'albero decisionale globalmente ottimale.

Per la luminosità notturna, questo tipo di modello è molto utile per interpretare le variazioni di Y oltre determinate soglie di valore delle feature. Ad esempio, quanta luminosità ci si aspetta di avere oltre un certo valore di altitudine o di pendenza, oppure eventuali salti di Y oltre specifici valori della densità di popolazione.

In questa tesi, dato che la variabile target è continua, è stato utilizzato il decision regressor tree.

Nel preprocessing del decision regressor tree, né la standardizzazione né la funzione logaritmo sono state applicate alle distribuzioni delle variabili, poiché raramente cambia la scelta degli split.

I suddetti modelli sono stati selezionati allo scopo di valutare quali features hanno più influenza sulla spiegazione della variabilità della luminosità notturna media. La regressione polinomiale permette di valutare i coefficienti e la significatività statistica di ogni feature. L'effetto, nonostante possa essere molto significativo, potrebbe avere un basso effetto sulla spiegazione della variabile target. Ciò è visualizzabile grazie alla regolarizzazione Lasso, che permette di analizzare, aumentando e diminuendo la penalità, quali sono le feature più importanti. Infine, i primi livelli del decision regressor tree sono i rami più rappresentativi, che coinvolgono le variabili maggiormente influenti sul modello.

3.3. Selezione degli iperparametri

Per poter eseguire l'addestramento del modello, è stato eseguito uno split del dataset in base alla disponibilità temporale:

- Dataset di train: 2013-2021
- Dataset di test: 2022-2023

In base al tipo di modello, potrebbe essere necessario identificare e selezionare gli iperparametri ottimali da utilizzare nell'addestramento. Gli iperparametri sono variabili da impostare in anticipo per gestire il processo di formazione di un modello di machine learning. Ogni algoritmo di apprendimento automatico favorisce il proprio set di iperparametri e non è necessario massimizzarli in tutti i casi. Ad esempio, il decision regressor tree è un modello che richiede la selezione e l'identificazione di molti iperparametri, per questo è necessario ottimizzare la loro ricerca attraverso gli strumenti adatti. Se questa fase viene eseguita correttamente, si riduce al minimo la funzione di perdita, e quindi le prestazioni del modello vengono addestrate per essere il più accurate possibili.

L'**Hyperparameter Optimization** è l'ambito che si occupa della scelta degli iperparametri ottimali per un modello di Machine Learning. In ogni modello di Machine Learning, l'analista può decidere i valori dei parametri da inserire, in modo da migliorare le sue performance. In genere, se questi parametri non sono modificati, sono utilizzati un preset di valori messi a disposizione. Non fare "ottimizzazione degli iperparemetri" significa quindi utilizzare questo preset, senza preoccuparsi di cercare i valori ottimali.

Ad esempio, il Decision Tree ha il problema di non riuscire bene a generalizzare con dati nuovi, ovvero di cadere nel fenomeno dell'overfitting. Gli iperparemetri hanno l'obiettivo di prevenirlo, come per esempio max_depth, min_samples_leaf e max_features.

Ci sono due tecniche principali per automatizzare la ricerca degli iperparametri: la Grid Search e la Random Search.

La **Grid Search** permette di inserire un insieme di valori per i parametri che vogliamo ottimizzare e prova tutte le possibili combinazioni (un modello per ogni combinazione). Ad esempio, nel caso del DecisionTree, posso dire al modello di addestrarsi con criterion = "gini", criterion = "entropy" e max_depth = [3, 4, 5], e la Grid Search farà un modello per ognuna delle possibili combinazioni di questi iperparametri. La tecnica ha il vantaggio di trovare la migliore combinazione tra le combinazioni proposte; tuttavia, il suo calcolo può risultare molto lento soprattutto se il numero di iperparametri è molto alto e se il dataset è grande. Inoltre, può portare all'overfitting, poiché trova la soluzione migliore per il dataset di training ma non generalizza ai nuovi dati.

Il **Random Search**, invece, seleziona un certo numero di randomiche combinazioni di iperparametri estratte da specifiche distribuzioni, piuttosto che dei valori ben definiti. Ad esempio, nel caso di variabili continue, posso dire al modello di addestrarsi con max_depth = uniform(3,5) invece di specificare i valori. Nel caso di iperparametri categorici, i valori possono essere esplicitati e il Random Search li estrae in modo randomico. Il Random Search produce risultati più rapidi, pur trovando iperparametri ottimali o quasi ottimali, soprattutto quando lo spazio di ricerca è ampio. Campionando casualmente gli iperparametri da una distribuzione di valori, esplora lo spazio in modo più ampio, scoprendo combinazioni più performanti senza testare ogni singola possibilità.

In GridSearch e in RandomSearch, viene eseguita anche la convalida incrociata K-fold, utilizzata durante l'addestramento del modello. Come abbiamo detto, prima di addestrare il modello, i dati sono divisi temporalmente in due parti: dati di training (2013-2021) e dati di test (2022-2023). Nella cross validation, il processo divide ulteriormente i dati di training in due parti: dati di training e dati di validation. La K-fold cross validation è una procedura che consente di dividere il dataset di training

in k partizioni. A ogni iterazione, sono utilizzate k-1 partizioni per l'addestramento del modello e la k-esima per il testing. L'iterazione successiva imposterà la partizione successiva come dati di test e le restanti partizioni come dati di training. A ogni iterazione, sono registrate le performance ottenute dal modello e, infine, è applicato il refit sull'intero dataset in base agli iperparametri del modello da cui si ottengono le prestazioni migliori.

Date le dimensioni elevate del dataset dell'analisi in questione e il vasto numero di iperparametri del Decision Regressor Tree, si è preferita una strategia di ottimizzazione di tipo Random Search.

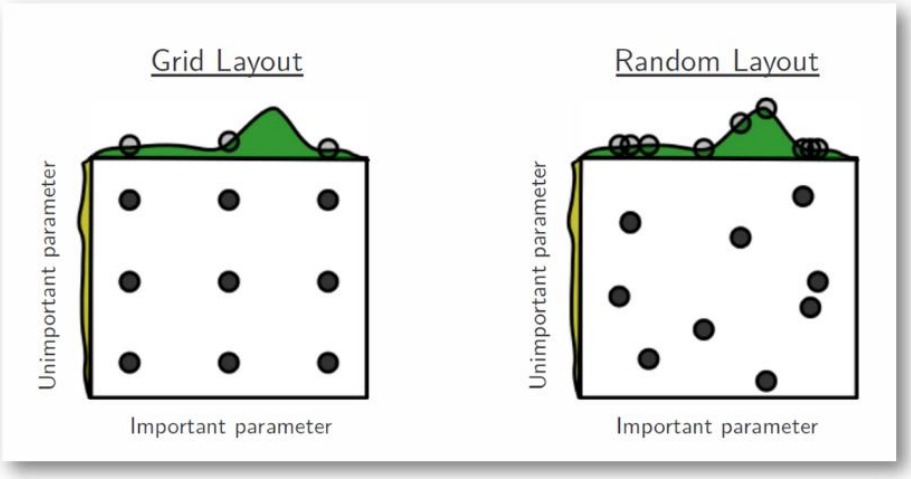

Figura 10: Rappresentazione grafica del GridSearch e RandomSearch

In particolare, gli argomenti della funzione di Random Search per il Decision Regressor Tree sono:

Tabella 3: Argomenti della funzione RandomSearchCV per il Decision Regressor Tree

estimator	pipe_tree
param_distributions	param_tree_distributions
n_iter	15
scoring	"MAE" e "R2"
refit	"R2"
cv	3
random_state	42
n_jobs	1

verbose	1

- Ottimizza la pipeline contenente il pre-processing dei dati e i modelli considerati;
- Param_distributions contiene i valori e le distribuzioni degli iperparametri selezionati;
- N_iter è il numero di combinazioni di iperparametri selezionate e valutate in modo randomico;
- Durante la cross validation sono calcolati il MAE e l'R-quadro, ma alla fine viene rifittato il modello con l'R-quadro ("R2") più alto su tutto il dataset;
- Cv = 3 è il numero di split del training set per la K-fold cross-validation;
- n_jobs = 1 indica che i calcoli verranno eseguiti da un solo core della CPU a causa della grande dimensione del dataset. La parallelizzazione del lavoro potrebbe causare un superamento del limite della RAM.

Per il modello di regressione lineare con regolarizzazione Lasso:

Tabella 4: Argomenti della funzione RandomSearchCV per la regolarizzazione Lasso

estimator	pipe_lasso
param_distributions	param_lasso
n_iter	50
scoring	"MAE" e "R2"
refit	"R2"
cv	3
random_state	42
n_jobs	1
verbose	1

L'unica differenza rispetto alla funzione dell'albero decisionale è nell'aumento delle iterazioni (50) compensato dal numero ridotto di iperparametri da trovare (param_lasso).

La selezione degli iperparametri (param_distributions e param_lasso) riguarda il modello di regressione lineare con regolarizzazione Lasso e il Decision Regressor Tree. La regressione polinomiale (senza regolarizzazione) non richiede la selezione di iperparametri.

In particolare, param_lasso:

Tabella 5: Distribuzioni utilizzate per la ricerca degli iperparametri della regolarizzazione Lasso nella funzione RandomSearchCV

model__alpha	loguniform(1e-5, 1e1)

- Alpha (o lamba) controlla la forza della regolarizzazione nella regressione lineare.

Invece, param_tree_distributions:

Tabella 6: Distribuzioni utilizzate per la ricerca degli iperparametri del Decision Regressor Tree nella funzione RandomSearchCV

model__criterion	["squared_error", "friedman_mse"]
model__max_depth	randint(2, 15)
model__min_samples_split	randint(10, 30)
model__min_samples_leaf	randint(5, 20)
model__max_features	[None, "sqrt", "log2", 0.5, 0.7, 1.0]
model__splitter	['best', 'random']
model__ccp_alpha	loguniform(1e-5, 1e-1)

- criterion: la funzione per misurare la qualità dello split nella cross-validation. Squared_error minimizza la varianza residua, mentre friedman_mse utilizza il mean squared error con dei miglioramenti per potenziali split;
- max_depth: la massima profondità oltre la quale l'albero decisionale non può andare (da regolare per evitare overfitting);
- min_samples_split: definisce il numero minimo di campioni necessari per suddividere un nodo interno. Se il numero di campioni in un nodo è inferiore a min_samples_split, il nodo non verrà suddiviso e si trasformerà in un nodo foglia;
- min_samples_leaf: il numero minimo di campioni che devono trovarsi in un nodo foglia;

- max_features: il numero di features da considerare quando si cerca il migliore split. Ad esempio, se “sqrt”, allora $max\_features = \sqrt{n\_features}$;
- splitter: la strategia utilizzata per scegliere lo split in ciascun nodo. Le opzioni sono "best" per scegliere lo split migliore e "random" per scegliere lo split casuale migliore;
- ccp_alpha: parametro di complessità utilizzato per Minimal Cost-Complexity Pruning. Verrà scelto il sottoalbero con la maggiore complessità di costo, inferiore a ccp_alpha.

3.4. Metriche di performance

In questa tesi, sono state considerate due metriche di performance per valutare i risultati ottenuti: l’R^2 e il MAE.

L’$\boldsymbol{R^2}$ è una misura statistica che rappresenta quanto si adatta bene il modello di regressione ai dati. Il valore varia tra 0 e 1. Se l’R^2 è uguale a 1, il modello è perfettamente adattato ai dati e non c’è nessuna differenza tra il valore predetto e il valore reale. Invece, se il valore è uguale a 0, il modello non predice nessuna variabilità e non impara nessuna relazione tra le variabili dipendenti e indipendenti. L’R^2è calcolato comparando la Sum of Squared of Errors (SSE) o il Sum of Squared Residuals (SSR) al Total Sum of Squared (SST). L'SSE è la somma delle differenze al quadrato tra i valori effettivi della variabile dipendente e i valori previsti dal modello di regressione. Rappresenta la variabilità non spiegata dalle variabili indipendenti. L'SST è la variazione totale della variabile dipendente e si calcola sommando le differenze al quadrato tra ciascun valore effettivo della variabile dipendente e la media di tutti i valori della variabile dipendente. Una volta calcolati SSE e SST, l’R-quadro viene determinato dividendo SSE per SST e sottraendo il risultato da 1. Il valore risultante rappresenta la proporzione della variazione totale nella variabile dipendente che è spiegata dalle variabili indipendenti:

Equazione 3: Formula dell'R-quadro

$$R^2 = 1 - \frac{SSE}{SST}$$

Per stabilire se un modello predittivo è stato adattato correttamente con un valore di R-quadro, è necessario considerare prima altri fattori come l'errore assoluto medio (MAE).

Il **MAE** è una metrica comunemente utilizzata per valutare l’accuratezza delle previsioni. Il MAE è la differenza in valore assoluto tra le previsioni di un modello e il valore reale, calcolata come media sull'intero set di dati. L'errore medio assoluto è ponderato su una scala lineare e quindi non attribuisce più importanza agli outliers. Questo fornisce una misura più uniforme delle prestazioni, ma significa

che errori grandi e piccoli sono ponderati allo stesso modo. Il MAE è la media dei valori assoluti degli errori del modello:

Equazione 4: Formula del MAE

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

Dove:

- n è il numero di dati,
- x_i è la previsione,
- x è il valore reale.

Per comprendere meglio quali sono i fattori che influenzano maggiormente la variabile dipendente, sono stati utilizzati i metodi SHAP e il Lasso Plot, complementari alle suddette metriche di performance.

Shapley Additive exPlanations (**SHAP**) consente di fornire spiegazioni sia locali che globali su come le singole features influenzano l'output del modello. Il metodo di interpretazione è model-agnostic, ovvero può essere usato per tutti i modelli di ML. In base al tipo di modello, ci sono diversi tipi di explainer; per questa tesi è stato utilizzato il Linear Explainer per il modello di regressione: utilizza una formula analitica per calcolare i contributi di Shapley, che si basa sulla decomposizione di Shapley, il quale suddivide il contributo di ciascuna variabile di input in un termine di "peso" e un termine di "discrepanza", che misura l'effetto della variabile sulla predizione del modello.

I valori SHAP hanno diverse proprietà utili che li rendono efficaci per l'interpretazione dei modelli:

- Additività: i valori SHAP sono additivi, il che significa che il contributo di ciascuna caratteristica alla previsione finale può essere calcolato indipendentemente e poi sommato. Questa proprietà consente un calcolo efficiente dei valori SHAP, anche per set di dati ad alta dimensionalità;
- Accuracy locale: i valori SHAP sommati rappresentano la differenza tra l'output previsto del modello e l'output effettivo per un dato input. Ciò significa che i valori SHAP forniscono un'interpretazione accurata e locale della previsione del modello per un dato input;
- Mancanza: i valori SHAP sono pari a zero per le caratteristiche mancanti o irrilevanti ai fini di una previsione. Questo rende i valori SHAP robusti ai dati mancanti e garantisce che le caratteristiche irrilevanti non distorcano l'interpretazione;

- Coerenza: i valori SHAP non cambiano quando il modello cambia, a meno che non cambi il contributo di una feature. Ciò significa che i valori SHAP forniscono un'interpretazione coerente del comportamento del modello, anche quando cambiano l'architettura o i parametri del modello. Nel complesso, i valori SHAP forniscono un modo coerente e oggettivo per ottenere informazioni su come un modello di machine learning effettua previsioni e quali feature hanno la maggiore influenza.

Il **Lasso plot** permette di visualizzare quali sono le features di maggiore importanza al variare dell'intensità della regolarizzazione, e quindi al variare di lamba (o alpha).

4. Risultati

In questo capitolo, saranno messi a confronto i risultati ottenuti dai modelli descritti nel capitolo precedente. Gli obiettivi possono essere riassunti in tre punti:

- Stimare i coefficienti di ogni feature presa in esame;
- Analizzare l'importanza relativa di ogni feature;
- Confrontare i risultati ottenuti dai diversi modelli.

4.1. Modello di regressione

Nel modello di regressione, la maggior parte delle variabili indipendenti sono sotto forma di logaritmo (come $\log(x)$), quindi un incremento dell'1% di x causa una variazione di $0{,}01 \cdot \beta$ delle luci notturne, a parità delle altre variabili. In presenza del termine quadratico, l'effetto marginale diventa $\beta 1 + 2 \cdot \beta 2 \cdot log(x)$. I risultati della regressione sono stati ottenuti dall'addestramento di 5 diversi modelli:

- Modello M1: considera solo log(densità di popolazione) che risulta essere statisticamente significativo con un coefficiente di 5,56 e una deviazione standard di 0,0014. Ciò significa che l'aumento dell'1% della densità di popolazione causa un aumento del 0,0556 della luminosità notturna. La variabile, da sola, spiega il 48,14% della varianza, affermandosi come la feature di maggiore importanza;
- Modello M2: rispetto al modello 1, viene anche incluso il termine quadratico di log(densità di popolazione), che assume un coefficiente positivo di 0,97, suggerendo un effetto marginale crescente rispetto alle luci notturne. Il termine lineare ha un coefficiente di 2,62, per cui si ottiene un effetto marginale pari a: $2{,}62 + 2 \cdot 0{,}97 \cdot log(Densità\ di\ popolazione)$. Ciò significa che un aumento della popolazione nelle aree ad alta densità causa un aumento maggiore di luminosità rispetto alle zone meno dense. L'R-quadro aumenta al 51,15%;
- Modello M3 (modello caratterisiche morfologiche): rispetto al modello 2, vengono incluse le variabili morfologiche della pendenza, dell'altitudine e dei laghi, con il loro termine quadratico. Tutti i loro coefficienti risultano negativi: -0.6121 l'altitudine, -0.9510 la pendenza e -1.4592 i laghi. Tuttavia, il quadrato del logaritmo della pendenza ha coefficiente positivo: la forma della relazione è convessa, cioè le zone più pendenti sono quelle meno luminose, nonostante il coefficiente positivo del termine quadratico riduca la penalità per le aree con maggiore pendenza. Questi risultati sono coerenti: il terreno non è edificabile nei laghi, nelle zone ad alta quota ed oltre una certa pendenza. Il coefficiente della densità diminuisce a 1,87 rispetto al modello 2, poiché una parte delle caratteristiche fisiche, spiegate dalla sola variabile della densità nel modello 2, è ora "assorbita" dalle variabili morfologiche. L'R-quadro aumenta al 53,55%;

- Modello M4 (modello caratteristiche socioeconomiche): rispetto al modello M2, include il Log(Reddito medio) e il Log(Numero esercizi ricettivi) con il loro termine quadratico. L'impatto è positivo e statisticamente significativo, con coefficienti di 1,0129 e 0,1959 rispettivamente. Il segno è coerente: più è alto il reddito medio e il turismo in una data area, più aumenta la sua luminosità. Nonostante l'effetto significativo, rispetto al modello 2, il MAE peggiora più nel modello 4 che nel modello 3: ciò dimostra che probabilmente c'è un effetto di collinearità tra la densità e le altre due variabili, e la variabilità dell'output è spiegata per la gran parte dalla prima. Inoltre, il coefficiente della densità di popolazione si riduce meno dal modello 2 al 3 (da 2,62 a 1,87), rispetto che dal 2 al 4 (da 2,62 a 2,31): ciò suggerisce che, nonostante la parte "socio-economica" abbia un effetto sulla spiegazione della variabilità dell'output del modello, la parte morfologica risulta essere più impattante. Ciò è coerente con gli studi di Henderson et al. (2017), secondo i quali gran parte della variabilità delle luci notturne è spiegata da "effetti fissi" spaziali;
- Modello M5: include tutte le feature: densità di popolazione, reddito medio, pendenza, laghi, numero di esercizi ricettivi, altitudine e rispettivi termini quadratici. L'R-quadro ottenuto è del 54,31% (migliore tra le regressioni). I risultati riconfermano che:
    - L'urbanizzazione, lo sviluppo economico e il turismo hanno un effetto positivo sulle luci notturne, dimostrato dai coefficienti positivi della densità di popolazione, numero di esercizi ricettivi e reddito medio;
    - L'altitudine ha un effetto negativo sulla luminosità: più un luogo si trova ad alta quota, meno è illuminato;
    - I laghi hanno un effetto negativo sulla luminosità: non essendo edificabili, non sono luminosi;
    - La pendenza, come accennato prima, mantiene il suo effetto negativo.

Nonostante la grande asimmetria della distribuzione della variabile delle luci notturne, il MAE del modello M5 è di circa 2,06.

Tabella 7: Coefficienti risultati dal modello di regressione polinomiale

Features	M1	M2	M3	M4	M5
Log(Densità popolazione)	5.5609***	2.6195***	1.8754***	2.3181***	1.9211***
	(0.0014)	(0.0029)	(0.0048)	(0.0047)	(0.0031)
Log(Densità popolazione)^2		0.9677***	1.1099***	1.0129***	1.0462***

		(0.0009)	(0.0014)	(0.0014)	(0.0009)
Log(Reddito medio)				0.4939***	0.3221***
				(0.0025)	(0.0016)
Log(Reddito medio)^2				0.0130***	0.0088***
				(0.0002)	(0.0001)
Log(Numero esercizi ricettivi)				0.1959***	0.2735***
				(0.0025)	(0.0016)
Log(Numero esercizi ricettivi)^2				0.2664***	0.1954***
				(0.0013)	(0.0008)
Log(Altitudine)			-0.6121***		-0.3071***
			(0.0044)		(0.0044)
Log(Altitudine)^2			-0.1422***		0.0436***
			(0.0017)		(0.0013)
Pendenza			-0.9510***		-1.0240***
			(0.0040)		(0.0025)
(Pendenza)^2			0.5180***		0.4607***
			(0.0021)		(0.0014)
Laghi (binaria)			-1.4592***		-1.8505***
			(0.0256)		(0.0160)
Intercetta	3.2174***	2.2497***	1.7465***	1.9296***	1.4757***
	(0.0014)	(0.0016)	(0.0035)	(0.0028)	(0.0025)
N	19725849	19725849	19725849	19725849	19725849
R-quadro (test)	0.4814	0.5115	0.5355	0.5223	0,5435
R-quadro (train)	0.4901	0.5274	0.543	0.5338	0,5536
MAE (test)	2.0021	1.9792	2.0032	2.1487	2,0641

Dall'analisi SHAP è possibile visualizzare l'impatto positivo o negativo di ogni feature sull'output del modello. In particolare, la feature di maggiore rilevanza è il quadrato del logaritmo della densità di popolazione (valore SHAP medio di 1,54) con un impatto fortemente positivo, seguita dal suo termine lineare (1,18). Dopo la densità, la pendenza incide maggiormente sull'output, come dimostrato precedentemente dall'analisi di regressione del modello 3. Di seguito ci sono il reddito e il numero di esercizi, coerenti con i risultati ottenuti dal modello 4. Tutti gli altri valori SHAP

supportano i risultati ottenuti dalla regressione: molte variabili hanno un effetto relativamente basso sull'output del modello, ma hanno un'elevata significatività.

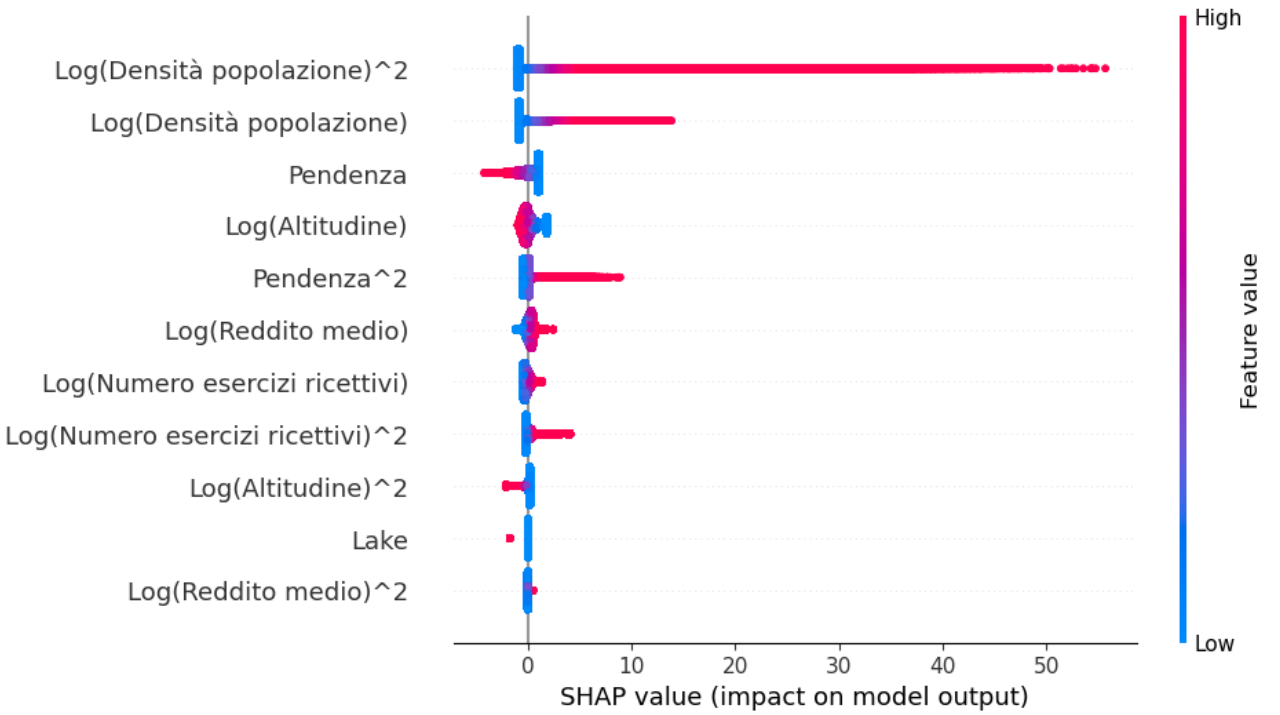

Figura 11: Valori SHAP delle features del modello

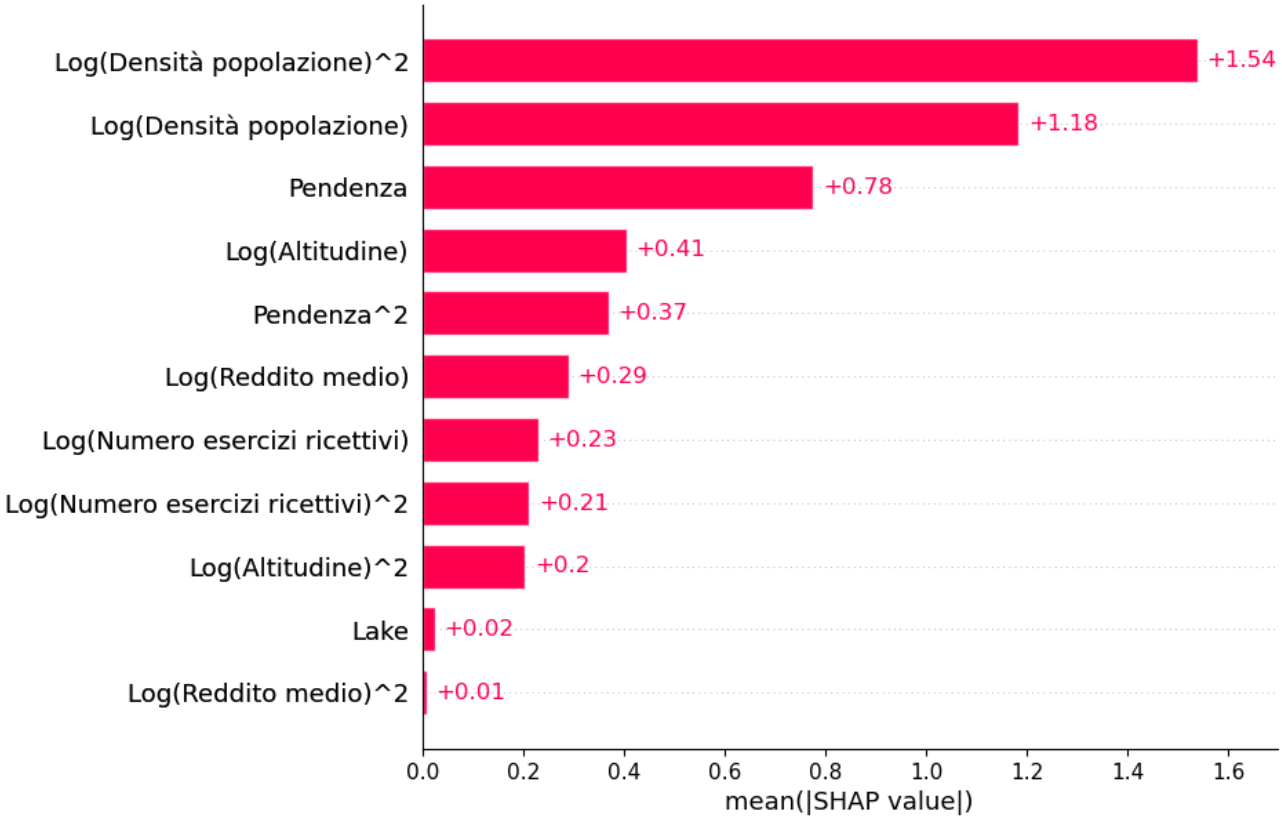

Figura 12: Media dei valori SHAP delle features del modello

4.2. Regressione Lasso

La regolarizzazione Lasso (o L1) è stata utile per prevenire l'overfitting, selezionare le feature più impattanti e capirne l'importanza relativa. A differenza della regressione polinomiale, in questo modello sono stati considerati solo i termini lineari, e quindi sono stati omessi quelli quadratici. L'R-quadro del modello risulta di 0,4989 sul dataset di test, molto simile al risultato ottenuto dalla regressione polinomiale. Il valore di alpha, trovato dal RandomSearchCV, risulta essere circa 1,33e-05. È possibile capire l'importanza relativa in base al momento dell'"entrata" della variabile dovuta al decremento dell'intensità della penalizzazione (alpha):

1. Densità di popolazione: entra ad alpha=5,25 (-log10(alpha)=-0,720);

2. Altitudine: entra ad alpha=0,851 (-log10(alpha)=0,070);

3. Pendenza: entra ad alpha=0,687 (-log10(alpha)=0,163);

4. Numero esercizi ricettivi: entra ad alpha=0,6512 (-log10(alpha)=0,186);

5. Reddito medio: entra ad alpha=0,3813 (-log10(alpha)=0,419);

6. Laghi.

Anche dal Lasso plot emerge che la densità di popolazione ha l'impatto maggiore rispetto alle altre variabili. Seguono l'altitudine, la pendenza, il numero di esercizi ricettivi, il reddito medio e i laghi. Le variabili morfologiche dimostrano di avere un effetto maggiore sull'output rispetto a quelle socioeconomiche, come precedentemente dimostrato nei modelli di regressione 3 e 4.

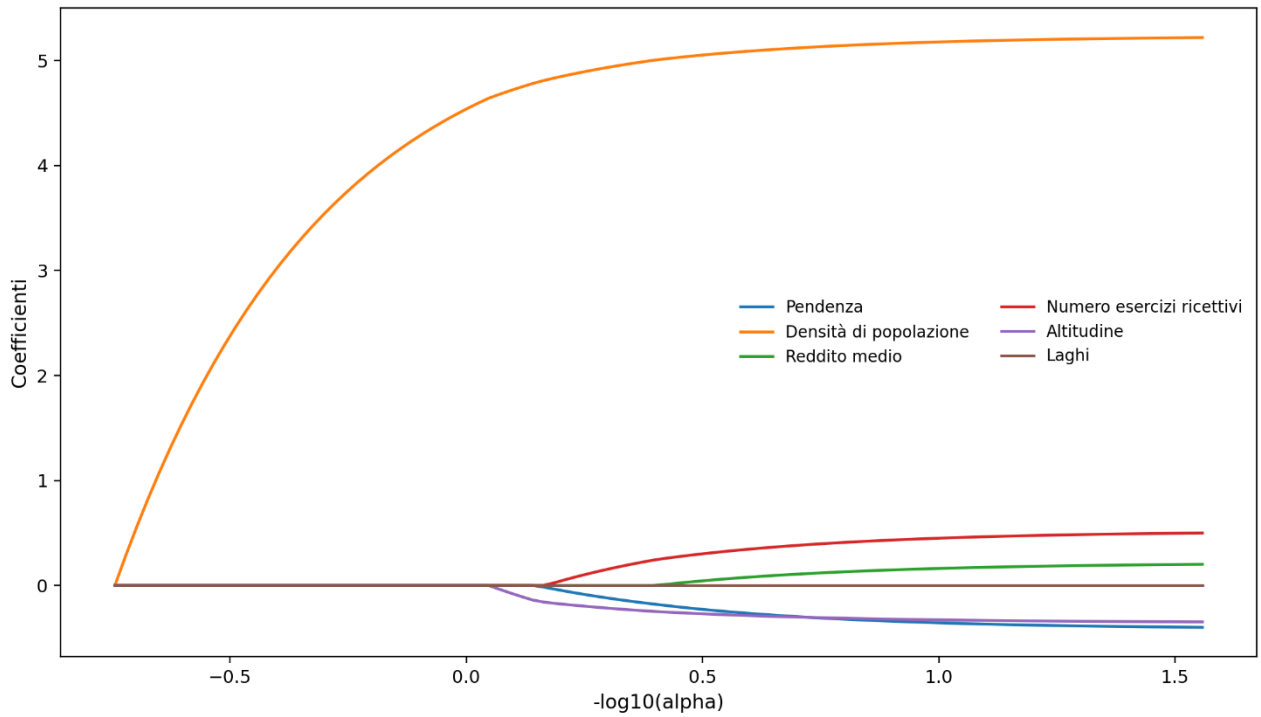

Figura 13: Percorsi dei coefficienti Lasso

4.3. Albero decisionale

L'albero decisionale permette di avere una visione differente sull'importanza delle feature rispetto la variabile delle luci notturne, determinata dall'ordine degli split all'interno dell'albero. In questa analisi, il risultato ottenuto dal modello è molto simile a quello ottenuto dai modelli precedenti: un R-quadro di 0,5294 sul dataset di test e 0,6681 sul dataset di train. La differenza tra i due risultati suggerisce un certo grado di overfitting. Il MAE è di 1,8976 sul dataset di test, risultato migliore rispetto a quello ottenuto dai modelli di regressione.

L'albero conferma nuovamente la densità di popolazione come feature principale del modello: i primi rami destro e sinistro si dividono in "alta" (> 11.142) e "bassa" (≤ 11.142) densità popolativa. Dopo il secondo livello, a sua volta diviso in base alla densità di popolazione, entrano in gioco altre variabili: la pendenza al terzo livello (soglia 0,038), il numero di esercizi ricettivi e l'altitudine al quarto livello.

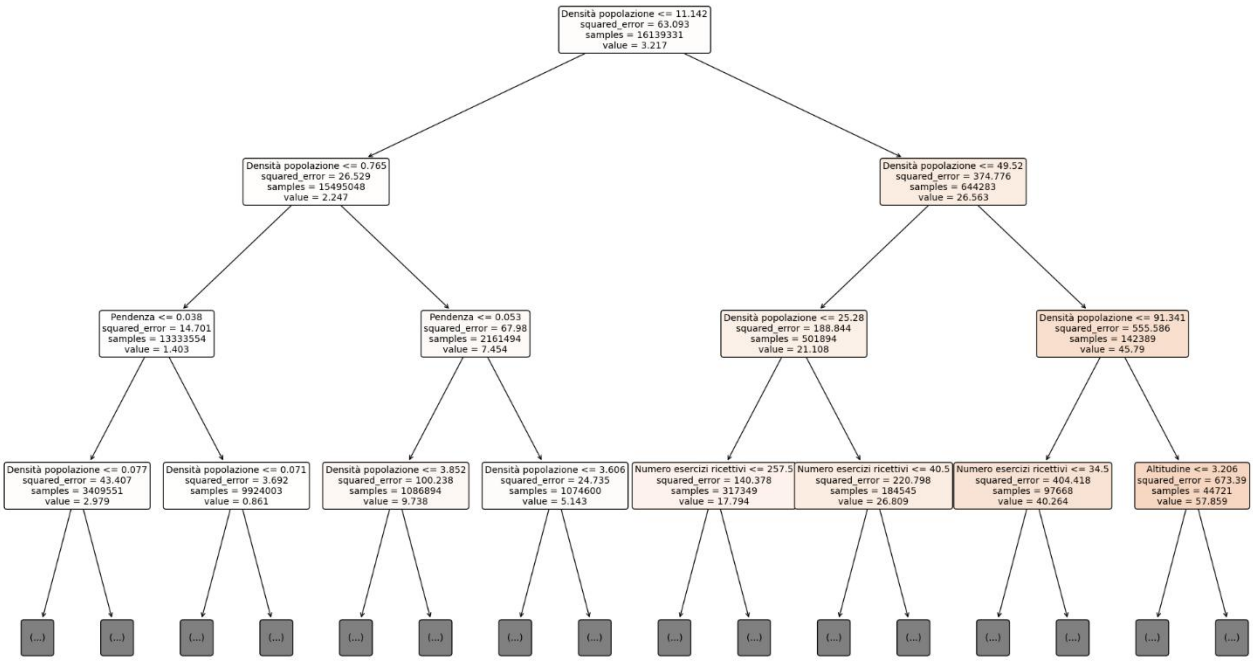

Figura 14: Primi tre livelli del Decision Regressor Tree

In sintesi, la densità di popolazione risulta essere la variabile di maggiore impatto sulla luminosità delle luci notturne, seguita dalla pendenza. Nonostante l'effetto della maggior parte delle feature considerate sia basso, tutte hanno un impatto statisticamente significativo. I risultati migliori sono stati ottenuti dal modello di regressione polinomiale con un R-quadro di 0,54, mentre l'albero decisionale ottiene il MAE migliore di 1,897.

Conclusioni

Il presente lavoro ha permesso di esplorare l'impatto di alcune features socioeconomiche e morfologiche sulla luminosità notturna (NTL) sul suolo italiano, grazie all'addestramento di più modelli di Machine Learning e alla comparazione dei loro risultati.

Ciò che è emerso è il ruolo principale della variabile della densità di popolazione per la spiegazione della variabilità dell'output del modello: da sola, riesce a spiegare circa il 51% e, insieme alle altre variabili, si ottiene un R-quadro del 54,35%. Come ci si aspettava, il reddito medio e il numero di esercizi ricettivi hanno avuto un impatto più basso rispetto alle variabili morfologiche: grazie alla regolarizzazione Lasso, è stato possibile analizzare l'importanza relativa delle variabili, risultando che la pendenza, l'altitudine e i laghi hanno maggiore impatto sul modello rispetto alle variabili economiche. Ciò viene anche confermato dall'albero decisionale, nel quale compaiono la densità di popolazione, la pendenza, l'altitudine e il numero di esercizi ricettivi nei primi 3 livelli.

Questa ricerca focalizzata sull'Italia ha deciso di escludere alcuni effetti fissi comunali, provinciali e regionali a causa di vincoli computazionali, lasciando, probabilmente, parte della varianza non spiegata. In più, sarebbe stato più corretto utilizzare il PIL invece del reddito medio: non è stato possibile introdurre a livello di singolo pixel il valore del PIL, che sarebbe stato, secondo articoli passati, di grande impatto sull'output del modello. Proprio per i suddetti motivi, nascono molteplici direzioni per la ricerca futura, con lo scopo di creare un modello sempre più preciso.

Per concludere, questa tesi vuole anche dimostrare che i dati delle luci notturne, apparentemente considerati come privi di informazione, in realtà sono una fonte di notevole valore scientifico. Infatti, nella ricerca scientifica, la conoscenza non risiede dove tutti puntano gli occhi, ma bensì dove è necessaria una visione più profonda, con lo scopo di dare un senso a quello che prima appariva come semplice "rumore".

Bibliografia, sitografia e citazioni

Elvidge, C. D., Baugh, K., Zhizhin, M., Hsu, F. C. and Ghosh, T. (2017). VIIRS night-time lights. International Journal of Remote Sensing, 38(21), 5860–5879

Gibson, J., Olivia, S. and Boe-Gibson, G. (2020). Night lights in economics: sources and uses. Journal of Economic Surveys, 34(5), 955–980

Gibson, J., Olivia, S., Boe-Gibson, G. and Li, C. (2021). Which night lights data should we use in economics, and where? Journal of Development Economics, 149, 102602

Gibson, J. (2021). Better night lights data, for longer. Oxford Bulletin of Economics and Statistics, 83(3), 770–791

Vernon Henderson, Tim Squires, Adam Storeygard, David Weil. THE GLOBAL DISTRIBUTION OF ECONOMIC ACTIVITY: NATURE, HISTORY, AND THE ROLE OF TRADE

J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring Economic Growth from Outer Space

Charlotta Mellander, José Lobo, Kevin Stolarick, Zara Matheson. Night-Time Light Data: A Good Proxy Measure for Economic Activity?

The case of Morocco - Mark Roberts. Tracking economic activity in response to the COVID-19 crisis using nighttime lights

L. Buzzacchi, A.M. De Marco, F.L. Milone. Estimating the economic impact of the russo-ukrainian conflict with nightlights data

Arsid Pambuku, Mario Elia, Alessandro Gardelli,, Vincenzo Giannico, Giovanni Sanesi, Angela Stefania Bergantino, Mario Intini, Raffaele Lafortezza. Assessing urbanization dynamics using a pixel-based nighttime light indicator.

Matteo Marcantonio, Sajid Pareeth, Duccio Rocchini, Markus Metz, Carol X. Garzon-Lopez, Markus Neteler. The integration of Artificial Night-Time Lights in landscape ecology: A remote sensing approach

https://pygis.io/docs/d_crs_what_is_it.html

https://www.istat.it/

https://www.finanze.gov.it/it/

https://tinitaly.pi.ingv.it/

https://scikit-learn.org/

https://www.diariodiunanalista.it/

https://www.yimp.it/

https://pulplearning.altervista.org/

https://it.mathworks.com/

Ringraziamenti

Per prima cosa, vorrei ringraziare il mio relatore, Professor Francesco Milone, per la sua disponibilità e per avermi supportato nella stesura di questo lavoro.

Vorrei ringraziare la mia famiglia, che ha reso possibile il mio percorso universitario e mi ha sempre sostenuto.

Un caloroso abbraccio va a tutti i miei amici: Jacopo, Francesco, Nico, Luca, Bruno, Mattia, Manchi, Caste, Alex, Ferro, Barni, Pg, il gruppo Comple Marti, il gruppo Skollonisi e il gruppo Asse Pinerolo Napoli. Senza di voi, tutto ciò non sarebbe stato possibile.