UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI INFORMATICA

Corso di laurea in Informatica e Tecnologie per la produzione del software

Tesi di laurea

PREDIZIONE DELLA RADIAZIONE SOLARE ATTRAVERSO ALGORITMI DI MACHINE LEARING

Relatore:
Prof. Donato Impedovo

Laureando:
Sorrenti Simone

ANNO ACCADEMICO 2017-2018

“L'Intelligenza Artificiale sarà la più importante conquista dell'uomo, peccato che potrebbe essere l'ultima.”

Stephen Hawking

Sommario

Prefazione

Oggigiorno le aziende agricole sono in grado di raccogliere ingenti quantità di dati, grazie soprattutto ai recenti progressi nel campo tecnologico, i quali hanno permesso di ridurre le incertezze riguardo i processi decisionali, ad esempio quali tipologie di colture piantare, in quale periodo ararle o irrigarle, etc.
Tali decisioni influenzano fortemente il tipo, la qualità e la quantità della produzione agricola e, di conseguenza, il guadagno economico dell'azienda; questi ultimi sono influenzati da fattori meteorologici, dalle precipitazioni alla temperatura dell'aria, dalla pressione atmosferica alle radiazioni solari.

Le variabili agrometeorologiche sopracitate possono essere misurate e raccolte facilmente, e ciò è reso possibile dalla rivoluzione tecnologica che stiamo vivendo, conosciuta come *Internet of things* (IoT). L'Internet delle cose permette agli oggetti di connettersi tra di loro in ogni parte del mondo: questa connessione gli permette di connettersi, identificarsi, localizzarsi, di elaborare dati, fornendo numerosi servizi innovativi.

Nelle coltivazioni questa tecnologia, in particolare l'utilizzo di sensori ad alta precisione connessi tra di loro, ha portato alla nascita di una strategia gestionale denominata agricoltura di precisione.

L'agricoltura di precisione consiste nella raccolta, nell'elaborazione e nell'analisi dei dati in tempo reale, i quali, insieme a tecnologie di automazione, consentono di massimizzare la produttività, servendosi di un utilizzo più mirato dei fertilizzanti e dei pesticidi, e un'irrigazione di maggior precisione per le colture.

Un'irrigazione efficace influenza la quantità di acqua da utilizzare per le colture, la quale è costituita da due processi ben specifici: l'*evaporazione*, nella quale l'acqua viene convertita in vapore acqueo (vaporizzazione), e la *traspirazione*, che consiste nella vaporizzazione dell'acqua contenuta nei tessuti vegetali.

Per poter determinare indirettamente *l'evapotraspirazione* attraverso altri dati meteorologici è possibile utilizzare l'equazione di *Penman-Monteith*, che è stata raccomandata *dall'Organizzazione delle Nazione Unite per l'alimentazione e l'agricoltura* (FAO): l'equazione permette di determinare l'evapotraspirazione potenziale, la quale è in funzione di quattro parametri

meteorologici, ossia temperatura dell'aria, umidità relativa, velocità del vento e radiazione solare.

Pertanto, un'irrigazione efficace ed efficiente è dettata dalla conoscenza dell'evapotraspirazione necessaria per i giorni futuri, e, per poterla stimare, bisogna essere in possesso dei parametri sopra citati: è dunque necessario predire i valori dei parametri meteorologici d'interesse per i giorni futuri.

In questo studio ci si è focalizzati sulla predizione della radiazione solare per diversi orizzonti temporali (orario, giornaliero, mensile). Tale tematica viene costantemente studiata e discussa a causa della sua imprevedibilità e variabilità, in quanto la radiazione solare è governata da fenomeni distribuiti e difficili da prevedere.

Attraverso una ricerca bibliografica sulla predizione della radiazione solare per orizzonti temporali multipli si evince che gli algoritmi di *machine learning* sono i più adatti: in particolare gli algoritmi più utilizzati, più accurati e più semplici da implementare nelle predizioni della radiazione solare, sono le *support vector machines*, le reti neurali artificiali *multi-layer perceptron* e *long short-term memory*.

Per individuare la combinazione di parametri meteorologici migliore, in modo tale da ottenere predizioni il più possibile accurate in input ai modelli predettivi, sono stati applicati algoritmi di *feature selection*, i quali permettono di selezionare le caratteristiche più rilevanti da un grande dataset. Vengono quindi utilizzati i seguenti algoritmi di *feature selection*: *pearson correlation, recursive feature elimination con SVM e random forest*.

Infine, mediante l'uso di tale studio si è cercato di identificare il miglior algoritmo di *machine learning* e la miglior combinazione di parametri meteorologici da fornire in input a tale algoritmo, assicurando predizioni della radiazione solare più accurate per ogni orizzonte temporale.

Capitolo 1: Agricoltura

1. Azienda agricola

Il termine *azienda agricola* descrive un complesso di beni organizzati dall'imprenditore agricolo per l'esercizio della sua attività. Ed è quindi con riferimento al concetto di attività di impresa agricola che trova la sua definizione. L'articolo 2135 codice civile qualifica attività agricole la coltivazione del fondo, la selvicoltura, l'allevamento di animali e le attività connesse, precisando che:

“per coltivazione del fondo, per selvicoltura e per allevamento di animali si intendono le attività dirette alla cura ed allo sviluppo di un ciclo biologico o di una fase necessaria del ciclo stesso, di carattere vegetale o animale, che utilizzano o possono utilizzare il fondo, il bosco o le acque dolci, salmastre o marine”. [1]

2. Tipologie di aziende e coltivazioni

È possibile individuare quattro principali tipologie di aziende agricole considerando i diversi fini della produzione:

- Coltivazioni di colture non permanenti: non durano più di due stagioni agricole, pertanto dipendono da condizioni climatiche e meteorologiche;
- Coltivazioni di colture permanenti: durano più di due anni, e pur morendo stagionalmente ricrescono in modo costante;
- Riproduzione delle piante:
- Attività di supporto all’agricoltura o successive alla raccolta. [2]

Inoltre, è possibile suddividere le coltivazioni nella seguente maniera (tabella 1): [3]

COLTIVAZIONI NON PERMANENTI	COLTIVAZIONI PERMANENTI
cereali, legumi e semi oleosi	Uva
ortaggi, meloni, radici e tuberi	Frutta tropicale e subtropicale
canna da zucchero	Agrumi
riso	pomacee e frutta a nocciolo
tabacco	frutti oleosi
piante tessili	piante per la produzione di bevande
floricoltura	piante aromatiche e farmaceutiche

Tabella 1: Tipologie di coltivazioni

3. Processo decisionale nell'agricoltura

Un responsabile di un'azienda agricola si occupa di avviare la maggior parte delle attività quotidiane di una coltivazione, prendendo delle decisioni. Tale decisioni sono chiamate *decisioni operative,* ed influenzano fortemente il tipo, la qualità e la quantità della produzione agricola e, di conseguenza, il guadagno economico dell'azienda.

Anche decidere di non fare nulla è una decisione e ha un impatto.

Quanti più agricoltori assumono un ruolo responsabile nei processi decisionali agricoli, tanto più è probabile che l'azienda agricola sarà sostenibile e redditizia.

Prendere delle decisioni esatte, per degli agricoltori, è sempre stata una fase importante e critica nel processo decisionale agricolo.

L'importanza nel prendere tali decisioni è aumentata, a causa della rivoluzione tecnologica che si sta vivendo, e della concorrenza commerciale nel settore agricolo.

Esempi di alcune delle decisioni quotidiane e fondamentali per il processo agricolo che gli agricoltori sono tenuti a fare sono: quali colture piantare, quando arare, quando seminare, quando irrigare o quanto raccolto vendere. [7]

Oggigiorno le aziende agricole sono in grado di raccogliere ingenti quantità di dati, grazie soprattutto ai recenti progressi nel campo tecnologico, i quali hanno permesso di ridurre le incertezze riguardo i processi decisionali.

La mancanza di dati disponibili, è evidente, non è più un problema per l'agricoltura moderna e, pertanto, non impedirà il suo progresso.

Piuttosto, la nuova sfida è il saper individuare quali dei tanti dati raccolti sono i più rilevanti, per poter ottimizzare il processo aziendale agricolo.

Pertanto, gli agricoltori devono essere capaci di comprendere di quali informazioni da raccogliere necessitano e, di pensare ai costi ed alle risorse indispensabili per poter raccogliere tali informazioni, prima di prendere delle decisioni. [8]

4. Ciclo di vita delle coltivazioni

Il ciclo di vita di qualsiasi coltivazione è caratterizzato da otto fasi principali (figura 1) che vanno dalla selezione del raccolto alla raccolta:

1. Selezione del raccolto
2. Preparazione del terreno
3. Selezione dei semi
4. Semina
5. Irrigazione
6. Crescita del raccolto
7. Fertilizzante
8. Raccolta

Figura 1: Ciclo di vita agricolo con le sue otto fasi

In ogni fase del processo agricolo, le aziende agricole hanno bisogno di informazioni, dalla selezione della coltura alla raccolta. Di seguito vengono elencate le informazioni necessarie in ciascuna delle fasi:

1. Selezione della coltura
    - I prezzi comparativi delle diverse colture
    - La domanda di mercato e potenziale di vendita del raccolto
    - Il budget richiesto per la coltivazione di ogni coltura
    - La fattibilità della coltura considerando il clima e la qualità del terreno
    - La produttività delle colture
2. Preparazione del terreno
    - Gli effetti di qualsiasi malattia dalla precedente coltivazione e misure necessarie per ridurre al minimo questo impatto

- I fertilizzanti necessari per portare il terreno alla sua normale fertilità a seconda delle colture precedenti e dei fertilizzanti usati
- La progettazione per un'irrigazione efficiente
- Le tecniche necessarie per livellare i campi e il loro costo

3. Selezione dei semi
    - Il prezzo e la quantità necessari per acro
    - Il rendimento medio
    - L'idoneità al terreno ed al clima
    - Il fabbisogno d'acqua
    - La resistenza alle malattie
4. Semina dei semi
    - Il tempo appropriato per seminare il seme
    - Le condizioni meteorologiche ottimali per seminare
    - Il miglior metodo per la semina dei semi
    - La profondità di semina
5. Irrigazione
    - Il momento critico per l'irrigazione
    - La quantità di acqua da fornire alle piante
    - Frequenza di irrigazione
6. Crescita della coltivazione
    - Il numero di piante per unità di superficie
    - Il tasso di crescita medio del raccolto in condizioni normali
    - Gli interventi necessari per mantenere la crescita prevista
    - Il tempo, la frequenza e il metodo corretto per l'aratura
    - Gli attacchi di parassiti e virus possibili con i sintomi di tali attacchi e le misure precauzionali che possono essere prese in anticipo per evitare questi attacchi
    - I pesticidi e la quantità da utilizzare per uccidere i parassiti e virus

7. Fertilizzante
    - La frequenza, la quantità e il modo di fertilizzazione
8. Raccolta
    - Il tempo e il metodo adeguati per la raccolta
    - La corretta conservazione del raccolto
    - Costo del trasporto
    - I prezzi di mercato comparativi. [4]

5. Agricoltura e meteorologia

Dal precedente paragrafo, è evidente come il clima e le condizioni meteorologiche siano fattori importanti e determinanti in diverse fasi del ciclo di vita agricolo, in particolare per le fasi della selezione del raccolto e della selezione dei semi.

Le condizioni climatiche sono uno dei fattori chiave della produttività agricola. Per esempio, i processi delle piante sono regolati da molte variabili climatiche come la temperatura, la radiazione solare, l'anidrite carbonica e la disponibilità di acqua.

Inoltre, le avversità meteorologiche possono danneggiare e causare danni alle coltivazioni, attraverso eventi climatici estremi come, le ondate di calore, le inondazioni, la siccità o le tempeste. [5]

In alcuni casi, è stato dichiarato che oltre l'80% della variabilità della produzione agricola è dovuta alla variabilità delle condizioni meteorologiche, come illustrato nella figura 2, in particolare per la disponibilità di acqua.

Le variabili agrometeorologiche che principalmente influenzano la produzione agricola sono la precipitazione, la temperatura dell'aria e la radiazione solare.

La temperatura dell'aria è la principale variabile meteorologica che regola la

velocità di crescita delle colture.
In molti casi, un aumento della temperatura dell'aria composta un incremento della velocità di crescita delle colture, però per temperature dell'aria estremamente elevate si verifica esattamente l'opposto, ovvero si ha un rallentamento della velocità di crescita delle colture.
La radiazione solare è fondamentale per le piante delle colture, in quanto fornisce ad esse l'energia necessaria per attuare specifici processi, ad esempio il processo della fotosintesi.
Le precipitazioni non influiscono direttamente su nessuno dei processi delle piante, piuttosto esse sono considerate un modificatore, che influisce indirettamente sulla crescita e sui processi di sviluppo delle colture.
Il problema della siccità delle colture è causato dai periodi caratterizzati da insufficienti precipitazioni, mentre, il problema opposto si verifica durante i periodi di abbondanti piogge causando allagamenti.
La siccità delle piante è dovuta alla combinazione di fattori, come l'evapotraspirazione, l'umidità relativa ed altri fattori vegetali e ambientali. La siccità può influire sulla velocità di crescita delle colture, aumentandola o rallentandola, a seconda del contesto.
A causa delle inondazioni o di eventi di pioggia intensa può succedere che le colture abbiano una maggiore quantità d'acqua, più del dovuto.
Annaffiare troppo le piante può causare una mancanza di ossigeno nella zona di radicazione, che è necessario per la crescita delle radici e la loro respirazione.
Applicare alle colture una quantità d'acqua più del dovuto comporta ad avere effetti simili a quelli causati dalla siccità, discussi in precedenza.
Altri fattori meteorologici che possono influire sulla produzione agricola sono la temperatura del suolo, la velocità del vento, l'umidità relativa o la temperatura del punto di rugiada.

In molte regioni, la temperatura del suolo è importante durante le prime fasi agricole, poiché colpisce la semina e la germinazione.

L'umidità relativa, la temperatura del punto di rugiada o la pressione del vapore sono fattori agrometeorologici simili, che esprimono la quantità di acqua presente nell'aria ed influenzano la traspirazione delle piante.

Il vento può anche avere un impatto multiplo sulla produzione del raccolto. Prima di tutto, può influenzare la traspirazione, inoltre può influenzare il trasporto e la distribuzione di insetti e di malattie nell'atmosfera. [6]

Variabili meteorologiche che influenzano la produzione del raccolto	temperatura dell'aria
	radiazione solare
	anidrite carbonica
	disponibilità d'acqua
	eventi meteorologici
	precipitazioni
	siccità
	umidità
	vento
	temperatura del suolo
	temperatura del punto di rugiada

Figura 2: variabili meteorologici e climatici che influenzano la produzione del raccolto

6. Smart Farming

Nella storia dell'umanità, l'agricoltura è stato uno dei più importanti settori industriali per la vita degli esseri umani, poiché mediante essa è garantita la produzione di risorse indispensabili come cibo, medicine, energia, fibra. Come qualsiasi altro settore industriale, anche l'industria agricola ha accelerato lo sviluppo, impiegando tecnologie e strumentazioni moderne. Oggigiorno, siamo vivendo una vera e propria rivoluzione tecnologia, conosciuta come *Internet of things* (IoT).

"Il termine IoT è stato coniato da Ashton nel 1999, e rappresenta il futuro di dell'informatica e della comunicazione, dove tutti gli oggetti nel mondo potranno essere connessi tra loro e condividere il loro stato e il loro ambiente, fornendo finalmente nuovi servizi innovativi per l'uomo (anche senza l'intervento umano)". [9]

Lo sviluppo di sensori capaci di effettuare misurazioni di alta precisione dell'ambiente nelle coltivazioni ha portato *all'agricoltura di precisione*. *L'agricoltura di precisione* consiste nella raccolta, nell'elaborazione e nell'analisi dei dati in tempo reale, i quali, insieme a tecnologie di automazione, consentono di massimizzare la produttività, servendosi di un utilizzo più mirato dei fertilizzanti e dei pesticidi, e un'irrigazione di maggior precisione per le colture.

L'agricoltura è altamente imprevedibile, a causa della sua grande dipendenza dalle condizioni atmosferiche e ambientali (ad esempio pioggia, temperatura, umidità, grandine), degli eventi imprevedibili (malattie, parassiti), nonché dalla volatilità dei prezzi dei mercati.

Ciò implica la necessità di framework di grandi dimensioni che sfruttano sensori, tecnologie automatiche e l'analisi dei dati, al fine di aiutare gli agricoltori, informandoli tempestivamente sulle condizioni e sui rischi delle loro

aziende agricole, al fine di prendere adeguate contromisure e proteggere le loro colture. [11]

Capitolo 2: Relazione tra l'irrigazione e la radiazione solare

1. Irrigazione delle coltivazioni

1.1 Irrigazione

L'irrigazione è l'applicazione artificiale dell'acqua al terreno ai fini della produzione agricola.
Un'irrigazione efficace influenzerà l'intero processo di crescita delle coltivazioni, dalla semina del seme alla qualità del raccolto. L'agricoltore ha un ruolo fondamentale nel decidere la quantità di acqua da fornire alle coltivazione e quando applicarla.
Decidere quali sistemi di irrigazione sono i più adatti per la vostra coltivazione richiede un'ampia conoscenza delle attrezzature, della progettazione del sistema, delle piante, della composizione del suolo e della formazione del terreno.
I sistemi di irrigazione dovrebbero assicurare la crescita delle piante riducendo al minimo gli squilibri di sale, le ustioni fogliari, l'erosione del suolo e la perdita d'acqua.
La perdita di acqua si verifica a causa dell'evaporazione, del vento, del deflusso e dell'acqua che filtra in profondità, sotto la zona della radice.

L'irrigazione consente:

- Avere maggiore flessibilità nelle operazioni poiché si ha la capacità di accedere all'acqua in qualsiasi momento.
- Produrre raccolti di qualità superiore.
- Avere sicurezza nei confronti di stagioni calde e quindi della siccità.

- Massimizzare i benefici delle applicazioni dei fertilizzanti.
- Utilizzare aree che risulterebbero meno produttive, in quanto troppo secche per coltivare.
- Coltivare produzioni fuori stagionale. [12]

1.2 Sistemi di irrigazione

Esistono numerosi e vari sistemi di irrigazione che possiamo suddividere e classificare nella seguente maniera:

- Sistemi di solco: questo sistema è composto da una serie di piccoli canali poco profondi utilizzati per guidare l'acqua lungo un pendio. I solchi sono generalmente dritti, ma possono anche essere curvi per seguire il contorno del terreno.
- Sistemi di controllo di piena o di confine: questi sistemi dividono la coltivazione in settori separati da solchi paralleli. Su terreni in forte pendenza, i solchi sono più ravvicinati e possono essere curvi per seguire il contorno del terreno.
- Sistemi di bacino: questi sistemi differiscono dai tradizionali sistemi di controllo di piena in quanto la pendenza del terreno è a livello e le estremità dell'area delle coltivazioni sono chiuse.
- Sistemi di irrigazione a perno centrale: un irrigatore a perno centrale è un sistema semovente, in cui una singola tubazione, sostenuta da torri mobili, è sospesa da 2 a 4 metri dal suolo.

  L'acqua viene pompata nel tubo centrale e mentre le torri ruotano lentamente attorno al punto di rotazione, viene irrigata una grande area circolare.

  Degli ugelli montanti su o sospesi dalla tubazione distribuiscono acqua sotto pressione mentre la tubazione ruota.

Gli ugelli sono graduati da piccoli a grandi in modo che il cerchio esterno che si muove più velocemente riceva la stessa quantità di acqua di quella più lenta che si muove all'interno.

- Sistemi di irrigazione a mano: sono una serie di sezioni di tubazioni leggere che vengono spostate manualmente.
  Le condotte laterali sono collegate ad una linea principale, che può essere portatile o interrata.
  I sistemi a mano sono spesso usati per piccole aree irregolari, in quanto i requisiti di manodopera sono più alti di quelli di tutti gli altri irrigatori.
- Sistema di irrigazione fisso: un sistema di irrigazione stazionario.
  Le tubazioni per l'approvvigionamento idrico sono generalmente fisse (di solito al di sotto della superficie del suolo) e gli ugelli sono sollevati sopra la superficie.
- Sistemi a spostamento lineare o laterale: sono simili ai sistemi a perno centrale, tranne per il fatto che la linea laterale e le torri si muovono in un percorso rettilineo continuo attraverso un campo rettangolare.
  L'acqua può essere fornita da un tubo flessibile o pressurizzata da un pozzo rivestito di calcestruzzo lungo il bordo del campo.
- Sistemi di irrigazione a basso flusso: utilizzano tubi di piccolo diametro posti sopra o sotto la superficie del terreno.
  Applicazioni frequenti e lente dell'acqua vengono applicate al terreno attraverso piccoli fori.
  L'acqua viene erogata direttamente nella zona della radice riducendo al minimo l'evaporazione.

A seguito dell'enorme diversità nei tipi di tecnologie e sistemi di irrigazione esistenti, per poter decidere qual è il miglior sistema di irrigazione da utilizzare bisogna considerare diversi fattori, ovvero:

- La tipologia del terreno
- La topografia del terreno
- La disponibilità di fonti di energia
- La disponibilità di acqua
- Le fonti d'acqua
- La dimensione dell'area da irrigare
- Le risorse finanziarie
- La disponibilità di manodopera. [12]

2. Fattori che incidono sulla programmazione dell'irrigazione

La pianificazione dell'irrigazione è il processo mediante il quale un irrigatore determina il tempo e la quantità di acqua da applicare alla coltura. La sfida consiste nel valutare il fabbisogno idrico delle colture per le diverse fasi di crescita e le condizioni climatiche.

Per evitare l'eccesso o l'inumidimento, è importante sapere quanta acqua è disponibile per la pianta e quanto efficientemente la pianta può usarla. I metodi disponibili per determinare se una coltura ha bisogno d'acqua includono:

- L'osservazione delle piante: cambiamenti visibili nelle caratteristiche della pianta, come il colore delle foglie, l'arricciamento delle foglie e infine l'appassimento possono indicare lo stress idrico delle piante e quindi la necessità di irrigazione.

- L’aspetto del suolo: l’osservazione del suolo viene utilizzata per monitorare i livelli di umidità delle colture.
  Un campione del suolo può essere ottenuto usando una sonda per il terreno, una trivella o una vanga.
- L’utilizzo di dispositivi di monitoraggio dell'umidità del suolo: L’umidità del suolo può essere misurata ed è un fattore utile per i coltivatori da utilizzare nella pianificazione delle loro irrigazioni.
- Stimare l'acqua disponibile dai dati meteorologici: esistono due sistemi di pianificazione basati sul clima e sugli eventi meteorologici, usati per misurare la quantità di acqua persa da un raccolto. Questi sono:
  - L’evaporazione da una superficie di acqua
  - Dati climatici storici come l’umidità relativa, la temperatura, la velocità del vento e le ore di sole. [12]

3. L’evapotraspirazione

La quantità di acqua persa, e quindi necessaria da fornire alle colture, è causata da due processi:

- L’evaporazione: è il processo mediante il quale l’acqua viene convertita in vapore acqueo (vaporizzazione).
- La traspirazione: consiste nella vaporizzazione dell’acqua contenuta nei tessuti vegetali. Le colture perdono prevalentemente la loro acqua attraverso gli stomi. Queste sono piccole aperture sulle foglie della pianta attraverso cui passano i gas e il vapore acqueo.

L'evaporazione e la traspirazione avvengono simultaneamente e non esiste un semplice modo per distinguere i due processi.

L'evaporazione dell'acqua dal terreno è principalmente causata della radiazione solare che raggiunge la superficie del suolo.

La radiazione che raggiunge il suolo diminuisce rispetto al periodo di crescita delle piante, infatti man mano che le piante crescono coprono sempre più la superficie del terreno, creando ombra.

Pertanto, quando il raccolto è nelle prime fasi di crescita, l'acqua è prevalentemente persa a causa della evaporazione dal suolo, ma una volta che il raccolto è ben cresciuto e copre completamente il terreno, la traspirazione diventa il processo principale, tale procedimento è illustrato graficamente mediamente la figure 3 e la figura 4:

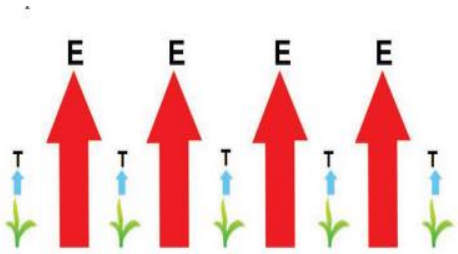


Figura 3: Grandezza relativa delle componenti dell'evapotraspirazione prima del periodo di crescita. [13]

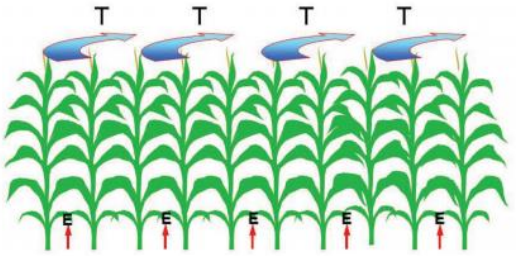


Figura 4: Grandezza relativa delle componenti dell'evapotraspirazione durante il periodo di crescita. [13]

I fattori che principalmente influenzano l'evapotraspirazione sono:

- Parametri meteorologici: poiché lo scopo principale della traspirazione è raffreddare la pianta, è prevedibile che le condizioni climatiche sono

le forze trainanti per tale processo.

La temperatura dell'aria e la radiazione solare sono i due parametri meteorologici principali che influiscono sulla velocità con cui avviene la traspirazione.

Quando la temperatura dell'aria e la radiazione solare aumentano anche la traspirazione aumenta.

Inoltre, l'evapotraspirazione aumenta con l'aumentare della velocità del vento, però quando la velocità del vento diventa troppo elevata gli stomi si chiudono come naturale meccanismo di difesa e di conseguenza la traspirazione diminuisce.

Al contrario, con l'aumentare dell'umidità relativa, la traspirazione diminuisce.

- Caratteristiche delle colture: il tipo di coltura, la varietà e lo stadio di sviluppo devono essere considerati nel valutare l'evapotraspirazione. Differenze di resistenza alla traspirazione, altezza del raccolto, copertura del terreno e caratteristiche di radicazione delle colture producono diversi livelli di ET in diversi tipi di colture in condizioni ambientali identiche.
- Gestione e aspetti ambientali: fattori quali la salinità del suolo, la scarsa fertilità del terreno, l'applicazione limitata di fertilizzanti, la presenza di terreno impenetrabile, l'assenza di controllo di malattie e parassiti e una cattiva gestione del suolo possono limitare lo sviluppo del raccolto e ridurre l'evapotraspirazione. [13] [14]

4. Metodi per determinare l'evapotraspirazione

Ci sono vari metodi per calcolare l'evapotraspirazione, per esempio:

- Misurazione diretta attraverso i lisimetri, dispositivi concepiti per la rilevazione dell'evapotraspirazione.
- Indirettamente da dati meteorologici: radiazione solare, temperatura dell'aria, velocità del vento e umidità possono essere utilizzati per calcolare l'ETo, (chiamato anche ET potenziale) ovvero l'evapotraspirazione potenziale.
- Stimare il consumo d'acqua attraverso i dati degli evaporimetri.

Per poter determinare indirettamente l'evapotraspirazione attraverso altri dati meteorologici è possibile utilizzare il metodo Penman-Monteith. Tale metodo è raccomandato dalla 'Organizzazione delle Nazione Unite per l'alimentazione e l'agricoltura' (FAO), inoltre è stato testato a livello internazionale e può essere utilizzato in un'ampia varietà di ambienti. [15] L'equazione di Penman-Monteith è la seguente:

$$ET_0 = \frac{0{,}408\,\Delta\,(Rn - G) + \gamma\left(\frac{900}{T_k}\right)U_2(e_a - e_d)}{\Delta + \gamma\,(1 + 0{,}34\,U_2)}$$

Dove:

- ETo = evapotraspirazione di riferimento (mm d^{-1})
- Δ = pendenza della curva che esprime la tensione di vapore saturo in funzione della temperatura (kPa $°C^{-1}$)
- R_n = radiazione netta (MJ m^{-2} d^{-1})
- T_k = temperatura assoluta media a 2 m dal suolo (°K)
- U_2 = velocità del vento a 2 m dal suolo (m s^{-1})
- G = flusso di calore dal suolo (MJ m^{-2} d^{-1})

- e_a-e_d = deficit di pressione di vapore dell'aria (kPa)
- γ = costante psicrometrica (kPa °C^{-1})

L'evapotraspirazione potenziale può essere calcolata per diversi orizzonti temporali: orario, giornaliero e mensile.

Per il calcolo giornaliero dell'evapotraspirazione sono necessari i dati quotidiani della radiazione solare, della temperatura massima e minima, della umidità dell'aria massima e minima, e infine della velocità del vento. Tali dati meteorologici sono facilmente misurabili o reperibili da delle stazioni meteorologiche e vengono elaborati nella seguente maniera:

- La pressione di vapore saturo (e_a):

$$e_a = 0.6108\, e^{(\frac{17.27\, T}{T+237.3})}$$

  dove T è la temperatura dell'aria

- La pressione di vapore effettiva (e_d):

$$e_d = e_a \frac{RH}{100}$$

  dove RH è l'umidità relativa

- Il valore di Δ:

$$\Delta = \frac{4098\, [0.6108 \exp(\frac{17.27\, T}{T + 237.3})]}{(T + 237.3)^2}$$

  Dove T è la temperatura media giornaliera

- Il valore di R_n:

$$Rn = Rns - Rnl$$

  dove:

  - ✓ R_{ns} è la radiazione netta a onda corta
  - ✓ R_{nl} è la radiazione netta a onda lunga

- La radiazione netta può essere stimata con la formula:

$$Rns = (1 - \alpha)Rs$$

  dove:

- ✓ α è il coefficiente di riflessione della superficie che per una coltura erbacea di riferimento è pari a 0.23
- ✓ R_s è la radiazione solare

- La radiazione R_{nl}:

$$Rnl = \left(0{,}1 + 0{,}9\frac{n}{N}\right)(0.34 - 0.14\sqrt{e_d})(\sigma\, {T_k}^4)$$

dove:

- ✓ n/N è l'eliofania relativa
- ✓ σ è pari a $4{,}903\text{x}10^{-9}$
- ✓ e_d è la tensione di vapore effettiva
- ✓ T_k è la temperatura assoluta

- Il flusso G per l'orizzonte temporale è zero. Per periodi più lunghi:

$$G = 0{,}14(T_i + T_{i-1})$$

dove:

- ✓ T_i è la temperatura media nel periodo considerato
- ✓ T_{i+1} è la temperatura media nel periodo precedente

- Il valore della costante psicrometrica γ:

$$\gamma = 0{,}00163\,\frac{P}{\lambda}$$

dove:

- ✓ P = pressione atmosferica:

$$P = 101.3\,(\frac{293 - 0.0065\,z}{293})^{5.26}$$

dove z è la quota (m) sul livello del mare

- ✓ λ = calore latente di vaporizzazione:

$$\lambda = 2{,}501 - (2{,}361\text{x}10^{-3})T$$

dove T è la temperatura media.

- La velocità del vento (m s^{-1}) va misurata a 2 metri di altezza. Per altezze differenti il valore può essere convertito:

$$U_2 = U_z \frac{4.87}{\ln(67.8z - 5.42)}$$

Attraverso l'equazione di Penman-Monteith si determina l'evapotraspirazione di riferimento o potenziale (ETo).
Successivamente bisogna utilizzare un coefficiente di correzione dell'ETo definito come coefficiente colturale (Kc), specifico per ogni coltura e diverso per ogni suo stadio vegetativo. [16]
Pertanto ETc, ovvero la domanda evapotraspirativa dell'ambiente può essere determinata nella seguente maniera:

$$ETc = ETo \ Kc$$

5. Parametri meteorologici

Nel precedente paragrafo è stato raccomandato dalla FAO il metodo di Penman-Monteith per poter determinare l'evapotraspirazione potenziale in funzione di quattro parametri meteorologici, ovvero:

- Temperatura dell'aria: grandezza fisica che caratterizza l'atmosfera dei pianeti gassosi
- Umidità relativa: quantità d'acqua o di vapore acqueo contenuta nell'atmosfera
- Velocità del vento: movimento di una massa d'aria atmosferica da un'area con alta pressione a un'area con bassa pressione
- Radiazione solare: energia radiante emessa dal Sole

La temperatura dell'aria, l'umidità relativa, la velocità del vento e la radiazione solare sono comunemente misurati dalle stazioni agro-meteorologiche e

pertanto reperibili.

Per poter determinare l'evapotraspirazione per un orizzonte temporale futuro bisogna di conseguenza ottenere i dati della temperatura dell'aria, dell'umidità relativa, della velocità del vento e della radiazione solare per il medesimo orizzonte temporale futuro per cui si vuole determinare l'evapotraspirazione, pertanto è necessario predire i valori futuri dei parametri meteorologici. Ci sono essenzialmente due modi per affrontare il problema della predizione delle variabili meteorologiche:

- Modelli di previsioni meteorologiche numerici: sono affidabili, ma anche abbastanza complessi e richiedono informazioni in tempo reale, solitamente disponibili solo dalle agenzie meteorologiche. Inoltre, sono necessari computer molto potenti per poter risolvere le equazioni differenziali coinvolte.
- Approccio statistico di modellazione: basati sull'uso di dati storici raccolti. Questi metodi, confrontati ai precedenti, richiedono meno sforzi computazionali, ma sono appropriati solo per orizzonti temporali brevi.

La radiazione solare e la velocità del vento sono due delle variabili meteorologiche più imprevedibili e variabili rispetto alle altre in quanto sono governate da fenomeni distribuiti e difficili da prevedere, pertanto sono anche le più complesse da predire.

Infatti, la radiazione solare è fortemente influenzata dalle caratteristiche dell'ombreggiatura derivata dalle nuvole (dimensioni, velocità e numero) e altre variabili, compresa la trasmittanza atmosferica, la torbidità del cielo e il livello di inquinamento.

Allo stesso modo, la velocità del vento dipende dalla differenze di pressione che si verificano in varie aree, ma è fortemente influenzata anche da fenomeni abbastanza complessi che si verificano nell'atmosfera. [17]

Capitolo 3: Data Analysis

1. Internet of things

L'Internet of Things è composto da dispositivi "intelligenti" che utilizzano la tecnologia wireless per parlare tra loro e con gli utenti.
Questi dispositivi collegati offrono esperienze più intelligenti ed efficienti per gli utenti, influenzando il business, la produzione, l'assistenza sanitaria, la vendita al dettaglio, la sicurezza e i trasporti.
I prodotti IoT possono essere trovati in casa, in ufficio, nelle industrie e nelle nostre città, con numerose applicazioni.
Oltre ad una connessione che consente loro di comunicare, questi dispositivi di solito hanno sensori digitali per raccogliere dati rilevanti e un processore per elaborare i dati.
A partire dal 2015, l'IoT comprendeva 15 miliardi di dispositivi ed è destinato a crescere fino a 30+ miliardi entro il 2020, equivalente a 3 oggetti intelligenti per ogni essere umano sulla Terra.
I dati raccolti sono usati per una varietà di scopi che possono essere noti o sconosciuti all'utente. Per esempio, un orologio indossabile per il fitness può fornire dati sulla salute all'utente in tempo reale, mentre il fornitore del servizio contemporaneamente raccoglie tutti i dati dei suoi utenti, li aggrega, e li monetizza. La proprietà dei dati e la sicurezza informatica sono due dei temi principali al centro del dibattito sull'IoT.
Un dispositivo IoT deve essere progettato in modo che possa sopravvivere all'ambiente in cui è collocato e in luoghi dove gli utenti potrebbero interagire con loro oppure non vederli.
L'obiettivo degli oggetti connessi è, in generale, quello di semplificarci la vita automatizzando processi o mettendoci a disposizione informazioni che prima

non avevamo. [18]

2. Big Data e Cloud Computing

2.1 Big Data

I *Big Data* e la loro analisi sono al centro della scienza moderna e degli affari. Questi dati sono generati online da transazioni, e-mail, video, audio, immagini, accessi, pubblicazioni, richieste di ricerca, cartelle cliniche, social network, interazioni, dati scientifici, sensori e telefoni cellulari.

Sono archiviati in database in aumento ed è difficile catturarli, gestirli, condividerli, analizzarli e visualizzarli attraverso l'uso di tipici strumenti software per database.

5 Exabyte (1018 byte) di dati sono stati generati fino al 2003. Oggi questa quantità di informazioni viene creata in due giorni.

Nel 2012, il mondo digitale dei dati è stato esteso a 2,72 Zettabyte (1021 byte). Si prevede di raddoppiare ogni due anni, raggiungendo circa 8 Zettabyte di dati entro il 2015. [20]

Gli autori tentano di definire chiaramente i Big Data presentando diverse definizioni, evidenziando che i Big Data sono prevalentemente associati all'archiviazione e all'analisi dei dati, termini che risalgono a tempi lontani. Gli autori sostengono anche che l'aggettivo grande implica significato, ma è difficile definire quantitativamente i Big Data.

Delle numerose e diverse definizioni, alcune definiscono le caratteristiche dei Big Data, altre si basano sull'aumento dei dati tradizionali con più fonti di dati non strutturati, e altri cercano di quantificare i Big Data.

Per concludere circa la somiglianza tra le definizioni, gli autori affermano che tutte le definizioni includono almeno uno dei seguenti aspetti: dimensione,

complessità, tecniche e tecnologie per elaborare grandi insiemi di dati complessi.
Software e hardware tradizionali non sono in grado di riconoscere, raccogliere, gestire o elaborare questo nuovo tipo di dati in tempi ragionevoli, pertanto il concetto di Big Data include l'archiviazione e l'analisi di grandi insiemi di dati complessi usando una serie di nuove tecniche.
I Big Data possono essere classificati come:

- dati strutturati (ad esempio, fogli di calcolo, database relazionali)
- semi-strutturati (ad es. accessi del server web e Extensible Markup Language - XML)
- non strutturati (ad es. post sui social media, audio, video, immagini).

Le tecnologie tradizionali possono presentare difficoltà significative per memorizzare ed elaborare i Big Data.

La maggior parte di questi dati non si adattano bene ai database tradizionali e ci deve essere un cambiamento di paradigma nel modo in cui le organizzazioni eseguono l'analisi per sistemare dati strutturati, semi-strutturati e non strutturati.
Un'altra caratteristica fondamentale è la velocità, riferendosi alla velocità con cui i dati sono generati o alla velocità di analisi e supporto decisionale.
I dati possono essere generati a velocità diverse.
Nel tempo sono emerse due caratteristiche aggiuntive: il valore e la veridicità.
Il valore rappresenta i risultati attesi dall'elaborare e dall'analisi dei Big Data, che di solito ha valore basso nel suo stato grezzo.
Dall'altra parte, la veridicità attira l'attenzione su possibili imprecisioni dei dati, poiché a volte l'analisi si basa su insiemi di dati con diversi gradi di precisione,

autenticità e affidabilità. [21]

2.2 Cloud Computing

Con il rapido sviluppo delle tecnologie di elaborazione e di archiviazione, e il successo di Internet, le risorse informatiche sono diventati più economici, più potenti e più onnipresenti.

Questa tendenza tecnologica ha permesso la realizzazione di un nuovo modello di calcolo chiamato cloud computing, in cui le risorse (ad es. CPU e la memoria) sono forniti come utilità generali che possono essere affittate e rilasciate dagli utenti attraverso Internet su richiesta.

In un ambiente di cloud computing, il tradizionale ruolo del fornitore di servizi è diviso in due: l'infrastruttura dei fornitori che gestiscono le piattaforme cloud, e il servizio offerto, ovvero noleggiare risorse da una o più infrastrutture. L'emergenza del cloud computing ha avuto un impatto enorme sull'industria dell'Information Technology (IT), dove grandi aziende come Google, Amazon e Microsoft si sforzano di fornire piattaforme cloud più potenti, affidabili ed a costi contenuti.

I servizi di cloud computing si possono suddividere in tre categorie principali:

- SaaS (Software as a Service): indica un software ospitato da un fornitore di terze parti al quale accedere via web (normalmente tramite un semplice log-in). Questo servizio viene generalmente pagato con un abbonamento per singolo utente (o singola postazione) e ciò differisce dal vecchio modello di acquisto e installazione manuale del software su una macchina o un server.
- IaaS (Infrastructure as a Service): è un'offerta con la quale un vendor di terze parti fornisce un'infrastruttura IT altamente automatizzata e scalabile.

• PaaS (Platform as Service): fornire una piattaforma con gli strumenti e le funzionalità necessarie per sviluppare e distribuire le applicazioni in modo sicuro. [22] [23]

3. Artificial intelligence e Machine Learning

3.1 Artificial Intelligence

L'obiettivo dell'intelligenza artificiale è sviluppare l'intelligenza umana nelle macchina. Tuttavia, un tale sogno può essere realizzato attraverso algoritmi di apprendimento che cercano di imitare come un cervello umano apprende. L'intelligenza artificiale (AI) porta con sé una promessa di genuina interazione uomo-macchina. Quando le macchine diventano intelligenti, possono comprendere richieste, collegare dati e trarre conclusioni. Possono ragionare, osservare e pianificare.
Nel complesso, l'intelligenza artificiale contiene molti sottocampi, tra cui:

- Machine Learning: automatizza la costruzione del modello analitico. Usa metodi delle reti neurali, della statistica, delle ricerche operative e della fisica per trovare informazioni nascoste nei dati senza essere programmato esplicitamente su cosa guardare o su cosa concludere.
- Reti neurali: è un tipo di apprendimento automatico ispirato al funzionamento del cervello umano. È un sistema di calcolo costituito da unità interconnesse (come i neuroni) che elabora le informazioni rispondendo a input esterni, trasmettendo informazioni tra ogni unità.
- Deep Learning: utilizza enormi reti neurali con molti strati di unità di elaborazione, sfruttando i progressi della potenza di calcolo e tecniche di allenamento migliorate per apprendere modelli complessi con grandi

quantità di dati. Le applicazioni comuni includono il riconoscimento dell'immagine e del parlato.

- Computer Vision: si basa sul riconoscimento di pattern e sull'apprendimento profondo per riconoscere cosa c'è in un'immagine o in un video. Quando le macchine possono elaborare, analizzare e comprendere le immagini, possono acquisire immagini o video in tempo reale e interpretare l'ambiente circostante.
- Natural Language Processing: è la capacità dei computer di analizzare, comprendere e generare il linguaggio umano, compreso il parlato. [24]

3.2 Machine Learning

Il Machine learning, che è un sottocampo cresciuto nel campo dell'intelligenza artificiale, è della massima importanza in quanto consente alle macchine di acquisire l'intelligenza umana senza una programmazione esplicita. Nell'area della ricerca sull'apprendimento automatico l'enfasi è dato di più sulla scelta o lo sviluppo di un algoritmo e sul condurre esperimenti sulla base dell'algoritmo.

Gli algoritmi di Machine Learning possono essere suddivisi in 2 tipologie, ovvero:

- Supervised Learning: questo processo mira ad apprendere una funzione associando ad ogni ingresso un'uscita basandosi su delle coppie input-output di esempio e migliora le proprie capacità confrontando l'output calcolato e l'output atteso.
- Unsupervised Learning: viene fornita una serie di input che saranno classificati ed organizzati in diversi cluster sulla base di caratteristiche comuni per cercare di effettuare ragionamenti e previsioni sugli input successivi. [25]

3.3 Deep Learning

Il Machine Learning alimenta molti aspetti della società moderna: dalle ricerche sul web e filtraggio dei contenuti sui social network alle raccomandazioni sui siti di e-commerce, ed è sempre più presente nei prodotti di consumo come fotocamere e smartphone.

I sistemi di Machine Learning sono utilizzati per identificare gli oggetti nelle immagini, trascrivere un discorso in testo, abbinare notizie, post o prodotti con interessi degli utenti e selezionare i risultati pertinenti alla ricerca. Sempre più spesso queste applicazioni fanno uso di una classe di tecniche chiamate Deep Learning.

I metodi di Deep Learning sono metodi di rappresentazione-apprendimento con più livelli di rappresentazione, ottenuti componendo moduli semplici ma non lineari, ciascuno di essi trasforma la rappresentazione in una rappresentazione di un livello più alto, leggermente più astratto. L'aspetto chiave del Deep Learning è che gli strati di funzionalità non sono progettati da ingegneri umani: loro vengono appresi dai dati utilizzando una procedura di apprendimento generale.

Il Deep Learning sta facendo importanti progressi nella soluzione di problemi che hanno resistito ai migliori tentativi da parte della comunità dell'intelligenza artificiale per molti anni.

Il Deep Learning si è dimostrato molto valido nello scoprire strutture complesse in molti domini di scienza, affari e governo.

Più sorprendentemente, il Deep Learning ha prodotto risultati estremamente promettenti per vari compiti nella comprensione del linguaggio naturale, in particolare nella classificazione degli argomenti, nell'analisi dei sentimenti, nel fornire delle risposte a delle domande.

Il Deep Learning avrà molti più successi nel prossimo futuro perché richiede pochissimi interventi ingegneristici da parte dell'uomo, quindi può facilmente

trarre vantaggio dagli aumenti della quantità di calcolo e dati disponibili. [38] Le principali caratteristiche del Deep Learning sono:

- vari livelli di unità non lineari a cascata per svolgere compiti di estrazione di caratteristiche e di trasformazione. Ciascun livello successivo utilizza l'uscita del livello precedente come input. Gli algoritmi possono essere sia di tipo supervisionato sia non supervisionato e le applicazioni includono l'analisi di pattern (apprendimento non supervisionato) e classificazione (apprendimento supervisionato).
- sono basati sull'apprendimento non supervisionato di livelli gerarchici multipli di caratteristiche (e di rappresentazioni) dei dati. Le caratteristiche di più alto livello vengono derivate da quelle di livello più basso per creare una rappresentazione gerarchica.
- fanno parte della più ampia classe di algoritmi di apprendimento della rappresentazione dei dati all'interno dell'apprendimento automatico.
- apprendono multipli livelli di rappresentazione che corrispondono a differenti livelli di astrazione. [39] [40]

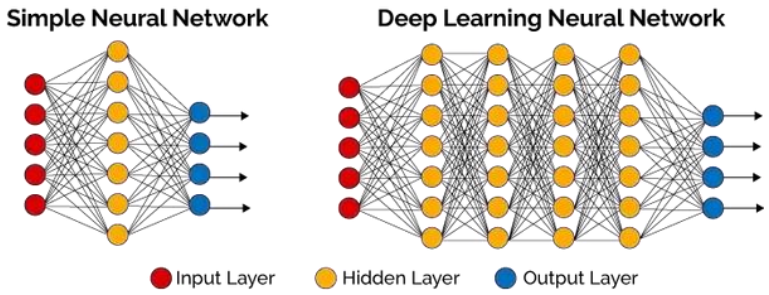


Figura 5: Confronto tra una rete neurale semplice e una rete neurale di Deep Learning. [41]

Capitolo 4: Machine Learning e Feature Selection per la predizione della radiazione solare

1. Stato dell'arte

1.1 Modelli di predizione

Per poter stimare l'evapotraspirazione, per un orizzonte temporale futuro, è necessario predire i valori futuri dei parametri meteorologici, fondamentali per calcolare l'evapotraspirazione attraverso l'equazione di Penman-Monteith, e tali parametri sono: la temperatura, l'umidità, la velocità del vento e la radiazione solare.

Lo studio si concentrerà sulla predizione della radiazione solare attraverso algoritmi di Machine Learning.

Per determinare quali dei tanti modelli di predizione fossero i più adatti, per predire una serie temporale della radiazione solare, è stata effettuata una ricerca bibliografica su tale argomento, riportata nella seguente tabella (tabella 2):

PREDIZIONE DELLA RADIAZIONE SOLARE				
ALGORITMO	ORIZZONTE	PARAMETRI METEOROLOGI	ACCURATEZZA DELLE PREDIZIONI	FONTE
MLP	giornaliero	temperatura media, umidità, ore di sole, giorno dell'anno, radiazione solare	RMSE: 0.0441	[37]
		temperatura massima, temperatura minima, radiazione solare	RMSE: 2.53	[27]
		temperatura media, temperatura massima, temperatura minima, differenza di temperatura, ore di sole, radiazione solare	RMSE: 2.925	[33]
		temperatura media, ore di sole, umidità, giorno dell'anno radiazione solare	RMSE: 0.044	[29]
SVM	orario	giorno dell'anno, anno, orario, radiazione solare	RMSE: 0.11	[31]

	giornaliero	ore di sole, temperatura minima, temperatura massima, radiazione solare	RMSE: 2.773	[30]
		temperatura minima, temperatura massima, radiazione solare, precipitazioni	RMSE: 2.142	[35]
		temperatura media, temperatura massima, temperatura minima, differenza di temperatura, ore di sole, radiazione solare	RMSE: 2.777	[33]
	mensile	temperatura massima, temperatura minima, ore di sole, radiazione solare	RMSE: 1.90	[28]
		temperatura minima, temperatura massima, differenza di temperatura, radiazione solare	RMSE: 0.87	[36]
		temperatura massima, temperatura minima, umidità, radiazione solare	RMSE: 0.619	[32]
LSTM	orario	giorno dell'anno, anno, orario, radiazione solare	RMSE: 0.086	[31]
ARMA	giornaliero	ore di sole, temperatura media, differenza di temperatura, pressione del vapore acqueo, umidità, radiazione solare	RMSE: 2.797	[34]
	mensile	ore di sole, temperatura media, differenza di temperatura, pressione del vapore acqueo, umidità, radiazione solare	RMSE: 1.99	[34]
W-SVM	giornaliero	ore di sole, temperatura minima, temperatura massima, radiazione solare	RMSE: 2.719	[30]
		ore di sole, temperatura media, differenza di temperatura, pressione del vapore acqueo, umidità, radiazione solare	RMSE: 2.647	[34]
	mensile	ore di sole, temperatura media, differenza di temperatura, pressione del vapore acqueo, umidità, radiazione solare	RMSE: 1.82	[34]
GP	giornaliero	ore di sole, temperatura media, differenza di temperatura, pressione del vapore acqueo, umidità, radiazione solare	RMSE: 2.781	[34]
	mensile	ore di sole, temperatura media, differenza di temperatura, pressione del vapore acqueo, umidità, radiazione solare	RMSE: 1.94	[34]
RBF	giornaliero	temperatura media, ore di sole, umidità, giorno dell'anno radiazione solare	RMSE: 0.013	[29]
ANFIS	giornaliero	temperatura minima, temperatura massima, radiazione solare, precipitazioni	RMSE: 2.423	[35]
	mensile	temperatura massima, temperatura minima, ore di sole, radiazione solare	RMSE: 2.06	[28]

Tabella 2: Stato dell'arte sulla predizione della radiazione solare.

Sono stati sviluppati diversi modelli per stimare la radiazione solare, basandosi su una serie di parametri meteorologici e geografici comunemente disponibili, compresa le ore del sole, la temperatura dell'aria, l'umidità relativa, la pressione, il livello sopra al mare, l'altitudine, la latitudine, la longitudine e tanti altri parametri meteorologici.

Tuttavia, trovare una relazione appropriata tra la radiazione solare e una o più parametri, da fornire in input ai modelli, è stata una sfida seria nel campo della predizione della radiazione solare.

Sebbene un vasto numero di modelli sia stato proposto per stimare la radiazione solare, sviluppando algoritmi appropriati e approcci per ottenere maggiore affidabilità, accuratezza ed efficienza nel processo computazionale è ancora tutto impegnativo.

Negli ultimi anni, varie intelligenze artificiali e tecniche di intelligenza computazionale sono state applicate dai ricercatori per il problema della predizione della radiazione solare in molte località di tutto il mondo. Ali Rahimikhoob [27] ha effettuato uno studio per predire la radiazione solare giornaliera utilizzando una rete neurale artificiale, in particolare il modello *Multi-Layer Perceptron.* I suoi risultati dimostrano che questa rete neurale artificiale può essere considerata un effettiva tecnica per stimare la radiazione solare.

Lanre et al. [28] hanno utilizzato la temperatura minima, la temperatura massima e le ore di sole come parametri da fornire in input al modello per predire la radiazione solare mensile usando i modelli *Support Vector Machines* (SVM) e *Adaptive Network-Based Fuzzy Inference System* (ANFIS). I risultati ottenuti mostrano che il modello *Support Vector Machine* è favorevolmente capace di stimare la radiazione solare su questi parametri. Mohamed Benghanem e Adel Mellit [29] hanno sviluppato i modelli *Radial Basis Function* e *Multi-Layer Percepton* per stimare la radiazione solare in

Arabia Saudita. Loro utilizzano diverse combinazioni di parametri in input come ore di sole, temperatura dell'aria, umidità relativa e giorno dell'anno. I loro risultati mostrano che un alta accuratezza è dipesa dalle ore di sole e dalla temperatura.

Ravinesh et al. [30] ha applicato il modello *Support Vector Machines* per stimare la radiazione solare fornendo come parametri in input le ore di sole, la temperatura minima e massima, la velocità del vento, l'evapotraspirazione e le precipitazioni. Esso ha trovato che stimando la radiazione solare usando il modello *Support Vector Machine* si ha un alta accuratezza. Ahmad Alzahrani et al. [31] hanno utilizzato due modelli: il *Long Short Term Memory* e il *Support Vector Machine* per predire la radiazione solare oraria usando come parametri il giorno dell'anno, l'anno e l'orario. Mennal e Selvakumar [32] per stimare la radiazione solare mensile in India hanno applicato il modello *Support Vector machine* con diverse combinazioni tra i seguenti parametri in input: mese, latitudine, longitudine, ore di sole, giorno dell'anno, umidità relativa, temperatura massima e minima, ottenendo un accuratezza ottima.

Belaid e Mellit [33] hanno predetto la radiazione solare giornaliera e mensile applicando una rete neurale artificiale, ovvero il *Multi-Layer Percepton* e il *Support Vector Machines* utilizzando la temperatura e le ore di sole come parametri meteorologi da fornire ai modelli.

Kasra Mohammadi et al. [34] hanno sviluppato un SVM ibrido, combinando SVM con la trasformata di Walet (W-SVM) per predire la radiazione solare giornaliera e mensile, e hanno confrontato i risultati ottenuti con i risultati determinato applicando *Genetic Programming* (GP) e *AutoRegressive Moving Average* (ARMA).

Victor Quej et al. [35] ha predetto la radiazione solare giornaliera nella penisola Tycatan in Messico attraverso i modelli*: ANFIS, SVM e MLP* usando

diversi parametri meteorologici e diverse combinazioni di essi. I risultati indicano che sia il modello *SVM* che il modello *MLP* forniscono predizioni accurate della radiazione solare.

Ji-Long Chen et al. [36] hanno predetto la radiazione solare mensile utilizzando solamente la temperatura massima e minima come parametri di input, logicamente provando le loro possibili combinazioni e differenze, utilizzando il modello *SVM*.

Benghamen, Mellit e Alamri [37] hanno sviluppato sei reti neurali artificiali per predire la radiazione solare giornaliere usando diverse combinazioni di input tra la temperatura dell'aria, l'umidità relativa, le ore di sole e il giorno dell'anno.

Di seguito sono stati sintetizzati in forma tabella i documenti scientifici individuati dalla ricerca bibliografica, sopra discussa, che utilizzavano i modelli: *Multi-Layer Percepton*, *Support Vector Machines* e *Long Short Term Memory.*

Lo studio si incentra sui modelli: *Multi-Layer Percepton*, *Support Vector Machines* e *Long Short Term Memory*, in quanto si evince dalla ricerca bibliografica condotta che sono maggiormente utilizzati e più accurati degli altri per predire la radiazione solare, e inoltre dispongono di librerie software per l'implementazione preesistenti.

Nella tabella sono stati riportati l'orizzonte temporale della predizione, i parametri meteorologici utilizzati in input ai modelli, l'accuratezza migliore ottenuti e la tecnica utilizzata per suddividere i dati in training set (insieme di dati che vengono utilizzati per addestrare un sistema supervisionato) e testing set (insieme di esempi utilizzati per valutare le prestazioni di un sistema).

PREDIZIONE DELLA RADIAZIONE SOLARE				
MODELLI: MLP, SVM, LSTM				
ORIZZONTE	INPUT	DATA SET	ACCURATEZZA	FONTE
orario	giorno dell'anno, anno, orario, radiazione solare	Training set: ~80% Testing set: ~20%	RMSE: 0.086	[5]
	giorno dell'anno, anno, orario, radiazione solare	Training set: ~80% Testing set: ~20%	RMSE: 0.11	[5]
giornaliero	temperatura media, umidità, ore di sole, giorno dell'anno, radiazione solare	Training set: 80% Testing set: 20%	RMSE: 0.044	[11]
	temperatura massima, temperatura minima, radiazione solare	Training set: 80% Testing set: 20%	RMSE: 2.53	[1]
	temperatura media, temperatura massima, temperatura minima, ore di sole, radiazione solare	Training set: ~70% Testing set: ~30%	RMSE: 2.925	[7]
	temperatura media, ore di sole, umidità, giorno dell'anno, radiazione solare	Training set: 80% Testing set: 20%	RMSE: 0.044	[3]
	ore di sole, temperatura massima, radiazione solare	Training set: ~70% Testing set: ~30%	RMSE: 2.773	[4]
	temperatura minima, temperatura massima, radiazione solare, precipitazioni	Training set: 70% Testing set: 30%	RMSE: 2.142	[9]

	temperatura media, temperatura minima, differenza di temperatura, ore di sole, radiazione solare	Training set: ~70% Testing set: ~30%	RMSE: 2.777	[7]
mensile	temperatura massima, temperatura minima, ore di sole, radiazione solare	Training set: 70% Testing set: 30%	RMSE: 1.90	[2]
	temperatura minima, temperatura massima, radiazione solare	Training set: ~70% Testing set: ~30%	RMSE: 0.87	[10]
	temperatura massima, temperatura minima, umidità, radiazione solare	Training set: ~70% Testing set: ~30%	RMSE: 0.619	[6]

Tabella 3: Parametri meteorologici, metodologia per suddividere il data set in training set e testing set, e accuratezza dei risultati per la predizione della radiazione solare.

Osservando la tabella 3, è evidente che nei documenti scientifici individuati attraverso la ricerca bibliografica, le varie predizioni della radiazione solare sono state effettuate suddividendo il data set attraverso due principali metodologie, ovvero:

- 70% training set e 30% testing set (60% delle fonti utilizzano tale metodologia).
- 80% training set e 20% testing set (40% delle fonti utilizzano tale metodologia).

1.2 Parametri in input per i modelli

Dai documenti scientifici individuati, mediante la ricerca bibliografica condotta, è possibile estrapolare informazioni riguardanti i parametri meteorologici e climatici utilizzati, e forniti in input ai modelli di predizione. È possibile individuare dalle prove effettuate nella letteratura, sopra riportata,

quali sono stati i parametri maggiormente utilizzati, e che hanno portato ad avere predizioni più accurate della radiazione solare, per differenti orizzonti temporali (orario, giornaliero, mensile).

Di seguito, raggruppiamo le informazioni riguardanti i parametri, considerando la loro frequenza d'uso nelle prove condotte e per ognuno di essi qual è stata l'accuratezza migliore e peggiore.

PREDIZIONE DELLA RADIAZIONE SOLARE				
MODELLI: MLP, SVM, LSTM				
ORIZZONTE	**PARAMETRI**	**FREQUENZA D'USO**	**RMSE ↓**	**RMSE ↑**
orario	radiazione solare	2 su 2	0.086	0.11
	giorno dell'anno	2 su 2	0.086	0.11
	anno	2 su 2	0.086	0.11
	orario	2 su 2	0.086	0.11
giornaliero	radiazione solare	7 su 7	0.044	2.925
	temperatura media	4 su 7	0.044	2.925
	temperatura massima	4 su 7	2.142	2.925
	temperatura minima	4 su 7	2.142	2.925
	ore di sole	4 su 7	0.044	2.925
	giorno dell'anno	2 su 7	0.044	0.044
	umidità	2 su 7	0.044	0.044
	precipitazioni	1 su 7	2.142	2.142
	differenza di temperatura	1 su 7	2.777	2.777
mensile	Radiazione solare	3 su 3	0.619	1.90
	Temperatura massima	3 su 3	0.619	1.90
	Temperatura minima	3 su 3	0.619	1.90
	Ore di sole	1 su 3	1.90	1.90
	umidità	1 su 3	0.619	0.619
Totale	radiazione solare	12 su 12	0.044	2.925
	temperatura massima	7 su 12	0.619	2.925
	temperatura minima	7 su 12	0.619	2.925
	ore di sole	5 su 12	0.044	2.925
	temperatura media	4 su 12	0.044	2.925
	giorno dell'anno	4 su 12	0.044	0.619
	umidità	3 su 12	0.044	0.044
	orario	2 su 12	0.086	0.11
	anno	2 su 12	0.086	0.11
	precipitazioni	1 su 12	2.142	2.142
	differenza di temperatura	1 su 12	2.777	2.777

Tabella 4: Variabili meteorologiche e climatiche usate, dalle fonti scientifiche, per predire la radiazione solare.

Analizzando la tabella 4 si evince che, i parametri meteorologici maggiormente utilizzati sono la temperatura massima, la temperatura minima, la temperatura media, le ore di sole e il giorno dell'anno. Invece, i parametri

che hanno permessi di ottenere predizioni delle radiazioni solari accurate sono stati: la temperatura minima, la temperatura massima, la temperatura media, le ore di sole e il giorno dell'anno, l'umidità, e infine l'orario e l'anno.

Pertanto, nello studio condotto saranno utilizzati i parametri indicati dai documenti scientifici individuati, ovvero:

- Temperatura massima
- Temperatura minima
- Temperatura media
- Differenza di temperatura
- Ore di sole
- Giorno dell'anno
- Umidità
- Precipitazioni
- Orario
- Anno

Inoltre, sono stati aggiunti due parametri meteorologici non utilizzati nelle prove condotte dai documenti scientifici della ricerca bibliografica, ovvero la pressione e la velocità del vento, per determinare se in qualche modo fossero correlati con la radiazione solare, e quindi portassero ad avere predizioni di essa maggiormente accurate.

2. Algoritmi di Machine Learning utilizzati

2.1 Reti neurali artificiali

Le reti neurali artificiali sono tra i modelli più accurati di predizione, ampliamente utilizzati in numerosi e vari domini (sociale, economico,

ingegneristico, ecc.).
Diverse caratteristiche rendono le reti neurali artificiali preziose e interessanti per compiti di predizione.
In primo luogo, a differenza dei modelli tradizionali, i modelli delle reti neurali artificiali sono auto-adattativi e guidati dai dati.
In secondo luogo, le reti neurali artificiali possono generalizzare. Dopo aver appreso i dati presentati (un campione), le reti neurali artificiali possono spesso inferire correttamente anche la parte invisibile di una popolazione se i dati di esempio contengono informazioni rumorose.
In terzo luogo, le reti neurali artificiali sono approssimatori di funzioni universali. È stato dimostrato che una rete può approssimare qualsiasi funzione continua a qualsiasi precisione desiderata.
Dati i vantaggi delle reti neurali artificiali, non è sorprendente che questa metodologia ha attirato l'attenzione in modo schiacciante nella predizione delle serie temporali. [42]
Esistono diversi tipi di reti neurali artificiali. Questi tipi reti sono implementate sulla base delle operazioni matematiche e di un set di parametri richiesti per determinare l'output.
Due dei vari tipi di reti neurali artificiali sono:

- Rete neurale feed forward
- Rete neurale ricorrente

La rete neurale feed forward è una delle forme più semplici di RNA, dove i dati o l'input viaggiano in una direzione. I dati passano attraverso i nodi di input ed escono dai nodi di output. Questa rete neurale può avere o meno i livelli nascosti.
Di seguito è riportata una rete di feed forward a livello singolo. Qui, la somma dei prodotti di input e dei pesi viene calcolata e alimentata dall'output. L'output è considerato se è superiore ad un certo valore, cioè una soglia (di solito 0) e

il neurone si attiva con un'uscita attivata (solitamente 1) e se non si attiva, viene emesso il valore disattivato (di solito -1). [43] [44]
Invece, la rete neurale ricorrente si basa sul principio di salvare l'output di un livello e reinserirlo nell'input.
Qui, il primo strato è formato in modo simile alla rete neurale feed forward con il prodotto della somma dei pesi e delle caratteristiche. Il processo della rete neurale ricorrente inizia una volta che questo è calcolato, ciò significa che da un passo temporale a quello successivo ogni neurone ricorderà alcune informazioni che aveva nel passo temporale precedente. Questo fa sì che ciascun neurone si comporti come una cella di memoria nell'esecuzione dei calcoli. [43]

2.1.1 Multi-Layer Percepton

Il *Multi-Layer Perceptron* è forse l'architettura di rete più popolare in uso oggi sia per la classificazione che per la regressione. Se tale rete è formata da un certo numero d di input:

$$y_j = y\left(\sum_{i=0}^{d} w_{ji} x_i\right)$$

dove x_i è il vettore degli input, w_{ji} sono i pesi di ogni input combinati con ogni output e y è l'output della rete.
Funzioni di attivazione sono utilizzate per questa architettura come funzione soglia, però essendo abbastanza limitante, viene sostituita la funzione di attivazione con la funzione logistica. Pertanto, si possono applicare le regole del calcolo differenziale essendo la funzione logistica differenziabile. [45]
È possibile aumentare i livelli e il numero di neuroni per ogni livello. Le unità dei livelli aggiuntivi sono dette unità nascoste e i livelli sono detti livelli

nascosti, in quanto non sono accessibili dall'esterno.
Considerando una rete di due livelli di unità di elaborazione con d ingressi e m uscite per il primo strato e c uscite per il secondo, gli output finali della rete sono:

$$z_k = z\left(\sum_{j=0}^{m} w'_{kj} y_j\right)$$

z_k è l'output, w'_{kj} sono i pesi per ogni unità di elaborazione e y_j il segnale inviato dalle unità nascoste.
Una rete a più livelli ha una capacità di approssimazione molto più ampia di quella a un livello solo, infatti con un numero sufficiente di unità nascoste possiamo rappresentare una qualsiasi funzione continua sul dominio delle variabili d'ingresso. [49]

2.1.2 Long Short-Term Memory

Il modello *Long Short-Termo Memory* (LSTM) è un'unità di una rete neurale ricorrente (RNN). Una rete neurale ricorrente composta da unità LSTM è spesso chiamata rete LSTM.
Le reti basate su LSTM sono ideali per la predizione e classificazione di sequenze temporali, e stanno soppiantando molti approcci classici di machine learning.
Un'unità LSTM comune è composta da:

- una cella
- una porta di input
- una porta di output
- una porta di dimenticanza

Pertanto, una rete LSTM è composta da celle concatenate tra loro. Ogni cella è composta dalle 3 tipi di porte (input, output, forget) che implementano rispettivamente le funzioni di scrittura, lettura e reset sulla memoria della cella.

Le porte non sono binarie ma analogiche e fanno uso di una funzione di attivazione che genera 1 per indicare la totale attivazione e 0 per la totale inibizione, inoltre sono di tipo moltiplicativo.

Tali porte permettono alle celle LSTM di ricordare le informazioni per un tempo indefinito: infatti, se la porta di input è sotto la soglia di attivazione, la cella manterrà lo stato precedente, mentre se è abilitato, lo stato corrente verrà combinato con il valore in ingresso.

Come suggerisce il nome, la porta di dimenticanza (forget) resetta lo stato corrente della cella (quando il suo valore viene portato a 0), e la porta di output decide se il valore all'interno della cella dev'essere portato in uscita o meno. [46] [50]

2.2 Support Vector Machines

L'algoritmo *SVM*, sviluppato da Vapnik, si basa sulla teoria dell'apprendimento statistico.

Dato un insieme di punti, l'obiettivo è di costruire un iperpiano che lascia la maggior parte dei punti di una stessa classe nello stesso semipiano e contemporaneamente massimizza la distanza dei punti delle classi dall'iperpiano.

Il modello di regressione ottenuto addestrandolo su un set di parametri può essere rappresentato come un ipertubo di raggio ε.

Nel caso ideale, viene individuata una funzione di regressione tramite le SVM che mappa tutti i dati di input con una deviazione massima pari proprio a ε,

dal valore di target, cosicché tutti i punti che sono stati forniti per addestrare le SVM si trovano all'interno del tubo di regressione.
Comunque, usualmente non è possibile inserire tutti gli oggetti all'interno del tubo e avere ancora un modello che abbia un qualche significato quindi nel caso generale le SVM in modalità di regressione considerano zero l'errore per gli oggetti all'interno del tubo mentre quelli all'esterno hanno un errore che dipende dalla distanza dal margine del tubo.
Il metodo che può essere utilizzato per costruire una mappatura in uno spazio di caratteristiche di grande dimensioni è mediante l'uso di kernel. Esistono numerosi kernel per costruire una mappatura, per esempio:

- Polynomial
- Gaussian Radial Basis Function
- Exponential Radial Basis Function
- Fourer Series
- Splines

Nello studio è stato utilizzato un SVM con kernel 'Gaussian Radial Basis Function', in quanto hanno ricevuto un'attenzione significativa negli ultimi decenni. [51] [52]

4. Feature Selection

4.1 Funzionalità e tipologia

Una caratteristica è una proprietà misurabile individuale del processo osservato. Utilizzando una serie di caratteristiche, qualsiasi algoritmo di Machine Learning può eseguire la classificazione.
Negli anni passati nelle applicazioni del machine learning o del pattern recognition, il dominio delle caratteristiche è stato esteso da decine a

centinaia di variabili utilizzate in tali applicazioni.
Diverse tecniche sono state sviluppate per affrontare il problema della riduzione delle variabili irrilevanti e ridondanti che sono un onere gravoso. La selezione delle caratteristiche aiuta a comprendere i dati, riducendo i requisiti di calcolo, riducendo l'effetto della maledizione della dimensioni e il miglioramento delle prestazioni predittive.
L'obiettivo della selezione delle caratteristiche è selezionare un sottoinsieme di variabili dall'input che possa descrivere in modo efficiente i dati di input riducendo al contempo gli effetti del rumore o delle variabili irrilevanti e fornisce ancora buoni risultati di previsione.
Le variabili dipendenti non forniscono ulteriori informazioni sulle classi e quindi possono essere rimosse riducendo la quantità di dati.
Per rimuovere una caratteristica irrilevante, è necessario un criterio di selezione delle caratteristiche che possa misurare la pertinenza di ciascuna caratteristica con la classe di uscita.
Una volta selezionato un criterio di selezione delle caratteristiche, è necessario sviluppare una procedura che trovi il sottoinsieme di funzioni utili. Quindi è necessario utilizzare una procedura subottimale che possa rimuovere i dati ridondanti con calcoli trattabili.
I metodi di eliminazione della caratteristiche sono stati classificati in:

- *Filter*
- *Wrapper*

I metodi *Filter* agiscono come preelaborazione, selezionando le variabili indipendentemente dal modello. Si basano solo su caratteristiche generali come la correlazione con la variabile da prevedere.
Nei metodi *Wrapper* il criterio di selezione delle caratteristiche è la performance del predittore, cioè il predittore è incluso in un algoritmo di

ricerca che troverà un sottoinsieme attraverso cui si ottiene l'accuratezza massima delle predizioni. [47]

Nello studio sono state applicate entrambe le tipologie, ovvero sono stati utilizzati sia metodi *Filter* sia metodi *Wrapper*.
La *correlazione di Pearson* è stata utilizzata come metodo *Filter*, mentre *l'eliminazione ricorsiva delle caratteristiche* (in inglese, Recursive feature elimination, RFE) con il modello *Support Vector Machine* e il modello *Random Forest* sono stati utilizzati come metodi *Wrapper*.

4.2 Pearson Correlation

Uno dei più semplici criteri per la selezione delle caratteristiche è il coefficiente di *correlazione di Pearson*, definito come:

$$R(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$$

dove x_i è la i_{th} variabile, Y è l'output (la classe), cov() è la covarianza e var() la varianza. Il ranking di correlazione può solo rilevare le dipendenze lineari tra le variabili e il target. [47]

4.3 Recursive feature elimination

Il metodo di selezione *RFE* è un processo ricorsivo che classifica le caratteristiche secondo una certa misura della loro importanza. Ad ogni iterazione viene misurata l'importanza delle caratteristiche e viene rimossa quella con l'importanza più bassa.
Un'altra possibilità, non usata nello studio, è di far rimuovere un gruppo di caratteristiche ad ogni iterazione, al fine di velocizzare il processo.

La ricorsione è necessaria perché per alcune caratteristiche, l'importanza può cambiare sostanzialmente quando sono valutate su un diverso sottoinsieme di caratteristiche.
L'ordine (inverso) con cui vengono eliminate le caratteristiche è usato per costruire un ranking finale delle importanza delle caratteristiche.
La procedure di selezione della caratteristiche consiste nel prendere le prime N caratteristiche dal ranking. [48]

4.4 Random Forest

Le foreste casuali sono tra i metodi di apprendimento automatico più popolari grazie alla loro buona accuratezza, robustezza e facilità d'uso.
La foresta casuale è composta da un numero di alberi decisionali.
Ogni nodo negli alberi decisionali è una condizione su una singola caratteristica, progettata per dividere il set di dati in due, in modo che i valori di risposta simili finiscano nello stesso insieme.
La misura in base alla quale viene scelta la condizione ottimale si chiama impurità. Per la classificazione, è tipicamente l'impurità di Gini e per gli alberi di regressione è la varianza.
Quindi, quando si allena un albero, si può calcolare di quanto ogni caratteristica diminuisce l'impurità ponderata in un albero.
Per una foresta, la diminuzione delle impurità causata da ciascuna caratteristica può essere mediata e le caratteristiche sono classificate in base a questa misura. [53]

Capitolo 5: Tools, database e task

1. Tools

1.1 RapidMiner

Per poter predire la radiazione solare, su orizzonti temporali differenti, è necessario gestire e manipolare l'enorme quantità di dati delle rilevazioni orarie fornite dal database Sysman, applicare ed eseguire algoritmi di Machine Learning per effettuare le predizioni, e realizzare grafici per poter analizzare la correlazione tra i vari parametri meteorologici e l'accuratezza delle predizioni, pertanto sono stati considerati vari strumenti di analisi dei dati, individuati attraverso una ricerca e riportati nella tabella 5:

TOOL	DESCRIZIONE	LICENZA
Knime	La piattaforma KNIME è la soluzione leader per l'innovazione basata sui dati, che ti aiuta a scoprire il potenziale nascosto nei tuoi dati ed a estrarre nuove idee. Con oltre 1000 moduli, centinaia di esempi pronti all'uso, una gamma completa di strumenti integrati e la più ampia scelta di algoritmi avanzati disponibili, la piattaforma di analisi KNIME è la cassetta degli attrezzi perfetta per qualsiasi scienziato dei dati.	open source
R Language	Il linguaggio R è ampiamente utilizzato tra i minatori di dati per lo sviluppo di software statistici e analisi dei dati. La facilità d'uso e l'estensibilità hanno notevolmente aumentato la popolarità di R negli ultimi anni. Oltre al data mining fornisce tecniche statistiche e grafiche, tra cui la modellazione lineare e non	open source

	lineare, i test statistici classici, l'analisi delle serie temporali, la classificazione, il clustering e altro.	
Orange	Orange permette l'analisi dei dati per principianti ed esperti e offre flussi di lavoro interattivi con una grande quantità di strumenti per creare flussi di lavoro interattivi per analizzare e visualizzare i dati. Orange è ricco di visualizzazioni diverse, da grafici a dispersione, grafici a barre, alberi e reti.	open source
RapidMiner	RapidMiner opera attraverso la programmazione visiva ed è in grado di manipolare, analizzare e modellare i dati. RapidMiner rende i team di Data Science più produttivi attraverso una piattaforma open source per la preparazione dei dati, l'apprendimento automatico e la distribuzione dei modelli. La sua piattaforma unificata per la scienza dei dati accelera la costruzione di flussi di lavoro analitici completi, dalla preparazione dei dati all'apprendimento automatico fino alla validazione del modello fino all'implementazione, in un unico ambiente, migliorando drasticamente l'efficienza e riducendo il tempo necessario per i progetti di scienza dei dati.	prova gratuita per un anno
Weka	Weka, un software open source, è una raccolta di algoritmi di apprendimento automatico per l'attività di data mining. Gli algoritmi possono essere applicati direttamente a un set di dati o richiamati dal proprio codice JAVA. È anche adatto per lo sviluppo di nuovi schemi di apprendimento automatico, poiché è stato completamente implementato nel linguaggio di programmazione JAVA, oltre a supportare diverse attività di data mining standard. Weka con la sua interfaccia grafica fornisce la transizione più semplice nel mondo della Data Science. Essendo scritti in Java, quelli con esperienza Java possono chiamare la libreria anche nel loro codice.	open source

Tabella 5: I principali strumenti per l'analisi dei dati. [54]

Secondo la società Gartner, leader mondiale nella consulenza strategica, ricerca e analisi nel campo dell'Information Technology, i software per l'analisi dei dati più specializzati e importanti sono RapidMiner e Knime, mettendoli a confronto considerando varie e numerose caratteristiche.
I risultati del confronto tra Knime e RapidMiner condotto dalla società Gartner sono evidenti e chiari, ovvero RapidMiner è il software più indicato per l'analisi dei dati, in quanto ottiene un punteggio superiore rispetto a Knime per ogni caratteristica considerata, dall'accesso e manipolazione ai dati alla visualizzazione di essi, dalla predizione all'ottimizzazione, dalle prestazioni e scalabilità all'esperienza dell'utente.
Inoltre, è vero che Knime è uno strumento open source ma RapidMiner mette a disposizione la licenza 'Educational', ovvero una prova gratuita senza limitazioni per un anno, pertanto nello studio è stato utilizzato RapidMiner. [55]

Secondo Bloor Research, RapidMiner fornisce il 99% di una soluzione analitica avanzata attraverso framework basati su template che velocizzano i processi e riducono gli errori eliminando quasi la necessità di scrivere codice. RapidMiner offre procedure di data mining e machine learning che includono: caricamento e trasformazione dei dati, preelaborazione e visualizzazione dei dati, analisi predittiva e modellazione statistica, valutazione e implementazione.
RapidMiner è scritto nel linguaggio di programmazione Java. RapidMiner fornisce una GUI per progettare ed eseguire flussi di lavoro analitici. Questi flussi di lavoro sono chiamati processi in RapidMiner e sono costituiti da più operatori.
Ogni operatore esegue una singola attività all'interno del processo e l'output di ciascun operatore costituisce l'input di quello successivo. RapidMiner fornisce schemi di apprendimento, modelli e algoritmi e può essere esteso utilizzando gli script R e Python. [56]

1.2 Anaconda

Nello studio, per predire la radiazione solare attraverso vari modelli di predizioni si è fatto uso di 'Anaconda', una distribuzione open source del linguaggio di programmazione Python.

Anaconda è particolarmente popolare per l'analisi e l'elaborazione scientifica di dati su larga scala, in quanto mette a disposizione numerosi moduli e implementazioni che mirano a semplificare l'ambiente complesso del Data Science e del Machine Learning, e realizza ciò attraverso modulo come: Scikit-learn, SciPy e NumPy.

Inoltre, è stato necessario scaricare e installare la libreria Pandas per poter utilizzare Python su RapidMiner, in quanto grazie a tale libreria è possibile convertire gli Example Set (tipologia particolare degli output generati dagli operatori di RapidMiner) in input comprensibili, leggibili e manipolabili dagli script in Python. [57]

1.3 Astral

Dalla ricerca bibliografica condotta nel capitolo precedente (capitolo 4, paragrafo 1, sotto paragrafo 1.2) sono stati evidenziati numerosi parametri meteorologici utilizzati dalle fonti scientifiche da fornire in input ai modelli di predizioni per poterli addestrare.

Uno dei parametri meteorologici principalmente utilizzato e fornito per addestrare i modelli di Machine Learning erano le ore di sole che si avevano in una giornata.

Tale parametro non è presente nel database Sysman, pertanto è stato necessario determinarlo in maniera indiretta, e ciò è stato possibile ricorrendo all'ausilio del modulo 'Astral' di Python, che permette di calcolare i tempi di vari aspetti del sole e della luna in ogni luogo del pianeta.

Attraverso il modulo 'Astral' è possibile determinare l'orario dell'alba e del tramonto, e successivamente effettuare la differenza per ottenere le ore di sole di una specifica giornata in un preciso luogo del pianeta.
Per poter determinare l'orario dell'alba e del tramonto di un esatto luogo del pianeta ed in un preciso giorno dell'anno, il modulo 'Astral' necessita di cinque parametri principalmente, ovvero:

- Giorno
- Mense
- Anno
- Latitudine
- Longitudine

Tutti e cinque i parametri sono presenti nel database Sysman. [58]

1.4 Keras-TensorFlow

La rete neurale artificiale Long Short-Term Memory è stata utilizzata per predire la radiazione solare per differenti orizzonti temporali.
Per poter implementare tale rete neurale è state scaricata ed installata la libreria 'Keras'.
La libreria Keras include implementazioni di reti neurali di alto livello, scritte in Python ed è in grado di funzionare su TensorFlow. [59]
Mentre TensorFlow è una libreria software open source per l'apprendimento automatico (machine learning), che fornisce moduli testati ed ottimizzati utili nella realizzazione di algoritmi per diversi tipi di compiti percettivi e di comprensione del linguaggio. [60]

2. Database

Per effettuare le predizioni della radiazione solare orarie, giornaliere e mensili sono stati utilizzate le rilevazioni giornaliere, fornite dall'azienda SysMan, che lavora principalmente in quattro grandi macro-aree:

- System integration e sviluppo
- Consulenza IT
- Service Assurance & Hardware Maintenance
- Research & Development

Uno degli ultimi progetti portati avanti da Sysman si chiama 'Bluleaf' e si occupa dell'agricoltura di precisione.
Tale progetto mira a monitorare le coltivazioni, utilizzando strumenti tecnologici come il pluviometro per monitorare le precipitazioni, il termometro per misurare la temperatura dell'aria, ecc.
Questi strumenti permettono di registrare le seguenti informazioni:

- Radiazione solare
- Temperatura dell'aria
- Umidità
- Velocità del vento
- Precipitazioni
- Pressione

Gli strumenti effettuano e registrano le rilevazioni ogni 15 minuti, e vengono poi associate altre informazioni come l'id della stazione, le coordinate geografiche e l'ora esatta del rilevamento.

3. Task

3.1 Operazioni di preparazione

Nei precedenti paragrafi, sono stati evidenziati i parametri meteorologici e climatici che sono stati registrati mediante le rilevazioni orarie, ovvero: la radiazione solare, la temperatura dell'aria, l'umidità, la velocità del vento, le precipitazioni e la pressione.

Tale dati sono stati raccolti e memorizzati in un database, messo a disposizione dall'azienda Sysman.

Per predire la radiazione solare, per diversi orizzonti temporali, sono stati forniti come parametri di input, ai modelli predittivi, i parametri meteorologici disponibili nel database Sysman e altri due parametri non presenti in tale database, ovvero: le ore di sole e la differenza di temperatura, utilizzati dalle fonti scientifiche, individuate mediante la ricerca bibliografica. Pertanto, non avendo a disposizione le ore di sole e la differenza di temperatura nel database Sysman, tali parametri sono stati determinati indirettamente, mediante appositi script in Python.

Per determinare le ore di sole in un giorno specifico dell'anno ed in un preciso luogo del pianeta è stata utilizzata la libreria 'Astral' che, permetteva di determinare l'ora dell'alba e l'ora del tramonto fornendo cinque principali informazioni, ovvero: la latitudine e la longitudine per determinare il luogo ed il giorno, il mese e l'anno per determinare il giorno specifico dell'anno, in quanto tutte queste informazioni influenza le ore di sole in una giornata. Lo script in Python scritto ed utilizzato è il seguente (figura 6):

```python
import pandas as pd
import numpy as np
import astral
import datetime

# rm_main is a mandatory function,
# the number of arguments has to be the number of input ports (can be none)
def rm_main(data):
    citta = data['poi'].tolist()
    latitudine = data['latitude'].tolist()
    longitudine = data['longitude'].tolist()
    altitudine = data['altitude'].tolist()
    date = data['date_time'].tolist()
    giorno = data['date_time_day'].tolist()
    mese = data['date_time_month'].tolist()
    anno = data['date_time_year'].tolist()
    ore_sole = np.empty(len(citta))
    for i in range(len(citta)):
        ore_sole[i] = get_ore_sole(citta[i], latitudine[i], longitudine[i], altitudine[i], giorno[i], mese[i], anno[i])
    data['ore_sole'] = ore_sole
    return data

def get_ore_sole(citta, latitudine, longitudine, altitudine, giorno, mese, anno):
    loc = astral.Location([citta, 'Italy', latitudine, longitudine, 'Europe/Rome', altitudine])
    loc.solar_depression = 'civil'
    sun = loc.sun(date=datetime.date(int(anno), int(mese), int(giorno)), local=True)
    sunrise = sun['sunrise']
    sunset = sun['sunset']
    hour_sunrise = int(sunrise.strftime('%H'))
    minute_sunrise = int(sunrise.strftime('%M'))
    hour_sunset = int(sunset.strftime('%H'))
    minute_sunset = int(sunset.strftime('%M'))
    difference = (minute_sunset + hour_sunset*60) - (minute_sunrise + hour_sunrise * 60)
    return (difference/60)
```

Figura 6: Script per determinare le ore di sole di una specifica giornata dell'anno e di un determinato luogo nel pianeta

Ricavate le due informazioni, ovvero: l'ora dell'alba e l'ora del tramonto, mediante la funzione sun() della libreria 'Astral', successivamente è stata determinata la differenza di tempo tra il tramonto e l'alba, così da ottenere il tempo di sole in una giornata.

Infine, tale valore è stato elaborato, manipolato e convertito per ottenere il tempo in ore, e quindi disporre come parametro delle ore di sole di una specifica giornata dell'anno e di un determinato luogo del pianeta.

Per quanto riguarda la differenza di temperatura, è stata determinata utilizzando la temperatura dell'aria massima e minima rilevate ad ogni ora, calcolando la differenza tra le due mediante un operatore di RapidMiner, che permette di generare nuovi attributi (parametri) specificando il nome che si vuole dare al nuovo attributo e la funzione o l'espressione che utilizza i valori degli attributi preesistenti per determinare i valori del nuovo attributo.

3.2 Feature Selection

Per determinare la predizione della radiazione solare, per diversi orizzonti temporali, avendo a disposizione 12 parametri meteorologici e climatici, è possibile determinare ben 4095 combinazioni di parametri, da poter fornire in input ai modelli predittivi per l'addestramento.
Infatti, abbiamo:

- Combinazioni con un solo parametro: 12
- Combinazioni con due parametri: 66
- Combinazioni con tre parametri: 220
- Combinazioni con quattro parametri: 495
- Combinazioni con cinque parametri: 792
- Combinazioni con sei parametri: 924
- Combinazioni con sette parametri: 792
- Combinazioni con otto parametri: 495
- Combinazioni con nove parametri: 220
- Combinazione con dieci parametri: 66
- Combinazione con undici parametri: 12
- Combinazioni con dodici parametri: 1

Inoltre, bisognerebbe applicare le 4095 combinazioni di parametri per ognuno degli orizzonti temporali e per ognuno dei modelli predettivi, pertanto, il numero totale delle combinazioni va moltiplicato per 3 (il numero degli orizzonti temporali), e successivamente di nuovo per 3 (il numero dei modelli di Machine Learning), così che si ottengono ben 36855 prove da effettuare per la predizione della radiazione solare.
Logicamente, per motivi di tempo e di potenza di calcolo, elaborare 36855 prove per predire la radiazione solare non è fattibile, pertanto, l'obiettivo è di ridurre drasticamente il numero delle combinazioni, e ciò è reso possibile

attraverso l'applicazione di algoritmi di Feature Selection.
Gli algoritmi di Feature Selection, come discusso nel capitolo precedente, si occupano di ridurre in numeri di parametri in ingresso da fornire ai modelli di predizioni, individuando le caratteristiche maggiormente rilevanti rispetto alle altre per la predizioni di un determinato parametro target.
Per questo studio, sono stati utilizzati tre algoritmi di feature selection per determinare i parametri meteorologici e climatici più rilevanti per effettuare le predizioni della radiazione solare, e questi sono:

- Pearson correlation (tipo filter)
- Recursive feature elimination attraverso il modello SVM (tipo wrapper)
- Random Forest (tipo wrapper)

Ciascuno dei tre algoritmi di Feature Selection è stato applicato per ogni orizzonte temporale, quindi orario, giornaliero e mensile, così da poter determinare quali fossero le caratteristiche più rilevanti per ogni orizzonte temporale secondo ogni algoritmo di feature selection.
Infine, tale procedimento è stato ripetuto per ben due volte, ovvero per ognuno dei due training set ottenuti adottando due criteri differenti per suddividere il data set (tali criteri sono stati adottati dalle fonti scientifiche individuate dalla ricerca bibliografica), e tali criteri sono i seguenti:

- Training set: ~70% del data set; Testing set: ~30% del data set
- Training set: ~80% del data set; Testing set: ~20% del data set

Pertanto, adottando il primo criterio, ovvero il training set è il 70% del data set e il resto è il testing set, otteniamo le seguenti partizioni delle rilevazioni per orizzonte temporale:

- Orizzonte orario → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018

- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017-Dicembre 2017, Testing set: Gennaio 2018 – Febbraio 2018

Mentre, con il secondo criterio, ovvero il training set è l'80% del data set e il resto è il testing set, otteniamo:

- Orizzonte orario → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017-Dicembre 2017, Testing set: Gennaio 2018 – Febbraio 2018

Pertanto, è possibile determinare i seguenti sotto task per applicare gli algoritmi di feature selection (tabella 6):

SOTTO TASK	ORIZZONTE	TRAINING E TESTING SET	FEATURE SELECTION
1	orario	~70% Training set ~30% Testing set	Pearson Correlation RFE-SVM Random Forest
		~80% Training set ~20% Testing set	
2	giornaliero	~70% Training set ~30% Testing set	Pearson Correlation RFE-SVM Random Forest
		~80% Training set ~20% Testing set	
3	mensile	~70% Training set ~30% Testing set	Pearson Correlation RFE-SVM Random Forest
		~80% Training set ~20% Testing set	

Tabella 6: Sotto task per applicare gli algoritmi di Feature Selection.

3.3 Predizione della radiazione solare

Per predire la radiazione solare, per multipli orizzonti temporali, sono stati applicati 3 algoritmi di Machine Learning, ovvero:

- *Multi-Layer Percepton*
- *Support Vector Machine*
- *Long Short-Term Memory*

Ciascuno dei tre algoritmi di predizione sono stati applicati per ognuno dei tre orizzonti temporali (orario, giornaliero, mensile).
Le combinazioni dei parametri da fornire in input ai modelli, per ogni orizzonte temporale, sono state determinate mediante gli algoritmi di feature selection, precedentemente discussi.
Per ogni orizzonte temporale, sono stati forniti ai modelli da un minimo di un parametro ad un massimo di quattro parametri in input dei parametri più rilevanti secondo gli algoritmi di feature selection.
Infine, tale procedimento è stato ripetuto per ben due volte, ovvero per ognuno dei due training set ottenuti adottando i due criteri differenti per la suddivisione del data set, e tali criteri sono i seguenti:

- Training set: ~70% del data set; Testing set: ~30% del data set
- Training set: ~80% del data set; Testing set: ~20% del data set

Pertanto, adottando il primo criterio, ovvero il training set è il 70% del data set e il resto è il testing set, otteniamo le seguenti partizioni delle rilevazioni per orizzonte temporale:

- Orizzonte orario → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018
- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018

- Orizzonte mensile → Training set: Luglio 2017-Dicembre 2017, Testing set: Gennaio 2018 – Febbraio 2018

Mentre, con il secondo criterio, ovvero il training set è l'80% del data set e il resto è il testing set, otteniamo:

- Orizzonte orario → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017-Dicembre 2017, Testing set: Gennaio 2018 – Febbraio 2018

Pertanto, è possibile determinare i seguenti sotto task per la predizione della radiazione solare (tabella 7):

SOTTO TASK	ORIZZONTE	TRAINING E TESTING SET	ALGORITMI DI MACHINE LEARNING
1	orario	~70% Training set ~30% Testing set	Multi-Layer Perceptron Support Vector Machine Long Short-Term Memory
		~80% Training set ~20% Testing set	
2	giornaliero	~70% Training set ~30% Testing set	Multi-Layer Perceptron Support Vector Machine Long Short-Term Memory
		~80% Training set ~20% Testing set	
3	mensile	~70% Training set ~30% Testing set	Multi-Layer Perceptron Support Vector Machine Long Short-Term Memory
		~80% Training set ~20% Testing set	

Tabella 7: sotto task per la predizione della radiazione solare per orizzonti temporali multipli.

Capitolo 6: Risultati sperimentali

1. Feature Selection

In questo task, sono stati determinati i parametri meteorologici e climatici più rilevanti per ottenere predizioni più accurate della radiazione solare, per ogni orizzonte temporale (orario, giornaliero, mensile), utilizzando i tre algoritmi di feature selection (Pearson Correlation, Recursive Feature Elimination con SVM e Random Forest).
Infine, tale procedimento è stato ripetuto per ben due volte, ovvero per ognuno dei due training set ottenuti adottando i due criteri differenti per la suddivisione del data set, e tali criteri sono i seguenti:

- Training set: ~70% del data set; Testing set: ~30% del data set
- Training set: ~80% del data set; Testing set: ~20% del data set

Pertanto, adottando il primo criterio, ovvero il training set è il 70% del data set e il resto è il testing set, otteniamo:

- Orizzonte orario → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018
- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017-Dicembre 2017, Testing set: Gennaio 2018-Febbraio 2018

Mentre, con il secondo criterio, ovvero il training set è l'80% del data set e il resto è il testing set, otteniamo:

- Orizzonte orario → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018

- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017-Dicembre 2017, Testing set: Gennaio 2018-Febbraio 2018

Sono state utilizzate le rilevazioni orarie della stazione meteorologica con ID=186, presente nella zona di Molfetta.

1.1 Orizzonte orario

Il processo di selezione della caratteristiche (Feature Selection) è descritto negli esempi successivi (figura 7):

- Feature Selection orario: processo nel quale sono contenute tutte le componenti che si rivelano utili per il filtraggio dei dati su determinate stazioni e in determinati intervalli di tempo, per determinare parametri aggiuntivi e per applicare algoritmi di Feature Selection.

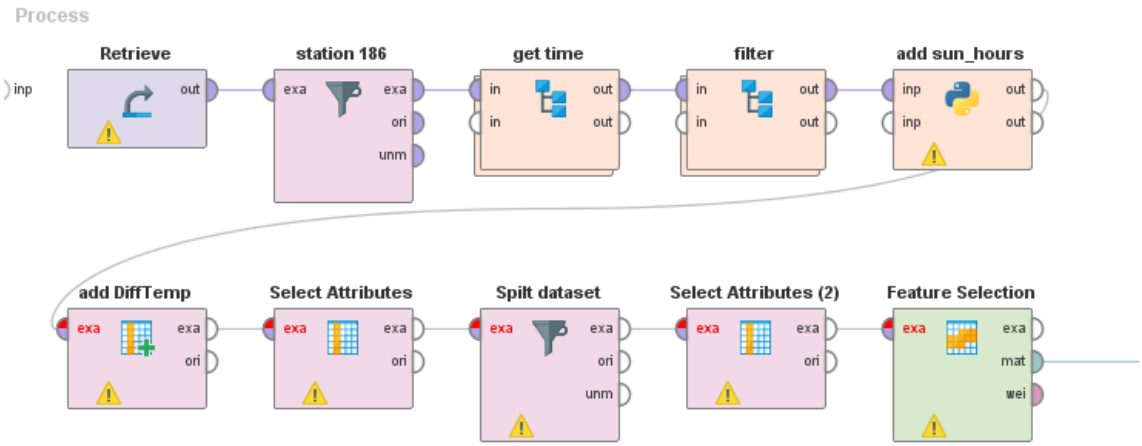


Figura 7: Processo di Feature Selection, per l'orizzonte orario, mediante RapidMiner.

- Retrieve: questo operatore permette di accedere e recuperare le rilevazioni orarie dei parametri meteorologici e climatici memorizzati nel database Sysman.
- Station 186: permette di filtrare le rilevazioni orarie mantenendo solamente le rilevazioni riguardanti la stazione 186 (zona di Molfetta)

- Get time: sotto processo che si occupa di scomporre la data delle rilevazione orarie in orario, giorno, mese e anno delle rilevazioni.
- Filter: sotto processo che si occupa di selezionare le rilevazioni degli anni 2017 e 2018, in quanto in tale intervallo non sono presenti rilevazioni errate o mancanti.
- Add sun hours: script in Python che permette di determinare e aggiungere alle rilevazioni le ore di sole in una specifica giornata dell'anno ed in un determinato luogo del pianeta attraverso la latitudine, la longitudine, il giorno, il mese e l'anno presenti nel database Sysman.
- Add DiffTemp: operatore che permette di generare un nuovo attributo, specificando il nome che si vuole dare e la formula che si vuole applicare per determinare i valori di esso, utilizzando i valori degli attributi già presenti. È utilizzato per determinare la differenza di temperatura, parametro non presente nel database Sysman, determinando la differenza tra la temperatura massima e minima.
- Select Attributes: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.
- Split dataset: operatore che si occupa di filtrare le rilevazioni di un determinato intervallo di tempo, per la precisione vengono selezionate solo le rilevazioni dal 1 Luglio 2017 al 17 Dicembre 2017 se il Training set è circa il 70% del data set, altrimenti dal 1 Luglio 2017 al 10 Gennaio 2018 se il Training set è circa l'80% del data set.
- Select Attributes 2: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione

solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.

- Feature Selection: script in Python che applica i due algoritmi di Feature Selection (Recursive Feature Elimination-SVM e Random Forest) o l'operatore di RapidMiner per applicare il Pearson Correlation per determinare i parametri più rilevanti per ottenere predizioni più accurate della radiazioni solari. Il codice sorgente viene riportato nelle figure 8 e 9.

```
import pandas as pd
import numpy as np
from sklearn.feature_selection import RFE
from sklearn.svm import SVR

# rm_main is a mandatory function,
# the number of arguments has to be the number of input ports (can be none)
def rm_main(data, data2):
     # train set
     X_train = data.drop(columns=['r_inc']).values.tolist()
     y_train = data['r_inc'].tolist()
     # feature selection
     estimator = SVR(kernel="linear")
     selector = RFE(estimator,1)
     selector = selector.fit(X_train, y_train)
     features = data.columns.tolist()
     features.remove('r_inc')
     values = {'features':features,'ranking':selector.ranking_}
     result = pd.DataFrame(values, columns=['features','ranking'])
     result = result.sort_values('ranking')
     return result
```

Figura 8: Algoritmo del recursive Feature Elimination con SVM

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor

# rm_main is a mandatory function,
# the number of arguments has to be the number of input ports (can be none)
def rm_main(data, data2):
     # train set
     X_train = data.drop(columns=['r_inc']).values.tolist()
     y_train = data['r_inc'].tolist()
     # feature selection
     rf = RandomForestRegressor()
     rf.fit(X_train, y_train)

     features = data.columns.tolist()
     features.remove('r_inc')
     values = {'features':features,'importanza':rf.feature_importances_}
     result = pd.DataFrame(values, columns=['features','importanza'])
     result = result.sort_values('importanza')
     return result
```

Figura 9: Algoritmo del Random Forest

1.1.1 Risultati

I risultati di seguito (tabella 8 e 9), mostrano le caratteristiche, in ordine di rilevanza, per la predizione della radiazione solare secondo ciascun algoritmo di Feature Selection, per l'orizzonte temporale orario.

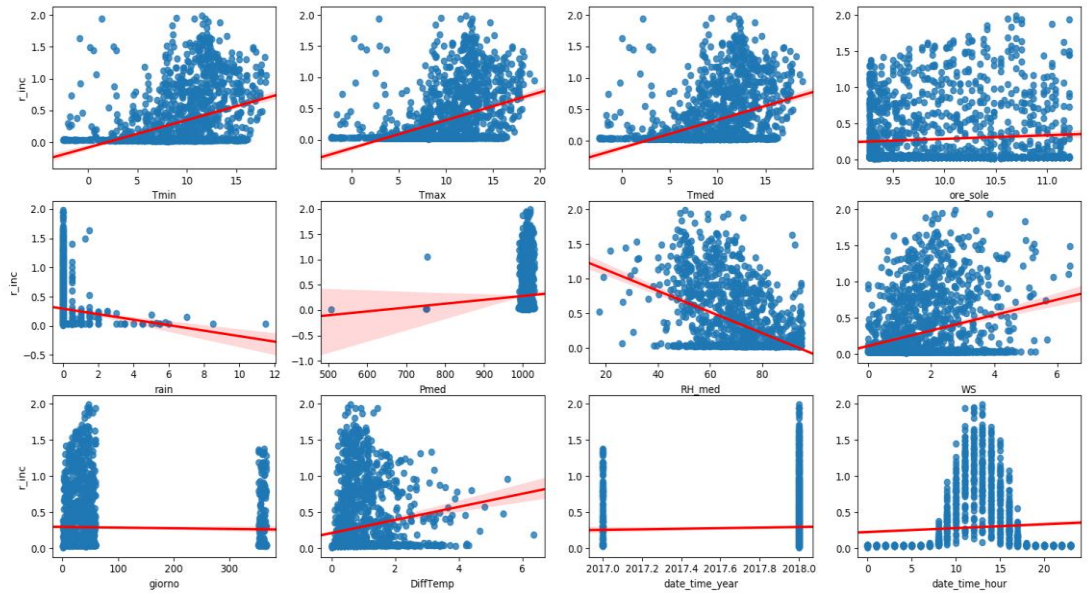
TRAINING SET: ~70% DEL DATA SET			
RANKING	PEARSON CORRELATION	RECURSIVE FEATURE ELIMINATION CON SVM	RANDOM FOREST
1	umidità	differenza di temperatura	orario
2	temperatura massima	ore di sole	umidità
3	temperatura media	temperatura minima	ore di sole
4	temperatura minima	precipitazioni	differenza di temperatura
5	velocità del vento	temperatura media	pressione
6	differenza di temperatura	anno	velocità del vento
7	orario	velocità del vento	giorno dell'anno
8	ore di sole	orario	temperatura minima
9	precipitazioni	umidità	temperatura massima
10	anno	temperatura massima	temperatura media
11	pressione	pressione	precipitazioni
12	giorno dell'anno	giorno dell'anno	anno

Tabella 8: Ranking delle caratteristiche, secondo 3 algoritmi di feature selection, sul training set pari al 70% del data set.

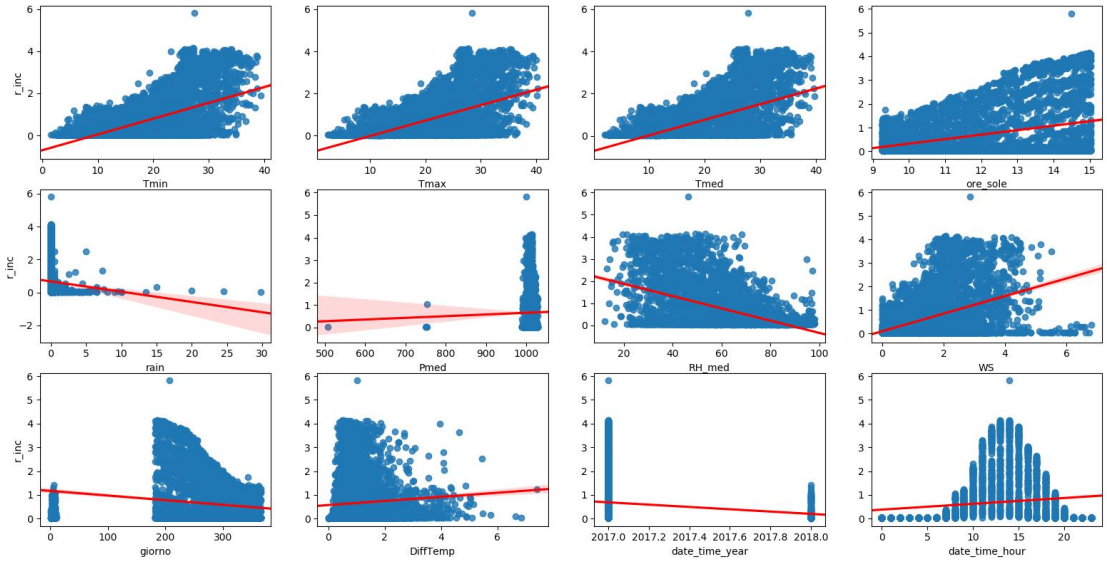
TRAINING SET: ~80% DEL DATA SET			
RANKING	PEARSON CORRELATION	RECURSIVE FEATURE ELIMINATION CON SVM	RANDOM FOREST
1	temperatura minima	velocità del vento	temperatura minima
2	temperatura media	ore di sole	orario
3	temperatura massima	temperatura massima	velocità del vento
4	umidità	temperatura media	ore di sole
5	velocità del vento	differenza di temperatura	giorno dell'anno
6	ore di sole	temperatura minima	pressione
7	orario	precipitazioni	umidità
8	giorno dell'anno	anno	temperatura massima
9	anno	orario	differenza di temperatura
10	differenza di temperatura	umidità	temperatura media
11	precipitazioni	pressione	precipitazioni
12	pressione	giorno dell'anno	anno

Tabella 9: Ranking delle caratteristiche, secondo 3 algoritmi di feature selection, sul training set pari al 80% del data set.

Correlazione tra la radiazione solare e gli altri parametri meteorologici con il Training set pari a circa il 70% del Data set



Correlazione tra la radiazione solare e gli altri parametri meteorologici con il Training set pari a circa l'80% del Data set



1.2 Orizzonte giornaliero

Il processo di selezione della caratteristiche (Feature Selection) è descritto negli esempi successivi (figura 10):

- Feature Selection giornaliero: processo nel quale sono contenute tutte le componenti che si rivelano utili per il filtraggio dei dati su determinate stazioni e in determinati intervalli di tempo, e per determinare parametri aggiuntivi e per applicare algoritmi di Feature Selection.

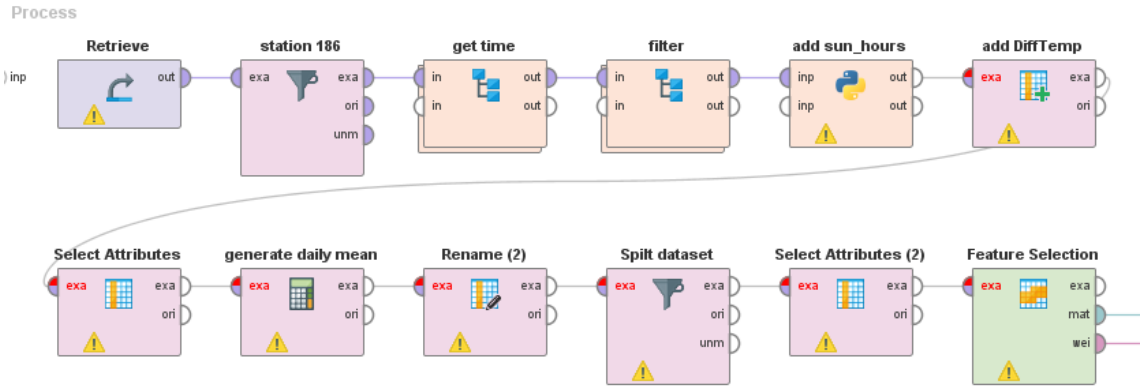


Figura 10: Processo di Feature Selection, per l'orizzonte giornaliero, mediante RapidMiner.

- Retrieve: questo operatore permette di accedere e recuperare le rilevazioni orarie dei parametri meteorologici e climatici memorizzati nel database Sysman.
- Station 186: permette di filtrare le rilevazioni orarie mantenendo solamente le rilevazioni riguardanti la stazione 186 (zona di Molfetta)
- Get time: sotto processo che si occupa di scomporre la data delle rilevazione orarie in orario, giorno, mese e anno delle rilevazioni.
- Filter: sotto processo che si occupa di selezionare le rilevazioni degli anni 2017 e 2018, in quanto in tale intervallo non sono presenti rilevazioni errate o mancanti.
- Add sun hours: script in Python che permette di determinare e aggiungere alle rilevazioni le ore di sole in una specifica giornata

dell'anno ed in un determinato luogo del pianeta attraverso la latitudine, la longitudine, il giorno, il mese e l'anno presenti nel database Sysman.

- Add DiffTemp: operatore che permette di generare un nuovo attributo, specificando il nome che si vuole dare e la formula che si vuole applicare per determinare i valori di esso, utilizzando i valori degli attributi già presenti. È utilizzato per determinare la differenza di temperatura, parametro non presente nel database Sysman, determinando la differenza tra la temperatura massima e minima.
- Select Attributes: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.
- Generate daily mean: operatore che permette di effettuare funzioni di aggregazione. In questo caso, sono determinate le medie giornaliere dei singoli parametri, raggruppando per giorno ed anno.
- Rename: operatore che permette di modificare i nomi degli attributi.
- Split dataset: operatore che si occupa di filtrare le rilevazioni di un determinato intervallo di tempo, per la precisione vengono selezionate solo le rilevazioni dal 1 Luglio 2017 al 17 Dicembre 2017 se il Training set è circa il 70% del data set, altrimenti dal 1 Luglio 2017 al 10 Gennaio 2018 se il Training set è circa l'80% del data set.
- Select Attributes 2: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.

- Feature Selection: script in Python che applica i due algoritmi di Feature Selection (Recursive Feature Elimination-SVM e Random Forest) o l'operatore di RapidMiner per applicare il Pearson Correlation per determinare i parametri più rilevanti per ottenere predizioni più accurate della radiazioni solari.

1.2.1 Risultati

I risultati di seguito (tabella 10 e 11), mostrano le caratteristiche, in ordine di rilevanza, per la predizione della radiazione solare secondo ciascun algoritmo di Feature Selection, per l'orizzonte temporale giornaliero.

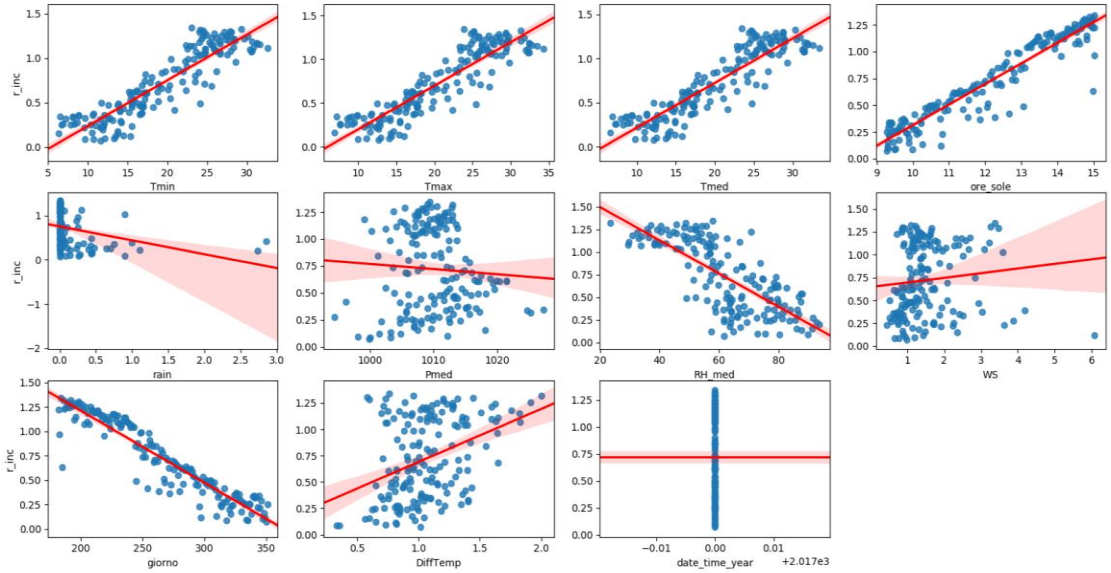
TRAINING SET: ~70% DEL DATA SET			
RANKING	PEARSON CORRELATION	RECURSIVE FEATURE ELIMINATION CON SVM	RANDOM FOREST
1	ore di sole	precipitazioni	differenza di temperatura
2	giorno dell'anno	anno	ore di sole
3	temperatura minima	velocità del vento	precipitazioni
4	temperatura media	temperatura massima	umidità
5	temperatura massima	temperatura minima	temperatura minima
6	umidità	differenza di temperatura	pressione
7	differenza di temperatura	temperatura media	temperatura massima
8	precipitazioni	ore di sole	velocità del vento
9	velocità del vento	umidità	giorno dell'anno
10	pressione	pressione	temperatura media
11	anno	giorno dell'anno	anno

Tabella 10: Ranking delle caratteristiche, secondo 3 algoritmi di feature selection, sul training set pari al 70% del data set.

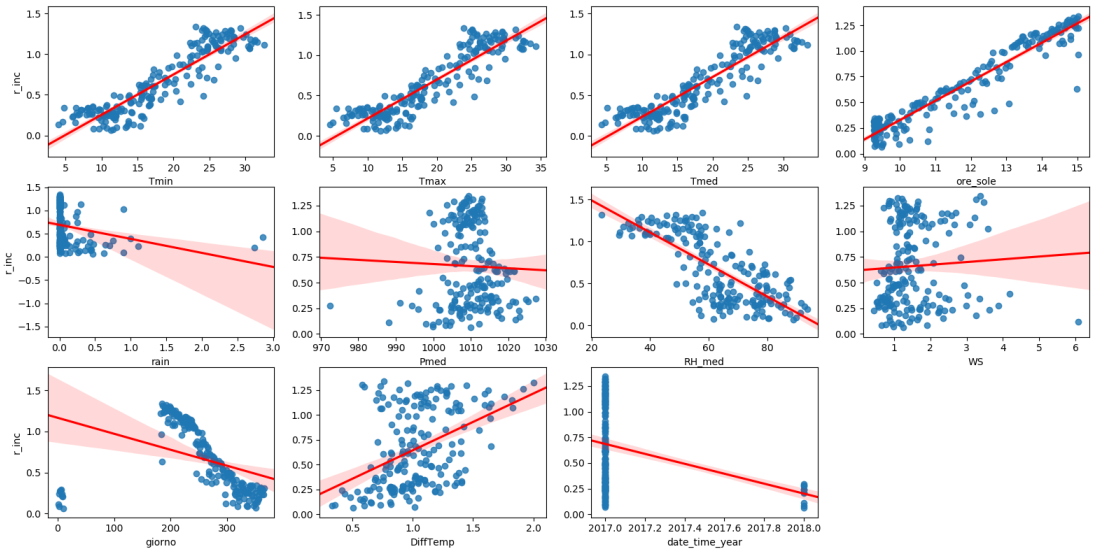
TRAINING SET: ~80% DEL DATA SET			
RANKING	PEARSON CORRELATION	RECURSIVE FEATURE ELIMINATION CON SVM	RANDOM FOREST
1	ore di sole	ore di sole	ore di sole
2	temperatura minima	precipitazioni	temperatura massima
3	temperatura media	velocità del vento	giorno dell'anno
4	temperatura massima	temperatura minima	umidità
5	umidità	differenza di temperatura	differenza di temperatura
6	differenza di temperatura	temperatura massima	temperatura minima
7	giorno dell'anno	anno	precipitazioni
8	anno	umidità	pressione
9	precipitazioni	pressione	temperatura media
10	velocità del vento	temperatura media	velocità del vento
11	pressione	giorno dell'anno	anno

Tabella 11: Ranking delle caratteristiche, secondo 3 algoritmi di feature selection, sul training set pari al 80% del data set.

Correlazione tra la radiazione solare e gli altri parametri meteorologici con il Training set pari a circa il 70% del Data set



Correlazione tra la radiazione solare e gli altri parametri meteorologici con il Training set pari a circa l'80% del Data set



1.3 Orizzonte mensile

Il processo di selezione della caratteristiche (Feature Selection) è descritto negli esempi successivi (figura 11):

- Feature Selection mensile: processo nel quale sono contenute tutte le componenti che si rivelano utili per il filtraggio dei dati su determinate stazioni e in determinati intervalli di tempo, e per determinare parametri aggiuntivi e per applicare algoritmi di Feature Selection.

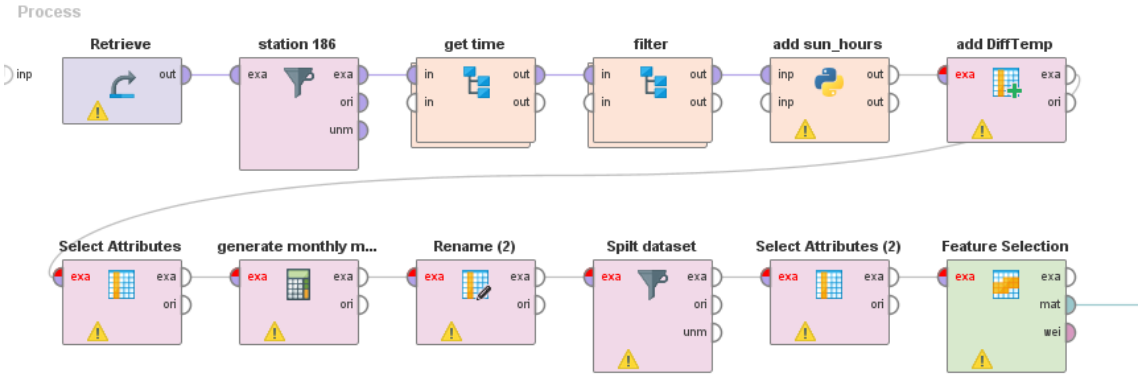


Figura 11: Processo di Feature Selection, per l'orizzonte mensile, mediante RapidMiner.

- Retrieve: questo operatore permette di accedere e recuperare le rilevazioni orarie dei parametri meteorologici e climatici memorizzati nel database Sysman.
- Station 186: permette di filtrare le rilevazioni orarie mantenendo solamente le rilevazioni riguardanti la stazione 186 (zona di Molfetta)
- Get time: sotto processo che si occupa di scomporre la data delle rilevazione orarie in orario, giorno, mese e anno delle rilevazioni.
- Filter: sotto processo che si occupa di selezionare le rilevazioni degli anni 2017 e 2018, in quanto in tale intervallo non sono presenti rilevazioni errate o mancanti.

- Add sun hours: script in Python che permette di determinare e aggiungere alle rilevazioni le ore di sole in una specifica giornata dell'anno ed in un determinato luogo del pianeta attraverso la latitudine, la longitudine, il giorno, il mese e l'anno presenti nel database Sysman.
- Add DiffTemp: operatore che permette di generare un nuovo attributo, specificando il nome che si vuole dare e la formula che si vuole applicare per determinare i valori di esso, utilizzando i valori degli attributi già presenti. È utilizzato per determinare la differenza di temperatura, parametro non presente nel database Sysman, determinando la differenza tra la temperatura massima e minima.
- Select Attributes: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.
- Generate monthly mean: operatore che permette di effettuare funzioni di aggregazione. In questo caso, sono determinate le medie mensili dei singoli parametri, raggruppando per mese ed anno.
- Rename: operatore che permette di modificare i nomi degli attributi.
- Split dataset: operatore che si occupa di filtrare le rilevazioni di un determinato intervallo di tempo, per la precisione vengono selezionate solo le rilevazioni dal 1 Luglio 2017 al 17 Dicembre 2017 se il Training set è circa il 70% del data set, altrimenti dal 1 Luglio 2017 al 10 Gennaio 2018 se il Training set è circa l'80% del data set.
- Select Attributes 2: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione

solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.

- Feature Selection: script in Python che applica i due algoritmi di Feature Selection (Recursive Feature Elimination-SVM e Random Forest) o l'operatore di RapidMiner per applicare il Pearson Correlation per determinare i parametri più rilevanti per ottenere predizioni più accurate della radiazioni solari.

1.3.1 Risultati

I risultati di seguito (tabelle 12 e 13), mostrano le caratteristiche, in ordine di rilevanza, per la predizione della radiazione solare secondo ciascun algoritmo di Feature Selection, per l'orizzonte temporale mensile.

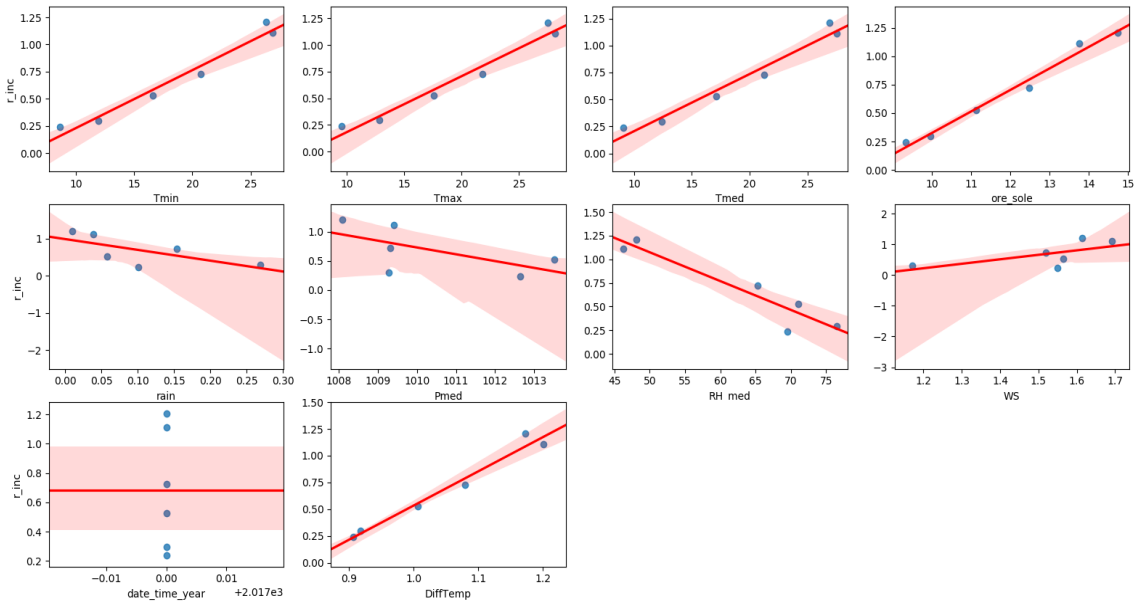
TRAINING SET: ~70% DEL DATA SET			
RANKING	PEARSON CORRELATION	RECURSIVE FEATURE ELIMINATION CON SVM	RANDOM FOREST
1	ore di sole	umidità	temperatura minima
2	differenza di temperatura	temperatura massima	temperatura massima
3	temperatura massima	temperatura media	umidità
4	temperatura minima	temperatura minima	differenza di temperatura
5	temperatura media	ore di sole	pressione
6	umidità	pressione	temperatura media
7	precipitazioni	differenza di temperatura	precipitazioni
8	velocità del vento	precipitazioni	ore di sole
9	pressione	velocità del vento	velocità del vento
10	anno	anno	anno

Tabella 12: Ranking delle caratteristiche, secondo 3 algoritmi di feature selection, sul training set pari al 70% del data set.

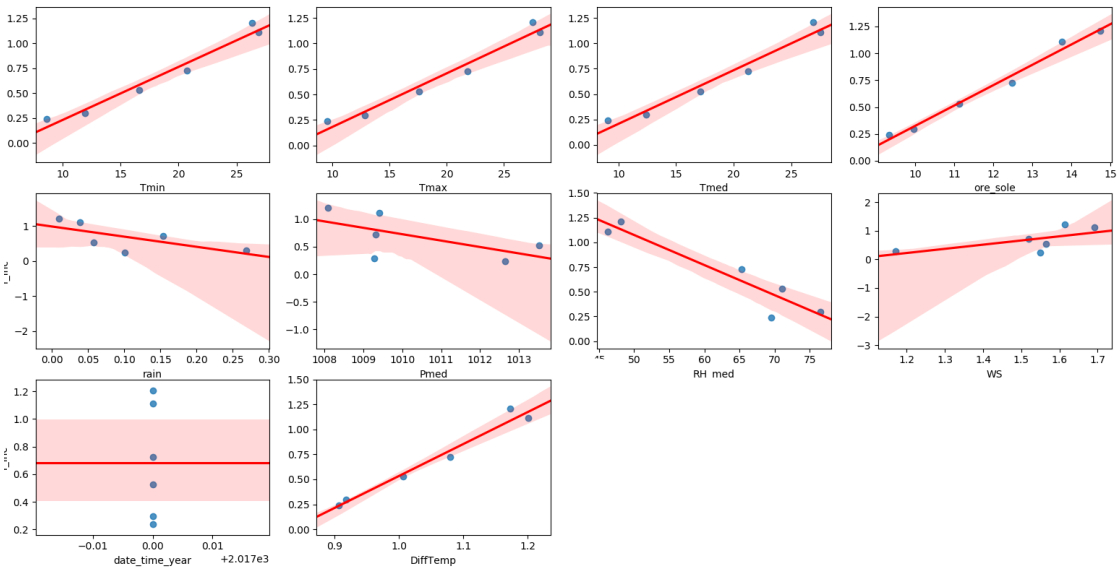
TRAINING SET: ~80% DEL DATA SET			
RANKING	PEARSON CORRELATION	RECURSIVE FEATURE ELIMINATION CON SVM	RANDOM FOREST
1	ore di sole	umidità	temperatura minima
2	differenza di temperatura	temperatura massima	temperatura massima
3	temperatura massima	temperatura media	umidità
4	temperatura media	temperatura minima	differenza di temperatura
5	temperatura minima	ore di sole	pressione
6	umidità	pressione	temperatura media
7	precipitazioni	differenza di temperatura	precipitazioni
8	velocità del vento	precipitazioni	ore di sole
9	pressione	velocità del vento	velocità del vento
10	anno	anno	anno

Tabella 13: Ranking delle caratteristiche, secondo 3 algoritmi di feature selection, sul training set pari al 80% del data set.

Correlazione tra la radiazione solare e gli altri parametri meteorologici con il Training set pari a circa il 70% del Data set



Correlazione tra la radiazione solare e gli altri parametri meteorologici con il Training set pari a circa l'80% del Data set



2. Predizione della radiazione solare

In questo task, sono state predette le radiazioni solari, per ognuno degli orizzonti temporali (orario, giornaliero, mensile), attraverso gli algoritmi di Machine Learning e riportate le accuratezze di esse mediante due metriche di errore:

- errore relativo:

$$Relative\ Error\ (E_r) = \frac{(X_i - X_t)}{X_t}$$

  Dove X_i sono i valori della radiazione solare predetti e X_t sono i valori osservati.

- radice quadrata dell'errore quadratico medio:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$$

  dove $\hat{y}_t$ sono i valori della radiazione solare predetti e y_t sono i valori osservati.

Gli algoritmi di Machine Learning utilizzati per la predizione della radiazione solare sono:

- Multi-Layer Perceptron
- Support Vector Machine
- Long Short-Term Memory

Per quanto riguarda i parametri meteorologici più rilevanti, da fornire in input ai modelli, per ottenere predizioni più accurate, sono stati determinati mediante gli algoritmi di feature selection, nel precedentemente paragrafo. Ad ogni modello di Machine Learning è stato fornito da un minimo di un parametro ad un massimo di 4 parametri in input, secondo i parametri più rilevanti determinati dagli algoritmi di feature selection.

Infine, tale procedimento è stato ripetuto per ben due volte, ovvero per ognuno dei due training set ottenuti adottando i due criteri differenti per la suddivisione del data set, e tali criteri sono i seguenti:

- Training set: ~70% del data set; Testing set: ~30% del data set
- Training set: ~80% del data set; Testing set: ~20% del data set

Pertanto, adottando il primo criterio, ovvero il training set è il 70% del data set e il resto è il testing set, otteniamo le seguenti partizioni delle rilevazioni per ciascun orizzonte temporale:

- Orizzonte orario → Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018
- Orizzonte giornaliero→ Training set: 1 Luglio 2017 – 17 Dicembre 2017, Testing set: 18 Dicembre 2017 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017 – Dicembre 2017, Testing set: Gennaio 2018 – Febbraio 2018

Mentre, con il secondo criterio, ovvero il training set è l'80% del data set e il resto è il testing set, otteniamo:

- Orizzonte orario → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte giornaliero → Training set: 1 Luglio 2017 – 10 Gennaio 2018, Testing set: 11 Gennaio 2018 – 28 Febbraio 2018
- Orizzonte mensile → Training set: Luglio 2017 – Dicembre 2017, Testing set: Gennaio 2018 – Febbraio 2018

Sono state utilizzate le rilevazioni orarie della stazione meteorologica con ID=186, presente nella zona di Molfetta.

2.1 Orizzonte orario

Il processo di predizione della radiazione solare oraria è descritto negli esempi successivi (figura 12):

- Prediction r_inc orario: processo nel quale sono contenute tutte le componenti che si rivelano utili per il filtraggio dei dati su determinate stazioni e in determinati intervalli di tempo, per determinare parametri aggiuntivi e per applicare algoritmi di Machine Learning per predire la radiazione solare.

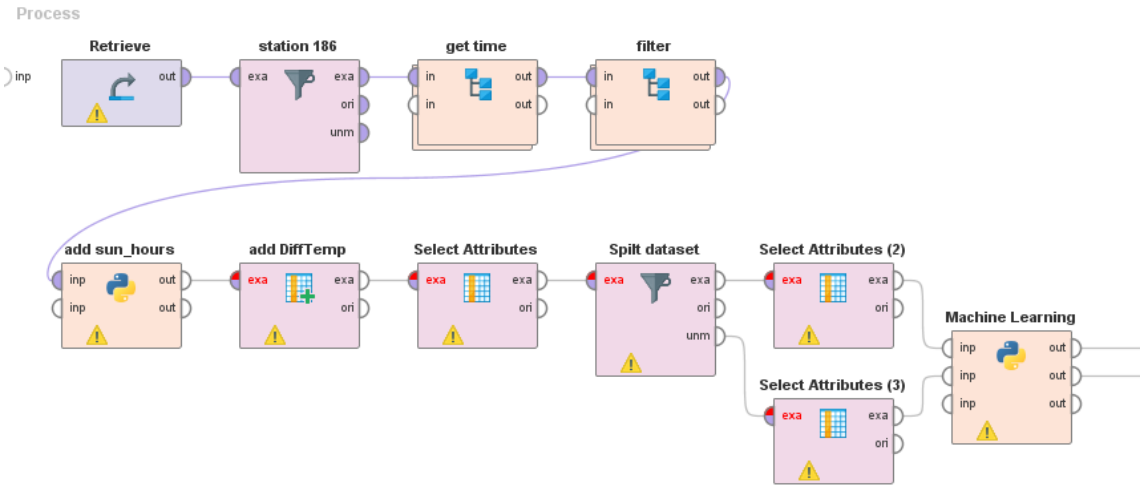


Figura 12: Processo di predizione della radiazione solare mediante modelli predittivi, per l'orizzonte orario, mediante RapidMiner.

- Retrieve: questo operatore permette di accedere e recuperare le rilevazioni orarie dei parametri meteorologici e climatici memorizzati nel database Sysman.
- Station 186: permette di filtrare le rilevazioni orarie mantenendo solamente le rilevazioni riguardanti la stazione 186 (zona di Molfetta)
- Get time: sotto processo che si occupa di scomporre la data delle rilevazione orarie in orario, giorno, mese e anno delle rilevazioni.

- Filter: sotto processo che si occupa di selezionare le rilevazioni degli anni 2017 e 2018, in quanto in tale intervallo non sono presenti rilevazioni errate o mancanti.
- Add sun hours: script in Python che permette di determinare e aggiungere alle rilevazioni le ore di sole in una specifica giornata dell'anno ed in un determinato luogo del pianeta attraverso la latitudine, la longitudine, il giorno, il mese e l'anno presenti nel database Sysman.
- Add DiffTemp: operatore che permette di generare un nuovo attributo, specificando il nome che si vuole dare e la formula che si vuole applicare per determinare i valori di esso, utilizzando i valori degli attributi già presenti. È utilizzato per determinare la differenza di temperatura, parametro non presente nel database Sysman, determinando la differenza tra la temperatura massima e minima.
- Select Attributes: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.
- Split dataset: operatore che si occupa di partizionare le rilevazioni del dataset set in due partizioni, secondo determinati intervalli di tempo, così che una partizione sarà il Training set e l'altra il Testing set. Per la precisione, il Training set sarà composto dalle rilevazioni che vanno dal 1 Luglio 2017 al 17 Dicembre 2017 ed il Testing set dal 18 Dicembre 2017 al 28 Febbraio 2018 se il Training set è circa il 70% del data set, altrimenti il Training set sarà composto dalle rilevazioni che vanno dal 1 Luglio 2017 al 10 Gennaio 2018 ed il Testing set dal 11 Gennaio 2018 al 28 Febbraio 2018 se il Training set è circa l'80% del data set.

- Select Attributes 2: permette di selezionare solo i parametri di interesse per il Training set da voler fornire in input al modello di predizione.
- Select Attributes 3: permette di selezionare solo i parametri di interesse per il Testing set da voler fornire in input al modello di predizione.
- Machine Learning: script in Python che applica i tre algoritmi di Machine Learning (Multi-Layer Perceptron, Support Vector Machine e Long Short-Term Memory), per predire le radiazioni solari e determinare le accuratezze di esse. Il codice sorgente viene riportato qui di seguito, nelle figure 13,14 e 15:

```
import pandas as pd
from sklearn import svm
from sklearn import metrics
import numpy as np
import sklearn

# rm_main is a mandatory function,
# the number of arguments has to be the number of input ports (can be none
def rm_main(data, data2):

     # architettura
     epsilon = 0
     gamma = 0
     error = 100
     relative_error = 0

     # train set
     X_train = data.drop(columns=['r_inc']).values.tolist()
     y_train = data['r_inc'].tolist()

     # test set
     X_test = data2.drop(columns=['r_inc']).values.tolist()
     y_test = data2['r_inc'].tolist()

     for l in range(100):
          for k in range(100):
               # create model SVM
               SVM = svm.SVR(epsilon=((l+1)*0.001), gamma=((k+1)*0.001))
               # training
               SVM.fit(X_train, y_train)
               # predizione e valori predetti
               y_pred = SVM.predict(X_test)
               # RMSE
               rmse = np.sqrt(metrics.mean_squared_error(y_test, y_pred))
               # select arch
               if rmse < error:
                    relative_error = er(y_test, y_pred)
                    error = rmse
                    epsilon = (l+1)*0.001
                    gamma = (k+1)*0.001

     print(epsilon)
     print(gamma)
     print(error)
     print(relative_error)

     return data, data2

def er(y_test, y_pred):
     er = 0
     for i in range(len(y_test)):
          er = er + (abs(y_test[i]-y_pred[i])/y_test[i])
     er = er/len(y_test)
     return er
```

Figura 13: Algoritmo per la predizione della radiazione solare mediante Support Vector Machine.

```python
import pandas as pd
from sklearn.neural_network import MLPRegressor
from sklearn.preprocessing import StandardScaler
from sklearn import metrics
import numpy as np
import sklearn

# rm_main is a mandatory function,
# the number of arguments has to be the number of input ports (can be none)
def rm_main(data, data2):

    # architettura
    lv_1 = 0
    lv_2 = 0
    error = 100
    relative_error = 0

    # train set
    X_train = data.drop(columns=['r_inc']).values.tolist()
    y_train = data['r_inc'].tolist()

    # test set
    X_test = data2.drop(columns=['r_inc']).values.tolist()
    y_test = data2['r_inc'].tolist()

    for i in range(100):
        # create model MLP
        mlp = MLPRegressor(hidden_layer_sizes=(i+1), solver='lbfgs', random_state=1)
        # training
        mlp.fit(X_train, y_train)
        # predizione e valori predetti
        y_pred = mlp.predict(X_test)
        # RMSE
        rmse = np.sqrt(metrics.mean_squared_error(y_test, y_pred))
        # select arch
        if rmse < error:
            error = rmse
            lv_1 = i+1
            lv_2 = 0

    for i in range(100):
        for l in range(100):
            # create model MLP
            mlp = MLPRegressor(hidden_layer_sizes=(i+1, l+1), solver='lbfgs', random_state=1)
            # training
            mlp.fit(X_train, y_train)
            # predizione e valori predetti
            y_pred = mlp.predict(X_test)
            # RMSE
            rmse = np.sqrt(metrics.mean_squared_error(y_test, y_pred))
            # select arch
            if rmse < error:
                relative_error = er(y_test, y_pred)
                error = rmse
                lv_1 = i+1
                lv_2 = l+1

    print(lv_1)
    print(lv_2)
    print(error)
    print(relative_error)

    return data, data2

def er(y_test, y_pred):
    er = 0
    for i in range(len(y_test)):
        er = er + (abs(y_test[i]-y_pred[i])/y_test[i])
    er = er/len(y_test)
    return er
```

Figura 14: Algoritmo per la predizione della radiazione solare mediante Multi-Layer Perceptron.

```
import pandas as pd
from sklearn import metrics
import numpy as np
import sklearn
np.random.seed(2018)
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM

def rm_main(data, data2):
     # architettura
     lv = 100
     epoche = 0
     error = 100
     relative_error = 0

     # train set
     X_train = np.array(data.drop(columns=['r_inc']).values.tolist())
     y_train = np.array(data['r_inc'].tolist())
     # test set
     X_test = np.array(data2.drop(columns=['r_inc']).values.tolist())
     y_test = np.array(data2['r_inc'].tolist())
     # reshape input to be 3D [samples, timesteps, features]
     X_train = X_train.reshape((X_train.shape[0], 1, X_train.shape[1]))
     X_test = X_test.reshape((X_test.shape[0], 1, X_test.shape[1]))

     for l in range(50):
          # design network
          model = Sequential()
          model.add(LSTM(lv, input_shape=(X_train.shape[1], X_train.shape[2])))
          model.add(Dense(1))
          model.compile(optimizer='rmsprop', loss='mse')
          # fit network
          model.fit(X_train, y_train, epochs=(l+1), verbose=0, shuffle=False)
          # predizione e valori predetti
          y_pred = model.predict(X_test)
          # RMSE
          rmse = np.sqrt(metrics.mean_squared_error(y_test, y_pred))
          # select arch
          if rmse < error:
               relative_error = er(y_test, y_pred)
               error = rmse
               epoche = l+1
     print(lv)
     print(epoche)
     print(error)
     print(relative_error)
     return data, data2

def er(y_test, y_pred):
     er = 0
     for i in range(len(y_test)):
          er = er + (abs(y_test[i]-y_pred[i])/y_test[i])
     er = er/len(y_test)
     return er
```

Figura 15: Algoritmo per la predizione della radiazione solare mediante Long Short-Term Memory.

2.1.1 Risultati

FEATURE SELECTION: PEARSON CORRELATION TRAINING SET: ~70% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
RH_med	0.961	4.85	0.98	3.65	0.944	6.98
RH_med + Tmax	0.886	7.87	0.924	7.04	1.03	11.86
RH_med + Tmax + Tmed	0.888	8.23	0.957	7.23	1.013	11.27
RH_med + Tmax + Tmed + Tmin	0.873	7.03	0.998	6.87	1.0015	10.84

Tabella 14: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte orario, sul training set pari al 70% del data set, fornendo in input i parametri più rilevanti secondo il Pearson Correlation.

FEATURE SELECTION: RECURSIVE FEATURE ELIMINATION (SVM) TRAINING SET: ~70% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
DiffTemp	1.121	3.69	1.216	1.71	1.117	4.396
DiffTemp + ore_sole	1.022	6.32	1.125	3.01	1.06	5.5
ore_sole + DiffTemp + Tmin	0.846	8.32	0.895	8.38	1.002	10.865
ore_sole + DiffTemp + Tmin + rain	0.835	7.32	0.894	8.34	1.002	11.63

Tabella 15: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte orario, sul training set pari al 70% del data set, fornendo in input i parametri più rilevanti secondo il Recursive Feature Elimination con SVM.

FEATURE SELECTION: RANDOM FOREST TRAINING SET: ~70% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
orario	0.916	1.31	0.932	0.405	0.895	0.952
orario + RH_med	0.841	2.10	0.901	0.99	0.808	1.888
orario + RH_med + ore_sole	0.476	1.73	0.828	2.41	0.579	1.752
orario + RH_med + ore_sole + DiffTemp	0.5337	2.37	0.826	2.43	0.550	1.843

Tabella 16: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte orario, sul training set pari al 70% del data set, fornendo in input i parametri più rilevanti secondo il Random Forest.

FEATURE SELECTION: PEARSON CORRELATION TRAINING SET: ~80% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
Tmin	0.456	1.68	0.522	0.655	0.557	2.876
Tmin + Tmed	0.446	2.16	0.518	0.606	0.574	2.294
Tmin + Tmed + Tmax	0.441	3.412	0.517	0.771	0.574	2.294
Tmin + Tmed + Tmax + RH_med	0.392	3.015	0.475	0.876	0.432	1.507

Tabella 17: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte orario, sul training set pari al 80% del data set, fornendo in input i parametri più rilevanti secondo il Pearson Correlation.

FEATURE SELECTION: RECURSIVE FEATURE ELIMINATION (SVM) TRAINING SET: ~80% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
WS	0.575	9.92	0.44	3.324	0.437	3.667
WS + ore_sole	0.459	4.848	0.449	3.183	0.450	3.570
WS + ore_sole + Tmax	0.426	2.412	0.486	0.924	0.564	3.179
WS + ore_sole + Tmax + Tmed	0.431	1.77	0.489	0.925	0.570	2.995

Tabella 18: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte orario, sul training set pari al 80% del data set, fornendo in input i parametri più rilevanti secondo il Recursive Feature Elimination con SVM.

FEATURE SELECTION: RANDOM FOREST TRAINING SET: ~80% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
Tmin	0.456	1.68	0.522	0.655	0.557	2.876
Tmin + date_time_hour	0.271	1.928	0.508	1.197	0.349	1.075
Tmin + date_time_hour + WS	0.265	1.58	0.488	0.902	0.335	1.201
Tmin + date_time_hour + WS + ore_sole	0.243	0.908	0.291	0.473	0.313	1.017

Tabella 19: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte orario, sul training set pari al 80% del data set, fornendo in input i parametri più rilevanti secondo il Random Forest.

2.2 Orizzonte giornaliero

Il processo di predizione della radiazione solare giornaliera è descritto negli esempi successivi (figura 16):

- Prediction r_inc giornaliero: processo nel quale sono contenute tutte le componenti che si rivelano utili per il filtraggio dei dati su determinate stazioni e in determinati intervalli di tempo, per determinare parametri aggiuntivi e per applicare algoritmi di Machine Learning per predire la radiazione solare.

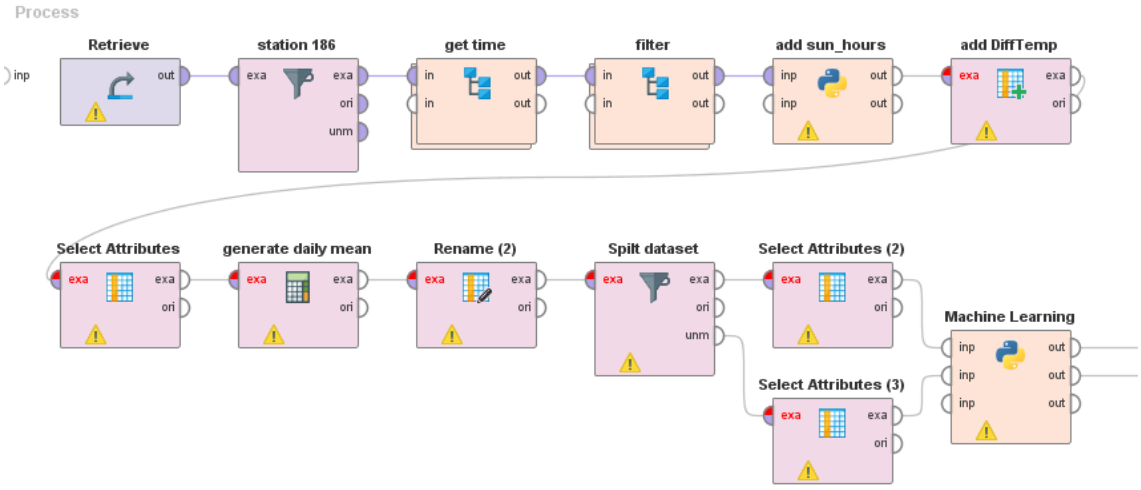


Figura 16: Processo di predizione della radiazione solare mediante modelli predittivi, per l'orizzonte giornaliero, mediante RapidMiner.

- Retrieve: questo operatore permette di accedere e recuperare le rilevazioni orarie dei parametri meteorologici e climatici memorizzati nel database Sysman.
- Station 186: permette di filtrare le rilevazioni orarie mantenendo solamente le rilevazioni riguardanti la stazione 186 (zona di Molfetta)
- Get time: sotto processo che si occupa di scomporre la data delle rilevazione orarie in orario, giorno, mese e anno delle rilevazioni.

- Filter: sotto processo che si occupa di selezionare le rilevazioni degli anni 2017 e 2018, in quanto in tale intervallo non sono presenti rilevazioni errate o mancanti.
- Add sun hours: script in Python che permette di determinare e aggiungere alle rilevazioni le ore di sole in una specifica giornata dell'anno ed in un determinato luogo del pianeta attraverso la latitudine, la longitudine, il giorno, il mese e l'anno presenti nel database Sysman.
- Add DiffTemp: operatore che permette di generare un nuovo attributo, specificando il nome che si vuole dare e la formula che si vuole applicare per determinare i valori di esso, utilizzando i valori degli attributi già presenti. È utilizzato per determinare la differenza di temperatura, parametro non presente nel database Sysman, determinando la differenza tra la temperatura massima e minima.
- Select Attributes: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.
- Generate daily mean: operatore che permette di effettuare funzioni di aggregazione. In questo caso, sono determinate le medie giornaliere dei singoli parametri, raggruppando per giorno ed anno.
- Rename: operatore che permette di modificare i nomi degli attributi.
- Split dataset: operatore che si occupa di partizionare le rilevazioni del dataset set in due partizioni, secondo determinati intervalli di tempo, così che una partizione sarà il Training set e l'altra il Testing set. Per la precisione, il Training set sarà composto dalle rilevazioni che vanno dal 1 Luglio 2017 al 17 Dicembre 2017 ed il Testing set dal 18 Dicembre 2017 al 28 Febbraio 2018 se il Training set è circa il 70% del data set,

altrimenti il Training set sarà composto dalle rilevazioni che vanno dal 1 Luglio 2017 al 10 Gennaio 2018 ed il Testing set dal 11 Gennaio 2018 al 28 Febbraio 2018 se il Training set è circa l'80% del data set.

- Select Attributes 2: permette di selezionare solo i parametri di interesse per il Training set da voler fornire in input al modello di predizione.
- Select Attributes 3: permette di selezionare solo i parametri di interesse per il Testing set da voler fornire in input al modello di predizione.
- Machine Learning: script in Python che applica i tre algoritmi di Machine Learning (Multi-Layer Perceptron, Support Vector Machine e Long Short-Term Memory), per predire le radiazioni solari e determinare le accuratezze di esse.

2.2.1 Risultati

FEATURE SELECTION: PEARSON CORRELATION TRAINING SET: ~70% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
ore_sole	0.346	0.405	0.447	0.481	0.519	0.538
ore_sole + giorno dell'anno	0.128	0.255	0.561	0.632	0.469	0.616
ore_sole + giorno dell'anno + Tmin	0.139	0.287	0.56	0.592	0.398	0.489
ore_sole + giorno dell'anno + Tmin + Tmed	0.140	0.289	0.562	0.595	0.431	0.486

Tabella 20: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte giornaliero, sul training set pari al 70% del data set, fornendo in input i parametri più rilevanti secondo il Pearson Correlation.

FEATURE SELECTION: RECURSIVE FEATURE ELIMINATION (SVM) TRAINING SET: ~70% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
rain	0.556	0.571	0.562	0.581	0.559	0.590
rain + anno	0.560	0.567	0.571	0.598	0.558	0.598
rain + anno + WS	0.561	0.592	0.571	0.598	0.558	0.598
rain + anno + WS + Tmax	0.181	0.311	0.552	0.555	0.558	0.597

Tabella 21: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte giornaliero, sul training set pari al 70% del data set, fornendo in input i parametri più rilevanti secondo il Recursive Feature Elimination con SVM.

FEATURE SELECTION: RANDOM FOREST TRAINING SET: ~70% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
DiffTemp	0.525	0.544	0.538	0.553	0.526	0.547
DiffTemp + ore_sole	0.223	0.275	0.411	0.433	0.496	0.505
DiffTemp + ore_sole + rain	0.213	0.271	0.406	0.423	0.448	0.487
DiffTemp + ore_sole + rain + RH_med	0.130	0.228	0.467	0.483	0.403	0.477

Tabella 22: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte giornaliero, sul training set pari al 70% del data set, fornendo in input i parametri più rilevanti secondo il Random Forest.

FEATURE SELECTION: PEARSON CORRELATION TRAINING SET: ~80% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
ore_sole	0.127	0.465	0.131	0.472	0.100	0.350
ore_sole + Tmin	0.103	0.369	0.103	0.364	0.168	0.487
ore_sole + Tmin + Tmed	0.104	0.381	0.0991	0.352	0.161	0.444
ore_sole + Tmin + Tmed + Tmax	0.0941	0.345	0.0959	0.336	0.170	0.466

Tabella 23: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte giornaliero, sul training set pari al 80% del data set, fornendo in input i parametri più rilevanti secondo il Pearson Correlation.

FEATURE SELECTION: RECURSIVE FEATURE ELIMINATION (SVM) TRAINING SET: ~80% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
ore_sole	0.127	0.465	0.131	0.472	0.100	0.350
ore_sole + rain	0.101	0.344	0.107	0.379	0.102	0.346
ore_sole + rain + WS	0.100	0.328	0.105	0.383	0.107	0.376
ore_sole + rain + WS + Tmin	0.0956	0.320	0.102	0.365	0.134	0.334

Tabella 24: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte giornaliero, sul training set pari al 80% del data set, fornendo in input i parametri più rilevanti secondo il Recursive Feature Elimination con SVM.

FEATURE SELECTION: RANDOM FOREST TRAINING SET: ~80% DEL DATA SET						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
ore_sole	0.127	0.465	0.131	0.472	0.100	0.350
ore_sole + Tmax	0.103	0.379	0.101	0.355	0.170	0.497
ore_sole + Tmax + giorno dell'anno	0.110	0.410	0.153	0.546	0.147	0.516
ore_sole + Tmax + giorno dell'anno+ RH_med	0.0928	0.327	0.155	0.569	0.117	0.340

Tabella 25: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte giornaliero, sul training set pari al 80% del data set, fornendo in input i parametri più rilevanti secondo il Random Forest.

2.3 Orizzonte mensile

Il processo di predizione della radiazione solare mensile è descritto negli esempi successivi (figura 17):

- Prediction r_inc mensile: processo nel quale sono contenute tutte le componenti che si rivelano utili per il filtraggio dei dati su determinate stazioni e in determinati intervalli di tempo, per determinare parametri

aggiuntivi e per applicare algoritmi di Machine Learning per predire la radiazione solare.

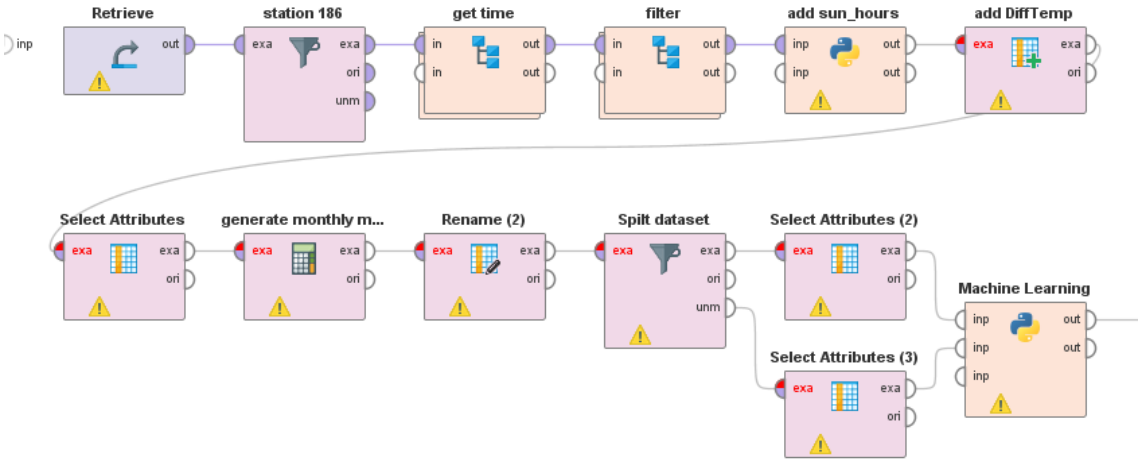


Figura 17: Processo di predizione della radiazione solare mediante modelli predittivi, per l'orizzonte mensile, mediante RapidMiner.

- Retrieve: questo operatore permette di accedere e recuperare le rilevazioni orarie dei parametri meteorologici e climatici memorizzati nel database Sysman.
- Station 186: permette di filtrare le rilevazioni orarie mantenendo solamente le rilevazioni riguardanti la stazione 186 (zona di Molfetta)
- Get time: sotto processo che si occupa di scomporre la data delle rilevazione orarie in orario, giorno, mese e anno delle rilevazioni.
- Filter: sotto processo che si occupa di selezionare le rilevazioni degli anni 2017 e 2018, in quanto in tale intervallo non sono presenti rilevazioni errate o mancanti.
- Add sun hours: script in Python che permette di determinare e aggiungere alle rilevazioni le ore di sole in una specifica giornata dell'anno ed in un determinato luogo del pianeta attraverso la latitudine, la longitudine, il giorno, il mese e l'anno presenti nel database Sysman.

- Add DiffTemp: operatore che permette di generare un nuovo attributo, specificando il nome che si vuole dare e la formula che si vuole applicare per determinare i valori di esso, utilizzando i valori degli attributi già presenti. È utilizzato per determinare la differenza di temperatura, parametro non presente nel database Sysman, determinando la differenza tra la temperatura massima e minima.
- Select Attributes: permette di selezionare solo i parametri di interesse, ovvero: la temperatura massima, la temperatura media, la temperatura minima, la pressione, la velocitò del vento, la radiazione solare, l'umidità, la differenza di temperatura, le ore di sole, il giorno dell'anno, l'anno, l'orario e le precipitazioni.
- Generate monthly mean: operatore che permette di effettuare funzioni di aggregazione. In questo caso, sono determinate le medie mensili dei singoli parametri, raggruppando per mese ed anno.
- Rename: operatore che permette di modificare i nomi degli attributi.
- Split dataset: operatore che si occupa di partizionare le rilevazioni del dataset set in due partizioni, secondo determinati intervalli di tempo, così che una partizione sarà il Training set e l'altra il Testing set. Per la precisione, il Training set sarà composto dalle rilevazioni che vanno da Luglio 2017 a Dicembre 2017 ed il Testing set dal Gennaio 2018 a Febbraio 2018.
- Select Attributes 2: permette di selezionare solo i parametri di interesse per il Training set da voler fornire in input al modello di predizione.
- Select Attributes 3: permette di selezionare solo i parametri di interesse per il Testing set da voler fornire in input al modello di predizione.
- Machine Learning: script in Python che applica i tre algoritmi di Machine Learning (Multi-Layer Perceptron, Support Vector Machine e

Long Short-Term Memory), per predire le radiazioni solari e determinare le accuratezze di esse.

2.3.1 Risultati

FEATURE SELECTION: PEARSON CORRELATION						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
ore_sole	0.0655	0.156	0.0537	0.159	0.0348	0.112
ore_sole + DiffTemp	0.01	0.0338	0.0532	0.158	0.0506	0.160
ore_sole + DiffTemp + Tmax	0.0002	0.00062	0.0002	0.00067	0.0348	0.115
ore_sole + DiffTemp + Tmax + Tmin	0.000808	0.00263	0.00026	0.0008	0.0408	0.115

Tabella 26: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte mensile, fornendo in input i parametri più rilevanti secondo il Pearson Correlation.

FEATURE SELECTION: RECURSIVE FEATURE ELIMINATION (SVM)						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
RH_med	0.0224	0.0758	0.0203	0.0687	0.0255	0.0731
RH_med + Tmax	0.000899	0.003	0.00758	0.0256	0.0266	0.0789
RH_med + Tmax + Tmed	0.000874	0.00264	0.00479	0.0128	0.0273	0.0850
RH_med + Tmax + Tmed + Tmin	0.000441	0.00144	0.00287	0.00949	0.00832	0.0266

Tabella 27: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte mensile, fornendo in input i parametri più rilevanti secondo il Recursive Feature Elimination con SVM.

FEATURE SELECTION: RANDOM FOREST						
PARAMETRI IN INPUT	MLP		SVM		LSTM	
	RMSE	ER	RMSE	ER	RMSE	ER
Tmin	0.00304	0.0982	0.00245	0.0818	0.0434	0.138
Tmin + Tmax	0.00724	0.0245	0.00197	0.0456	0.0511	0.155
Tmin + Tmax + RH_med	0.000469	0.00127	0.00444	0.0144	0.0270	0.0847

Tmin + Tmax + RH_med + DiffTemp	0.000929	0.00257	0.00474	0.0146	0.0184	0.0579

Tabella 28: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, per l'orizzonte mensile, fornendo in input i parametri più rilevanti secondo il Random Forest.

Capitolo 7: Conclusioni

Nel capitolo precedente, sono state predette le radiazioni solari, per diversi orizzonti temporali (orario, giornaliero, mensile), utilizzando tre algoritmi di Machine Learning, ovvero il Multi-Layer Perceptron, il Support Vector Machine e il Long Short-Term Memory.

L'obiettivo di questo studio era di individuare il modello predittivo e le combinazioni di parametri meteorologici, da fornire in input ai modelli, per ottenere le predizioni più accurate, per ogni orizzonte temporale.

1. Predizione della radiazione solare oraria

ACCURATEZZA DEI MODELLI PER FEATURE SELECTION UTILIZZATO							
PARTIZIONE DATA SET	FEATURE SELECTION	MLP		SVM		LSTM	
		RMSE	E.R.	RMSE	E.R.	RMSE	E.R.
~70% Training ~30% Testing	Pearson Correlation	0.90 ± 0.03	6.99 ± 1.51	0.96 ± 0.03	6.19 ± 1.70	0.99 ± 0.03	10.23 ± 2.21
	Recursive Feature Elimination (SVM)	0.95 ± 0.13	6.41 ± 1.99	1.03 ± 0.16	5.36 ± 3.50	1.04 ± 0.05	8.09 ± 3.67
	Random Forest	0.691 ± 0.21	1.87 ± 0.46	0.87 ± 0.05	1.55 ± 1.02	0.70 ± 0.16	1.60 ± 0.44
~80% Training ~20% Testing	Pearson Correlation	0.43 ± 0.02	2.56 ± 0.78	0.50 ± 0.02	0.72 ± 0.12	0.53 ± 0.06	2.24 ± 0.56
	Recursive Feature Elimination (SVM)	0.47 ± 0.06	4.73 ± 3.70	0.46 ± 0.02	2.08 ± 1.34	0.50 ± 0.07	3.35 ± 0.31
	Random Forest	0.30 ± 0.09	1.52 ± 0.43	0.45 ± 0.10	0.80 ± 0.31	0.38 ± 0.11	1.54 ± 0.89

Tabella 29: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, mediante i 3 algoritmi di Machine Learning, fornendo in input i parametri più rilevanti secondo i 3 algoritmi di Feature Selection.

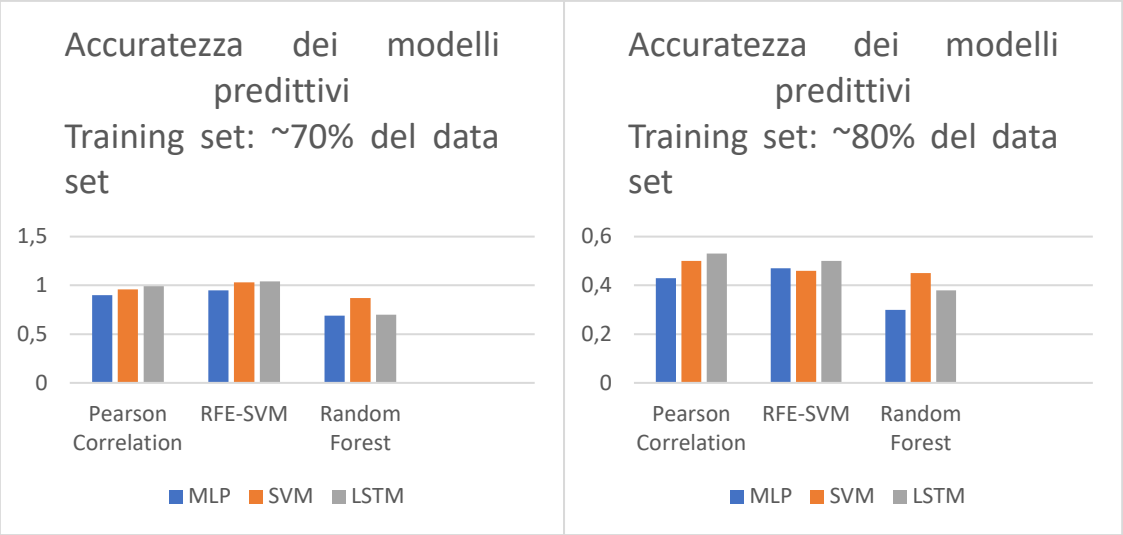


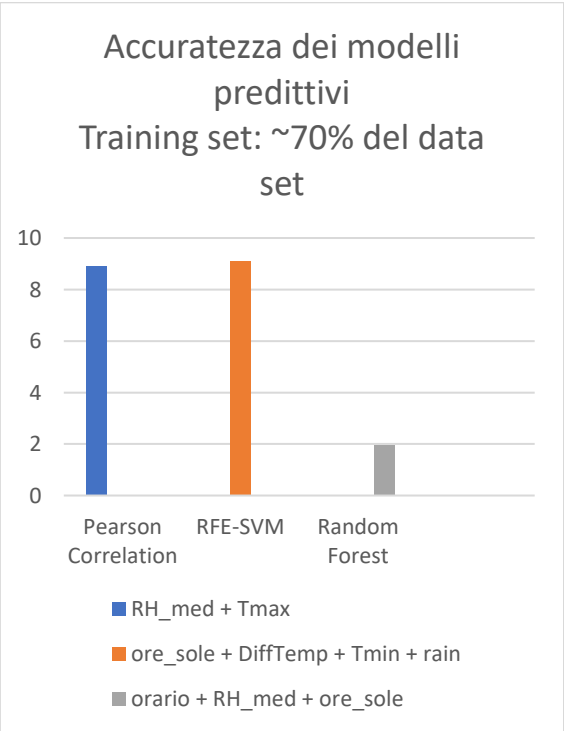
Grafico 1 e 2: Accuratezza delle predizione della radiazione solare per ogni algoritmo di Machine Learning.

Osservando la tabella 29 e 30, ed i grafici riassuntivi 1-4 delle accuratezze di ciascun modello predittivo, per la predizione della radiazione solare oraria, si evince che:

- L'algoritmo di Machine Learning che fornisce predizioni, della radiazione solare oraria, più accurate è stato il Multi-Layer Perceptron.
- Le migliori combinazioni dei parametri, da fornire in input ai modelli predittivi, per ottenere predizioni più accurate sono state individuate mediante l'algoritmo di Feature Selection: Random Forest.
- La suddivisione del data set in 80% Training set e 20% Testing set offre predizioni più accurate rispetto la suddivisione 70%/30%.
- La combinazione di parametri che permette di avere le predizioni orarie più accurate è: la temperatura minima, l'ora della rilevazione, la velocità del vento e le ore di sole.

ACCURATEZZA DEI MODELLI RISPETTO AI PARAMETRI IN INPUT				
PARTIZIONE DATA SET	FEATURE SELECTION	PARAMETRI IN INPUT	ACCURATEZZA	
			RMSE	E.R.
~70% Training ~30% Testing	Pearson Correlation	RH_med + Tmax	0.94 ± 0.07	8.92 ± 2.57
	Recursive Feature Elimination (SVM)	ore_sole + DiffTemp + Tmin + rain	0.91 ± 0.08	9.09 ± 2.25
	Random Forest	orario + RH_med + ore_sole	0.62 ± 0.18	1.96 ± 0.38
~80% Training ~20% Testing	Pearson Correlation	Tmin + Tmed + Tmax + RH_med	0.43 ± 0.04	1.79 ± 1.09
	Recursive Feature Elimination (SVM)	WS + ore_sole	0.45 ± 0.005	3.86 ± 0.87
	Random Forest	Tmin + orario + WS + ore_sole	0.28 ± 0.03	0.79 ± 0.28

Tabella 30: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, fornendo in input i parametri più rilevanti secondo i 3 algoritmi di Feature Selection.



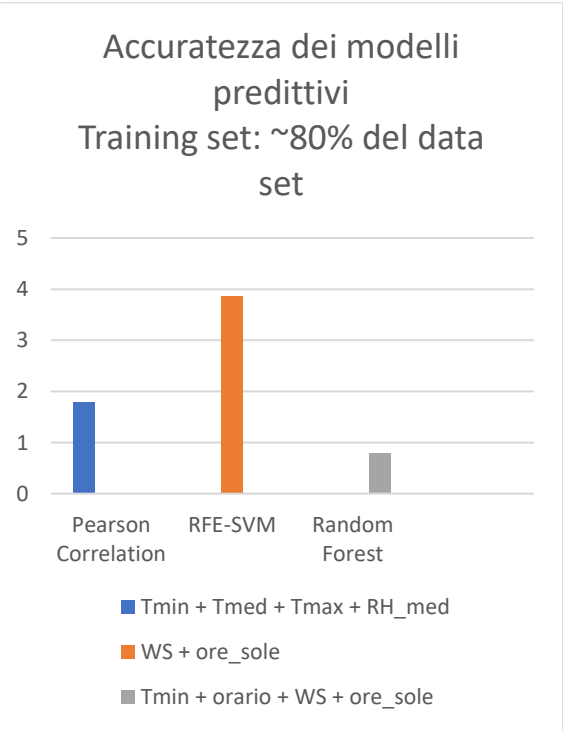
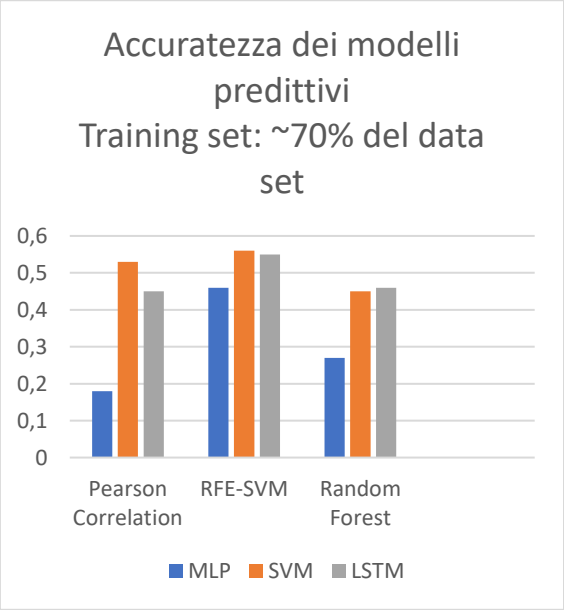


Grafico 3 e 4: Accuratezza delle predizione della radiazione solare per ogni algoritmo di Feature Selection.

2. Predizione della radiazione solare giornaliera

ACCURATEZZA DEI MODELLI PER FEATURE SELECTION UTILIZZATO							
PARTIZIONE DATA SET	FEATURE SELECTION	MLP		SVM		LSTM	
		RMSE	E.R.	RMSE	E.R.	RMSE	E.R.
~70% Training ~30% Testing	Pearson Correlation	0.18 ± 0.10	0.30 ± 0.06	0.53 ± 0.05	0.57 ± 0.06	0.45 ± 0.05	0.53 ± 0.06
	Recursive Feature Elimination (SVM)	0.46 ± 0.18	0.51 ± 0.13	0.56 ± 0.009	0.58 ± 0.02	0.55 ± 0.0005	0.59 ± 0.003
	Random Forest	0.27 ± 0.17	0.32 ± 0.14	0.45 ± 0.06	0.47 ± 0.05	0.46 ± 0.05	0.50 ± 0.03
~80% Training ~20% Testing	Pearson Correlation	0.10 ± 0.01	0.39 ± 0.05	0.10 ± 0.01	0.38 ± 0.06	0.14 ± 0.03	0.43 ± 0.06
	Recursive Feature Elimination (SVM)	0.10 ± 0.01	0.36 ± 0.06	0.11 ± 0.01	0.39 ± 0.04	0.11 ± 0.01	0.35 ± 0.01
	Random Forest	0.10 ± 0.01	0.39 ± 0.05	0.13 ± 0.02	0.48 ± 0.09	0.13 ± 0.03	0.42 ± 0.09

Tabella 31: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, mediante i 3 algoritmi di Machine Learning, fornendo in input i parametri più rilevanti secondo i 3 algoritmi di Feature Selection.



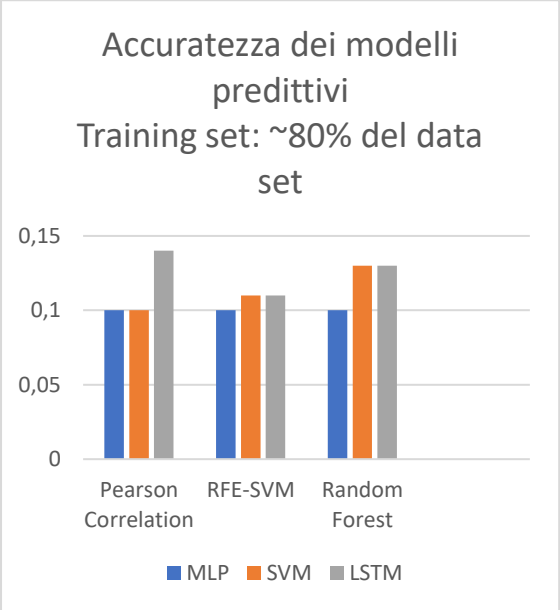


Grafico 5 e 6: Accuratezza delle predizione della radiazione solare per ogni algoritmo di Machine Learning.

Osservando la tabella 31 e 32, ed i grafici riassuntivi 5-8 delle accuratezze di ciascun modello predittivo per la predizione della radiazione solare giornaliera, si evince che:

- L'algoritmo di Machine Learning che fornisce predizioni della radiazione solare giornaliera più accurate è stato il Multi-Layer Perceptron.
- Le migliori combinazioni dei parametri, da fornire in input ai modelli predittivi, per ottenere predizioni più accurate sono state individuate mediante l'algoritmo di Feature Selection: Random Forest.
- La suddivisione del data set in 80% Training set e 20% Testing set offre predizioni più accurate rispetto la suddivisione 70%/30%.
- La combinazione di parametri che permette di avere le predizioni giornaliere più accurate è: le ore di sole, la temperatura massima, il giorno dell'anno e l'umidità relativa.

ACCURATEZZA DEI MODELLI RISPETTO AI PARAMETRI IN INPUT				
PARTIZIONE DATA SET	FEATURE SELECTION	PARAMETRI IN INPUT	ACCURATEZZA	
			RMSE	E.R.
~70% Training ~30% Testing	Pearson Correlation	ore_sole + giorno dell'anno + Tmin	0.36 ± 0.21	0.45 ± 0.15
	Recursive Feature Elimination (SVM)	rain + anno + WS + Tmax	0.43 ± 0.21	0.48 ± 0.15
	Random Forest	DiffTemp + ore_sole + rain + RH_med	0.33 ± 0.17	0.39 ± 0.14
~80% Training ~20% Testing	Pearson Correlation	ore_sole + Tmin + Tmed + Tmax	0.12 ± 0.04	0.42 ± 0.07
	Recursive Feature Elimination (SVM)	ore_sole + rain + WS + Tmin	0.13 ± 0.02	0.45 ± 0.02
	Random Forest	ore_sole + Tmax + giorno dell'anno+ RH_med	0.12 ± 0.03	0.41 ± 0.13

Tabella 32: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, fornendo in input i parametri più rilevanti secondo i 3 algoritmi di Feature Selection.

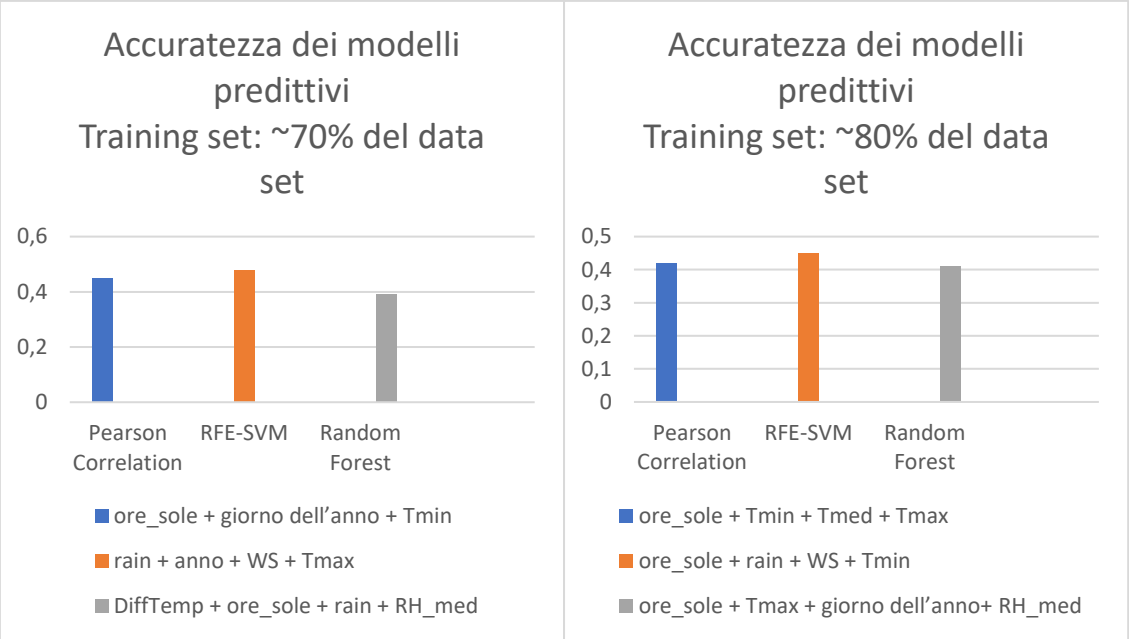


Grafico 7 e 8: Accuratezza delle predizione della radiazione solare per ogni algoritmo di Feature Selection.

3. Predizione della radiazione solare mensile

ACCURATEZZA DEI MODELLI PER FEATURE SELECTION UTILIZZATO						
FEATURE SELECTION	MLP		SVM		LSTM	
	RMSE	E.R.	RMSE	E.R.	RMSE	E.R.
Pearson Correlation	0.01 ± 0.03	0.04 ± 0.07	0.02 ± 0.03	0.07 ± 0.09	0.04 ± 0.007	0.12 ± 0.02
Recursive Feature Elimination (SVM)	0.006 ± 0.01	0.02 ± 0.03	0.008 ± 0.007	0.02 ± 0.02	0.02 ± 0.009	0.06 ± 0.02
Random Forest	0.002 ± 0.003	0.03 ± 0.04	0.003 ± 0.001	0.04 ± 0.03	0.03 ± 0.01	0.10 ± 0.04

Tabella 33: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, mediante i 3 algoritmi di Machine Learning, fornendo in input i parametri più rilevanti secondo i 3 algoritmi di Feature Selection.

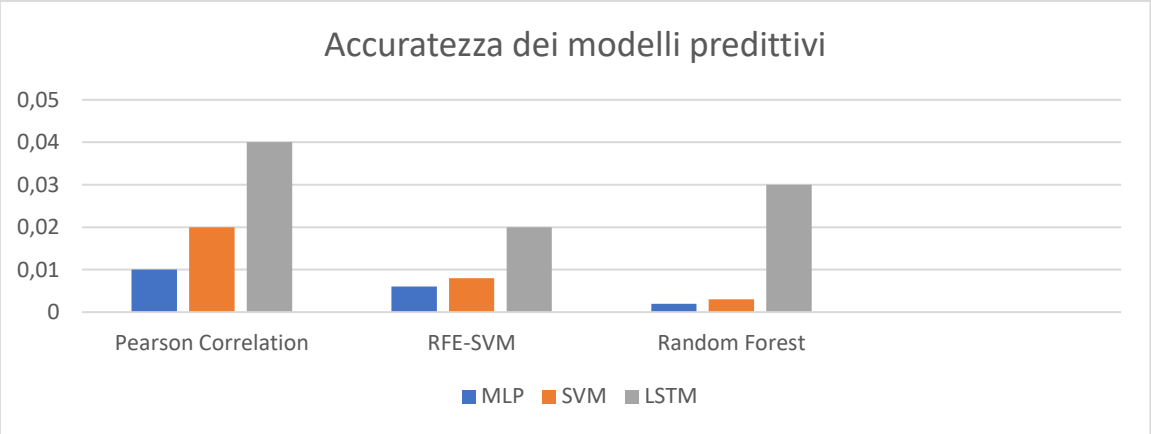


Grafico 9 e 10: Accuratezza delle predizione della radiazione solare per ogni algoritmo di Machine Learning.

Osservando la tabella 33 e 34, ed i grafici riassuntivi 9-12 delle accuratezze di ciascun modello predittivo per la predizione della radiazione solare mensile, si evince che:

- L'algoritmo di Machine Learning che fornisce predizioni della radiazione solare mensile più accurate è stato il Multi-Layer Perceptron.
- Le migliori combinazioni dei parametri, da fornire in input ai modelli predittivi, per ottenere predizioni più accurate sono state individuate mediante l'algoritmo di Feature Selection: Random Forest.
- La combinazione di parametri che permette di avere le predizioni giornaliere più accurate è: la temperatura minima, la temperatura massima, l'umidità relativa e la differenza di temperatura.

ACCURATEZZA DEI MODELLI RISPETTO AI PARAMETRI IN INPUT			
FEATURE SELECTION	PARAMETRI IN INPUT	ACCURATEZZA	
		RMSE	E.R.
Pearson Correlation	ore_sole + DiffTemp + Tmax	0.01 ± 0.02	0.03 ± 0.06
Recursive Feature Elimination (SVM)	RH_med + Tmax + Tmed + Tmin	0.003 ± 0.004	0.01 ± 0.01
Random Forest	Tmin + Tmax + RH_med + DiffTemp	0.002 ± 0.009	0.01 ± 0.02

Tabella 34: Root Mean Square Error e Relative Error delle predizioni della radiazione solare, fornendo in input i parametri più rilevanti secondo i 3 algoritmi di Feature Selection.

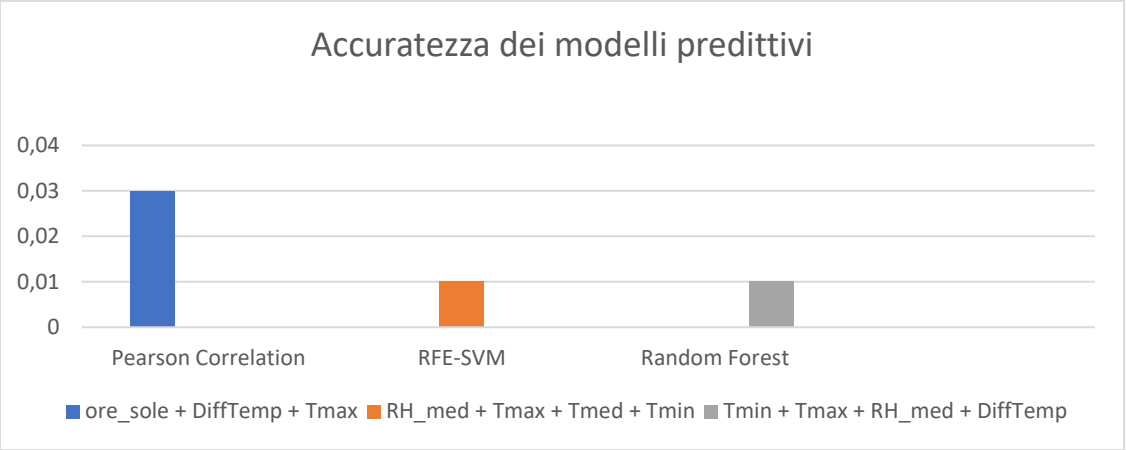


Grafico 11 e 12: Accuratezza delle predizione della radiazione solare per ogni algoritmo di Feature Selection.

4. Sviluppi futuri

In questo studio, per motivi di tempo e di risorse limitate, è stata predetta la radiazione solare per 3 orizzonti temporali, mediante 3 algoritmi di Machine Learning.
Inoltre, sono stati utilizzati 3 algoritmi di Feature Selection, per individuare le caratteristiche più rilevanti da fornire in input ai modelli predittivi.

Un possibile futuro sviluppo dello studio potrebbe essere quello di:

- Utilizzare modelli predittivi diversi da quelli già utilizzati
- Utilizzare algoritmi di Feature Selection diversi da quelli già utilizzati
- Determinare le predizioni per un orizzonte temporale annuale
- Predire la radiazione solare per le stazioni meteorologiche adiacenti
- Predire la temperatura, la velocità del vento e l'umidità relativa
- Determinare l'evapotraspirazione, mediante l'equazione di Penman-Monteith, utilizzando i valori predetti dei parametri meteorologici richiesti
- Utilizzare tecniche diverse da quelle già utilizzate per suddividere il data set in training set e testing set

Bibliografia

[1] Podetti D., “L’azienda agricola”, 2017.
[2] Caratteristiche tipologiche delle aziende agricole – 6° Censimento Generale dell’Agricoltura, Istat.
[3] “Classificazione delle attività agricole”, Ateco, 2010.
[5] Muhuddin Rajin Anwar, De Li Liu, Ian Macadam, Georgina Kelly. “Adapting agriculture to climate change: a review”. 2013
[6] Gerrit Hoogenboom. “Contribution of agrometeorology to the simulation of crop production and its applications”. 2000
[7] Roger Martin-Clouair. “Modelling Operational Decision-Making in Agriculture”. 2017.
[8] S. Fountas, D. Wulfsohn, B.S. Blackmore, H.L. Jacobsen, S.M. Pedersen. ” A model of decision-making and information flows for information-intensive agriculture”. 2006.
[9] Minwoo Ryu, Jaeseok Yun, Ting Miao, Il-Yeup Ahn, Sung-Chan Choi, Jaeho Kim. “Design and Implementation of a Connected Farm for Smart Farming System”. 2015
[11] Andreas Kamilaris, Feng Gao, Francesc X. Prenafeta-Boldu´ and Muhammad Intizar Ali. “Agri-IoT: A Semantic Framework for Internet of Things-enabled Smart Farming Applications”. 2016
[13] University of Nebraska. “Irrigation and Nitrogen Management”. 2008
[14] Richard G. Allen, Luis S. Pereira, Dirk Raes, Martin Smith. “Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56”. 1998
[15] S. L. Bithell and S. Smith. “The Method for Estimating Crop Irrigation Volumes for the Tindall Limestone Aquifer, Katherine, Water Allocation Plan”. 2011
[16] FAO (Food and Agriculture Organization of the United Nations). “Chapter 4 - Determination of ETo”.
[17] Silvia Nunnari. “Modeling solar radiation and wind speed time series for renewable energy applications”. 2015
[18] The Royal Society. “The Internet of Things: opportunities and threats”. 2017
[20] Seref Sagiroglu and Duygu Sinanc. “Bug Data: A Review”. 2013
[21] Carlos Costa and Maribel Yasmina Santos. “Big Data: State-of-the-art Concepts, Techniques, Technologies, Modeling Approaches and Research Challenges”. 2017
[22] Qi Zhang, Lu Cheng, Raouf Boutaba. “Cloud computing: state-of-the-art and research challenges”. 2010
[23] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. “A View of Cloud Computing”. 2010
[25] Sumit Das, Aritra Dey, Akash Pal, Nabamita Roy “. Applications of Artificial Intelligence in Machine Learning: Review and Prospect”. 2015
[27] Ali Rahimikhoob. “Estimating global solar radiation using artificial neural network and air temperature data in a semi-arid environment”. 2010
[28] Lanre Olatomiwa, Saad Mekhilef, Shahaboddin Shamshirband, Dalibor Petkovic. “Potential of support vector regression for solar radiation prediction in Nigeria”. 2015
[29] Mohamed Benghanem, Adel Mellit. “Radial Basis Function Network-based prediction of global solar radiation data:Application for sizing of a stand-alone photovoltaic system at Al-Madinah, Saudi Arabia”. 2010
[30] Ravinesh C. Deo, Xiaohu Wen, Feng Qi. “A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset”. 2016
[31] Ahmad Alzahrania, Pourya Shamsia, Cihan Daglib, and Mehdi Ferdowsia. “Solar Irradiance Forecasting Using Deep Neural Networks”. 2017
[32] R. Meenal, A. Immanuel Selvakumar. “Assessment of SVM, empirical and ANN based solar radiationprediction models with most influencing input parameters”. 2018
[33] S. Belaid, A. Mellit. “Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate”. 2016
[34] Kasra Mohammadi, Shahaboddin Shamshirband, Chong Wen Tong, Muhammad Arif, Dalibor Petkovic´, Sudheer Ch. “A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation”. 2015
[35] Victor H. Queja, Javier Almorox, Javier A. Arnaldo, Laurel Saito. “ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment”. 2017
[36] Ji-Long Chen, Hong-Bin Liu, Wei Wu, De-Ti Xie. “Estimation of monthly solar radiation from measured temperatures using support vector machines e A case study”. 2011

[37] M. Benghanem, A. Mellit, S.N. Alamri. “ANN-based modelling and estimation of daily global solar radiation data: A case study”. 2009
[38] Yann LeCun, Yoshua Bengio & Geoffrey Hinton. “Deep learning”. 2015
[39] Li Deng e Dong Yu. “Deep Learning: Methods and Applications”. 2014
[42] Mehdi Khashei, Mehdi Bijari. “An artificial neural network (p, d, q) model for timeseries forecasting”. 2010
[43] Kishan Maladkar. “6 Types of Artificial Neural Networks Currently Being Used in Machine Learning”. 2018
[44] Atsushi Yona, Tomonobu Senjyu, Ahmed Yousuf Saber, Toshihisa Funabashi, Hideomi Sekine, and Chul-Hwan Kim. “Application of Neural Network to One-Day-Ahead 24 hours Generating Power Forecasting for Photovoltaic System”. 2007
[45] Nesreen K. Ahmed, Amir F. Atiya, Neamat El Gayar, Hisham El-Shishiny. “An Empirical Comparison of Machine Learning Models for Time Series Forecasting”. 2014
[47] Girish Chandrashekar, Ferat Sahin. “A survey on feature selection methods”. 2013
[48] Pablo M. Granitto, Cesare Furlanello, Franco Biasioli, Flavia Gasperi. “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products”. 2006
[49] Giacomo Da Broi. “Modelli di reti neurali: multilayer perceptron e radial basis function”. 2012
[50] Dario Pavillo. “Riconoscimento real-time di gesture tramite tecniche di machine learning”. 2016
[51] Steve R. Gunn. “Support Vector Machines for Classification and Regression”. 1998
[52] Alex J. Smola And Bernhard Scholkopf. “A tutorial on support vector regression”. 2003
[53] A. Hapfelmeier, K. Ulm. “A new variable selection approach using Random Forests”. 2012
[56] David Norris, “RapidMiner - a potential game changer,” Bloor Research, 2013.

Sitografia

[4] https://mahtabrasheed.wordpress.com/2012/11/14/steps-a-farmer-performs-and-what-information-is-required-at-each-step/
[12] http://agriculture.vic.gov.au/agriculture/farm-management/soil-and-water/irrigation/about-irrigation
[24] https://www.sas.com/it_it/insights/articles/big-data/artificial-intelligence-machine-learning-deep-learning-and-beyond.html
[40] https://it.wikipedia.org/wiki/Apprendimento_profondo
[46] https://en.wikipedia.org/wiki/Long_short-term_memory
[54] http://bigdata-madesimple.com/top-30-big-data-tools-data-analysis/
[55] https://www.gartner.com/reviews/market/data-science-machine-learning-platforms/compare/knime-vs-rapidminer
[57] https://www.anaconda.com/what-is-anaconda/
[58] https://astral.readthedocs.io/en/stable/index.html
[59] https://keras.io/
[60] https://it.wikipedia.org/wiki/TensorFlow

Immagini

[41] https://qph.fs.quoracdn.net/main-qimg-f9151edaa922cf3af83d324fc6280e37

Ringraziamenti

Ringrazio il mio relatore, professor Donato Impedovo, per l'aiuto sempre attento e precisissimo che ha saputo darmi, per la competenza con cui mi ha indirizzato nelle occasioni di dubbio e per la pazienza davvero biblica che ha dimostrato nei miei confronti durante la gestazione lunghissima di questa tesi. Gli sono particolarmente grato di avermi fatto affacciare sull'abisso del Machine Learning.

Alla fine di questa fase del mio percorso formativo, così lungo e tormentato, debbo esprimere la mia più sincera gratitudine ai miei genitori per il sostegno materiale ed intellettuale che mi hanno sempre assicurato in questi anni, nonché per avermi sopportato e spronato. Mi rendo conto di come cioè non debba essere stato facile.

Ringrazio inoltre l'azienda SysMan, la quale ha fornito un dataset ricco di dati sul quale è stato possibile effettuare il lavoro di tesi.

Ringrazio infine ReCaS-Bari data center per aver fornito le risorse IT necessarie per l'elaborazione dei dati, rese disponibili attraverso due progetti finanziati dal MIUR (Ministero Italiano per l'Educazione, l'Università e la Ricerca) nel progetto "PON Ricerca e Competitività 2007-2013".