Università Politecnica delle Marche

Facoltà di Ingegneria

Corso di Laurea Magistrale in Ingegneria Meccanica

Tecniche di Parametric Cost Modelling per componenti di turbomacchine

Parametric Cost Modelling techniques for turbomachinery components

Relatore:

Prof. Michele Germani

Laureando:

Marco Pizzi

Correlatore:

Ing. Marco Mandolini

Anno Accademico 2019/20

Indice

Indice Figure

Indice Tabelle

Introduzione

Le crescenti pressioni concorrenziali che caratterizzano la maggior parte dei settori, hanno portato le imprese a sviluppare nuove strategie aziendali: livelli di qualità e di servizio più elevati, nonché la personalizzazione e l'innovazione continua devono essere rese compatibili con una riduzione dei costi.

In genere, il costo di un prodotto è dato dalla somma di diversi fattori di costo (materie prime, componenti, energia, macchinari, impianti, ecc.) difficili da quantificare nelle prime fasi del ciclo di vita data la ridotta quantità di informazioni e il basso livello di definizione del progetto.

Nonostante queste difficoltà, riuscire a stimare i fattori di costo più impattanti nella fase di progettazione concettuale garantirebbe lo sviluppo di un prodotto di successo in quanto la maggior parte dei costi sostenuti nelle fasi successive sono implicitamente determinati dalle scelte fatte durante la progettazione dettagliata del nuovo prodotto.

Fatta questa breve premessa, il presente lavoro ha come obiettivo lo sviluppo di tecniche di stima dei costi applicate a componenti di turbomacchine, seguendo un approccio tradizionale di regressione lineare e approcci innovativi di machine learning tra cui: reti neurali semplici, deep learning e random forest.

In particolare, queste tecniche sono state implementate partendo da un database storico di dati relative ai dischi di compressori assiali.

Il lavoro è stato svolto in collaborazione con l'azienda Baker Hughes di Firenze e rappresenta solo un primo studio di un progetto di lungo termine che ha come obiettivo lo sviluppo di un tool di supporto ai progettisti in grado di effettuare stime di costo parametriche con un margine di errore compatibile con l'accuratezza richiesta.

Nel capitolo 1 è stata fatta una panoramica sul concetto di stima dei costi e sono stati descritti i metodi utilizzati per le analisi.

Nel capitolo 2 è stato trattato il concetto di Should Cost Analysis, una tecnica che consente all'azienda di avere una visione più approfondita dei drivers di costo dei propri prodotti, con la duplice utilità di ottenere vantaggi in fase di contrattazione con i

fornitori, ma anche, qualora la parte venga prodotta internamente, di individuare quali siano le operazioni più gravose a livello di costo che impattano sul costo totale.

Nel capitolo 3 sono stati descritti i software utilizzati: LeanCOST® che è in grado di realizzare una Should Cost Analysis e RapidMiner che permette di implementare e validare gli algoritmi. Oltre a questi due software principali sono stati utilizzati anche Excel (per la sua semplicità e per realizzare i grafici) e SPSS per l'analisi MANOVA fattoriale.

Nel capitolo 4 è stata descritta la metodologia adottata per realizzare gli studi con annessa spiegazione di come può essere utilizzato RapidMiner.

Infine, nel capitolo 5 è stata introdotta l'azienda BH con informazioni riguardo le tendenze di mercato del settore Oil & Gas e la sua storia dalla nascita fino ad oggi e sono stati applicati i metodi di stima dei costi secondo le linee guida descritte nel capitolo 5 con la conseguente analisi dei risultati.

1 Stima dei costi

La stima dei costi permette di scartare o modificare i progetti non idonei il più velocemente possibile prima che siano state investite significative risorse economiche per la loro realizzazione. Solitamente viene determinata per mezzo di tecniche di modellizzazione durante le fasi preliminari di progetto, quando i costi dettagliati non possono ancora essere forniti.

Paradosso dei costi: La fase iniziale di progettazione concettuale impiega circa il 20% del budget stanziato per lo sviluppo di un nuovo prodotto. Le scelte intraprese durante questa fase condizionano l'80% delle spese di produzione (Figura 1)[1].

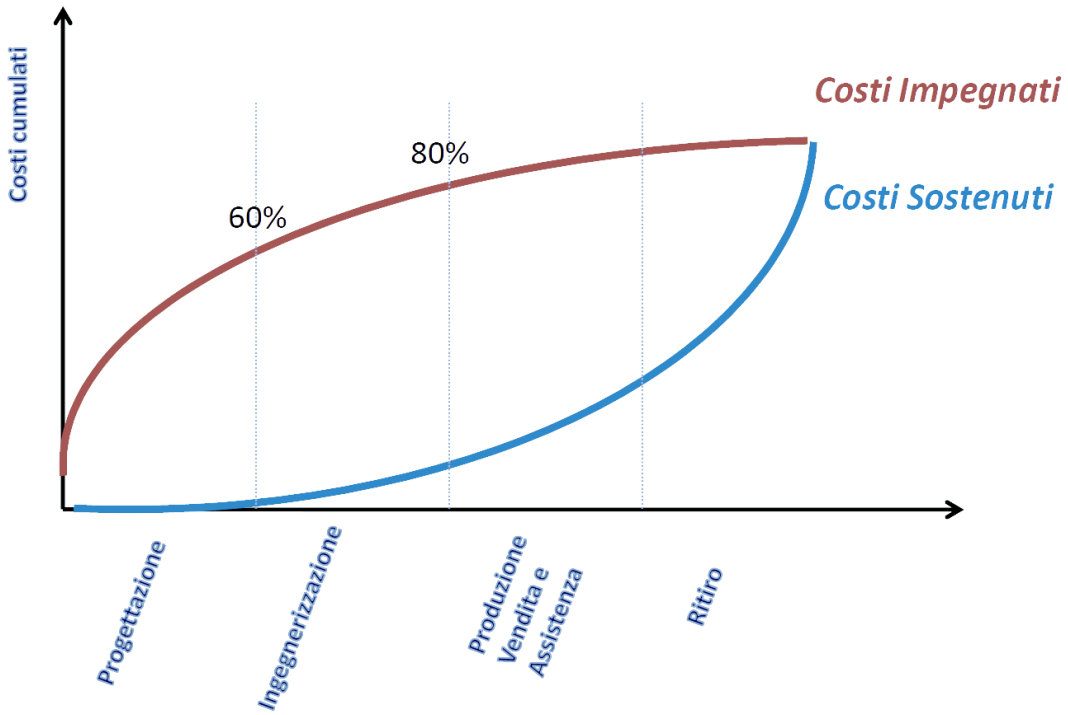

Figura 1: Paradosso dei costi

È evidente quindi che bisogna stimare ed ottimizzare i costi il più presto possibile, dove i gradi di libertà sono più elevati, poiché eventuali variazioni durante la fase di produzione potrebbero risultare molto costose.

Esistono diversi metodi per la stima dei costi e possono essere raggruppati in due famiglie principali[2]:

1. Metodi qualitativi: si basano principalmente su un'analisi comparativa di un nuovo prodotto e di uno esistente;
2. Metodi quantitativi: si basano su un'analisi dettagliata della progettazione di un prodotto, comprese le sue caratteristiche e i corrispondenti processi di produzione.

Tra i metodi esistenti per la stima dei costi, i metodi quantitativi sono la scelta più adatta per la valutazione dei costi del prodotto durante la fase di progettazione, in quanto forniscono migliori risultati in termini di affidabilità (Figura 2).

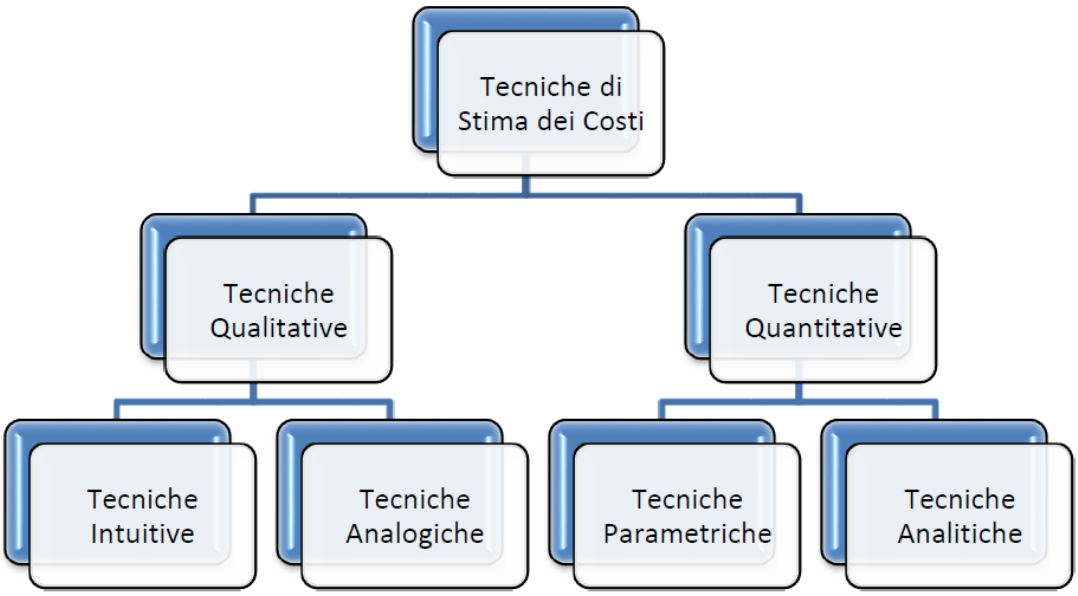

Figura 2: Metodi di stima dei costi

In questa tesi sono state utilizzate tecniche di stima parametrica e, in particolare, vi sono due approcci compatibili riguardanti la stima dei costi: approccio tradizionale e approccio innovativo. Nel primo caso il costo è stato espresso con delle funzioni analitiche lineari definite "Relazioni di stima di costo" (CER – Cost estimation relationship) e costruite per mezzo di applicazioni di metodologie statistiche. Nel

secondo caso il costo è stato stimato tramite tecniche di machine learning che possono essere reti neurali oppure foreste casuali (random forest).

Da questi metodi di predizione scaturiscono differenti modelli ed è utile metterli a confronto per un medesimo progetto per determinare quale di questi è più conveniente utilizzare; in tale modo si evidenziano sia le capacità predittive che le carenze di ciascun modello. Le stime di costo sono per loro stessa natura approssimative in quanto non consentono di determinare un valore puntuale del costo. È previsto che il valore individuato in questa fase per la valutazione della fattibilità economica possa fluttuare all'interno di un range più o meno esteso in funzione di quelle che sono le conoscenze specifiche del progetto in esame e dei riferimenti precedentemente accumulati nel corso del tempo.

Si vuole con questi sistemi valutare l'ordine di grandezza dei costi da affrontare: non essere in grado di svolgere questa valutazione significa andare incontro a gravi problemi nelle fasi di produzione di un progetto, quali la mancanza di fondi nel caso di sottostima, oppure, nel caso opposto, l'immobilizzo di capitali che potrebbero essere utilizzati per altri fini. Inoltre, è evidente che una stima di costo errata non permette di soddisfare le aspettative iniziali di progetto. Pertanto, questi sistemi e i modelli che verranno applicati sono strumenti di estrema utilità per decidere se procedere in una direzione o in una diversa nella produzione di un progetto; o addirittura se sia conveniente realizzare il progetto.

1.1 Stime di costo parametriche

Le stime parametriche dei costi (Parametric Cost Estimates) sono un approccio tradizionale di stima dei costi che determinano il costo per mezzo di relazioni statistiche tra variabili (CER). Il vantaggio principale dell'utilizzo di una metodologia parametrica è che la stima di solito può essere condotta rapidamente e può essere facilmente replicata.

1.1.1 Regressione

Il metodo principale con cui viene sviluppata la stima dei costi parametrici è la regressione. Questo metodo tenta di stabilire la natura della relazione tra le variabili fornendo un meccanismo di previsione. Ci sono due diversi tipi di variabili: dipendenti designate dal simbolo y e indipendenti designate dal simbolo x. La variabile y rappresenta un tipo di costo, mentre le variabili x rappresentano vari parametri del sistema. La regressione, quindi, è un ramo delle statistiche applicate che permette di quantificare la relazione tra la variabile dipendente e una o più variabili indipendenti e di descrive l'accuratezza di tale relazione. Sebbene possa essere non lineare, sono state analizzate sole le forme lineari di regressione (semplice e multipla). Anche se rimanere nel campo della linearità può sembrare una semplificazione eccessiva del problema, ci sono diverse buone ragioni per adottare questo approccio. I costi dovrebbero logicamente, e spesso lo fanno, variare in modo lineare con la maggior parte delle caratteristiche fisiche e prestazionali. Inoltre, molte funzioni curvilinee ed esponenziali possono essere trasformate in una forma lineare, prestandosi così all'analisi lineare[3].

1.1.1.1 Regressione lineare semplice

Si parla di regressione lineare semplice quando è necessaria una sola variabile indipendente per stimare il valore di y. La variabile y è correlata alla variabile x dalla seguente espressione:

$$y = \beta_0 + \beta_1 x \tag{1}$$

Dove:

β_0 = Intercetta

β_1 = Coefficiente angolare

X= variabile indipendente

Y=variabile dipendente

I coefficienti $\beta 1$ e $\beta 2$ possono essere calcolati graficamente tramite interpolazione (Figura 3).

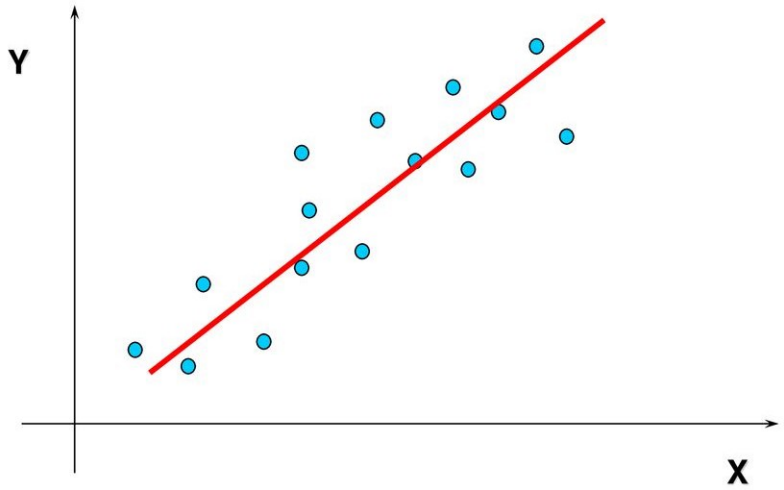

Figura 3: Regressione lineare semplice

1.1.1.2 *R*egressione lineare multipla

A differenza di quella semplice è caratterizzata da n variabili dipendenti. L'equazione assumerà la seguente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (2)$$

Dove

β_0 = Intercetta

β_1 = Coefficiente angolare di x_1

β_2 = Coefficiente angolare di x_2

X= variabile indipendente

Y= variabile dipendente

I coefficienti β_1 e β_2 possono essere calcolati matematicamente.

Sia per la regressione lineare semplice che per la regressione lineare multipla, il coefficiente di determinazione R^2 permette la misura dell'accuratezza delle equazioni. Può avere un valore compreso tra -1 e 1. In figura 4 si vede chiaramente che maggiore è il numero, migliore è l'adattamento della linea di regressione ai dati effettivi.

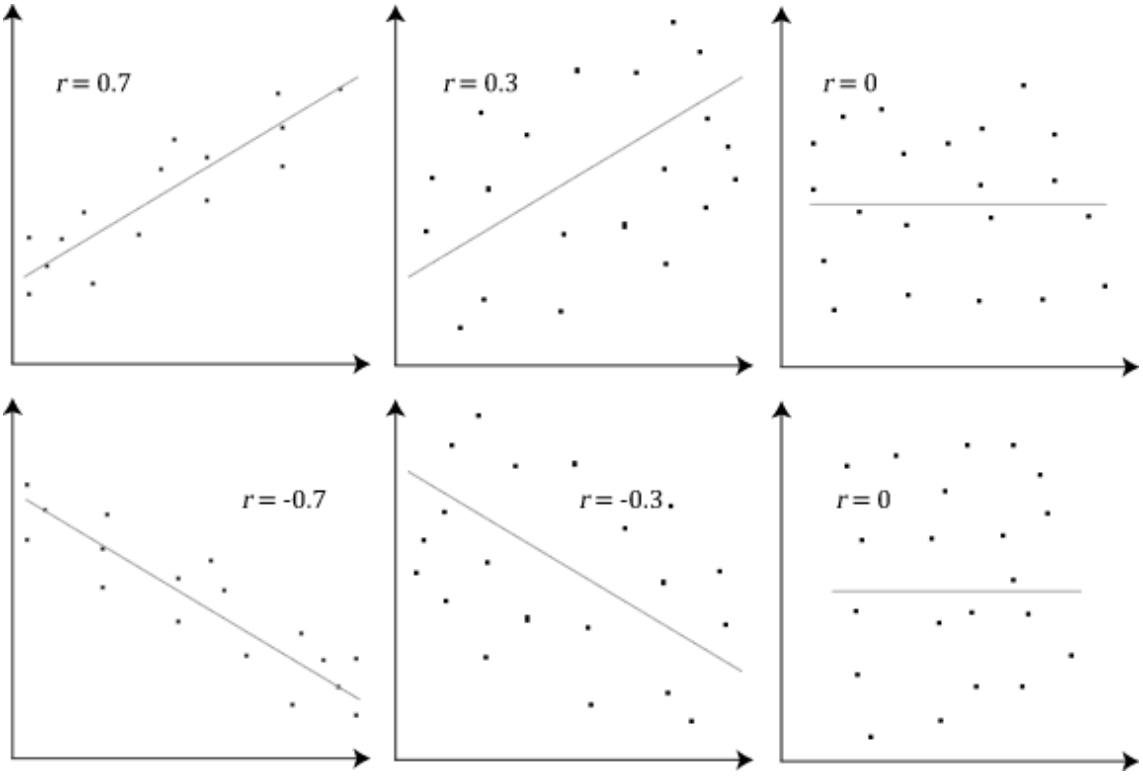

Figura 4: Andamento coefficiente R^2

1.2 Reti Neurali

Le reti neurali, definite dall'acronimo ANN (Artificial Neural Networks) sono un approccio innovativo di machine learning che permettono di sviluppare problemi computazionali complessi in maniera veloce. Esse si ispirano alle funzionalità del cervello umano ed alla sua struttura, che può essere rappresentata come una rete di neuroni interconnessi tra loro e in grado di accumulare nel tempo le conoscenze in maniera "distribuita": l'informazione viene codificata per mezzo di impulsi elettrici nei neuroni ed è immagazzinata modificando la struttura molecolare e fisica delle connessioni[4]. Analogamente, l'ANN è costituita da tante "unità" chiamate neuroni che, considerando l'architettura di rete più comune ("Multilayer Perceptron") (figura 5), sono disposti in strati successivi: ciascun neurone è tipicamente collegato a tutti i neuroni dello strato

successivo tramite connessioni pesate o sinapsi. Una connessione è un valore numerico (il "peso" appunto). È possibile individuare tre strati principali:

- Strato di input (input layer): costituito da neuroni che rappresentano la sorgente dei dati;
- Strato di output (output layer): costituito da neuroni che producono la risposta finale della rete;
- Strato intermedio (hidden layer): Può essere più di uno e contiene dei neuroni nascosti.

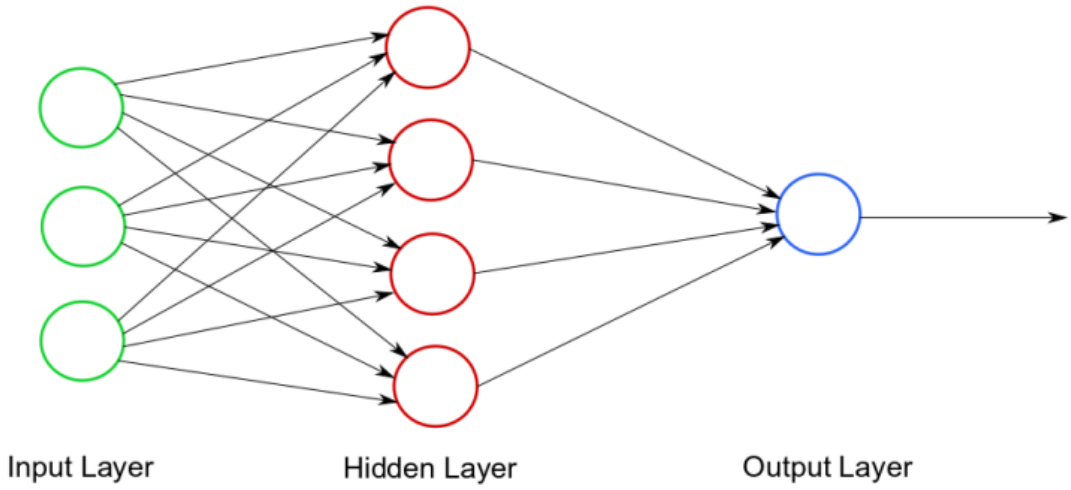

Figura 5. Architettura reti neurali

Ciascun neurone dello strato intermedio, somma i valori pesati di tutti i neuroni ad esso collegati e aggiunge un valore di bias. A questo risultato viene applicata una funzione di attivazione che trasforma matematicamente il valore prima di passarlo allo strato successivo. In questo modo i valori di input vengono propagati attraverso la rete fino a restituire l'output. Oltre alla struttura, un'altra similitudine fra la rete neurale ed il cervello umano e la capacità di apprendere. Un ANN per funzionare correttamente ha bisogno di essere allenato, cioè si richiede all'utilizzatore di immagazzinare una sempre maggior quantità di dati dettagliati e di informazioni che possano addestrare la rete neurale (tali dati sono chiamati "patterns"). Questa operazione permette di determinare la struttura definitiva della rete e quindi di determinare i valori dei parametri caratteristici (pesi e bias).

Da quanto detto, una rete può essere vista come un sistema a scatola chiusa in grado di fornire un output a fronte di un dato input.
Sono tre gli aspetti principali che descrivono e caratterizzano una ANN:

1. Architettura della rete: definita dal numero di neuroni per ogni strato, il numero di strati intermedi e le connessioni tra i neuroni;
2. Tipologia della funzione di attivazione: tipicamente non lineare (funzione sigmoidale o funzione tangente iperbolica)
3. Modalità di apprendimento: può essere di tipo supervisionato o non supervisionato. Nel primo caso i pesi vengono modificati sulla base dell'errore commesso dalla rete rispetto ai dati reali di output (ciò è possibile se si è a conoscenza di coppie di input-output). Nel caso del non supervisionato invece può essere di tipo non supervisionato dove i pesi variano nel corso dell'apprendimento in base ad una regola definita a priori che non utilizza l'errore rispetto al dato di actual.

1.2.1 Algoritmo di Backpropagation

Tra le diverse tecniche di apprendimento supervisionato la più utilizzata è l'algoritmo di Backpropagation, BP. L'obiettivo di questa tecnica è quello di determinare l'intensità delle connessioni tra i nodi e quindi i valori dei parametri caratteristici (pesi e bias). Tale algoritmo è costituito da due fasi: forward (in avanti) e backward (all'indietro). In particolare, nella fase forward i pattern di apprendimento (coppie di dati input-output) sono presentati ai nodi di input. La risposta della rete si sviluppa lungo i livelli intermedi fino ad arrivare ai nodi di output. Durante questa fase i pesi ed i bias rimangono invariati. Nella fase backward, l'errore esistente tra il processo reale e la risposta della rete viene calcolato e propagato all'indietro attraverso i nodi. Tramite opportune formule di aggiornamento i pesi ed i bias vengono modificati fino al primo livello della rete (quello di input). L'aggiornamento dei parametri caratteristici nel corso dell'addestramento può essere visto come un problema di ottimizzazione di una funzione obiettivo volta a minimizzare l'errore medio compiuto sull'insieme dei pattern di apprendimento. Un tipico problema di questa tecnica è il problema dell'overfitting (detto anche overlearning): si verifica quando la rete apprende in modo ottimo la risposta ai pattern

di input-output perdendo però la capacità di generalizzazione e di risposta a dati di input non ancora sperimentati. Il processo di apprendimento termina con una fase finale di convalida dove, tramite l'utilizzo di pattern diversi da quelli di apprendimento, si verifica se l'algoritmo fornisce risultati soddisfacenti.

1.2.2 Cross-validation

La cross-validation (o convalida incrociata) è una tecnica statistica usata nel machine learning per eliminare il problema dell'overfitting nei training-set. In particolare, la convalida incrociata cosiddetta k-fold si sviluppa nel seguente modo:

1. Si dividono i dati di input in k sottoinsiemi di dati (noti anche come fold);
2. Si addestra il modello su tutti i sottoinsiemi tranne uno (k-1) e quindi si valuta il modello sul sottoinsieme che non è stato utilizzato per l'addestramento. Ciò significa che ogni volta uno dei sottoinsiemi k viene usato come set di test, mentre gli altri sottoinsiemi k-1 sono messi insieme per formare un set di allenamento (Figura 6)[5];
3. Il processo viene ripetuto k volte, ogni volta con un diverso sottoinsieme riservato per la valutazione (ed escluso dall'addestramento).

L'esecuzione di una k-fold genera k modelli di machine learning, k origini dati di training e k origini dati di testing. Per ogni modello viene generato un parametro delle prestazioni e facendo la media si ottiene la prestazione complessiva.

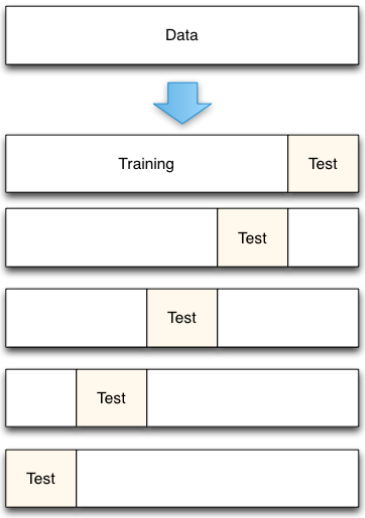

Figura 6: Cross-validation

1.3 Random Forest

Il Random Forest è un algoritmo di machine learning molto popolare per la sua semplicità, facilità d'uso e interpretabilità; è una metodologia di apprendimento supervisionato per ensemble che appartiene alla famiglia di algoritmi dell'albero delle decisioni. Un albero delle decisioni rappresenta un modello di classificazione o regressione in una struttura ad albero. Ogni nodo nella struttura ad albero rappresenta una particolare “domanda” su una caratteristica (feature), ogni ramo una decisione e ogni foglia alla fine di un ramo il valore di output corrispondente (Figura 7).

Per ottenere un risultato, partendo da uno specifico input, il processo decisionale inizia dal nodo radice (in cima) e percorre l'albero fino a raggiungere una foglia che contiene il risultato. In ogni nodo, il percorso da seguire dipende dai valori assunti dalle varie feature. Simile alle reti neurali, l'albero viene creato tramite un processo di apprendimento usando i dati di training.

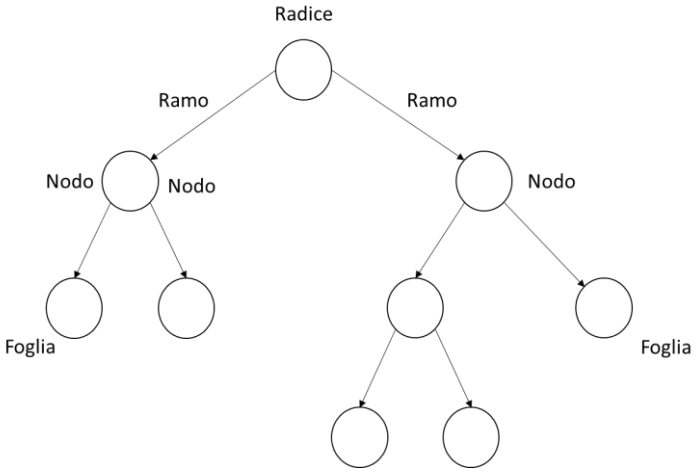

Figura 7: Albero delle decisioni

Il Random Forest è costituito da un insieme di alberi decisionali diversi tra loro ma con la stessa origine di dati di training. Ogni albero delle decisioni viene creato da un sottoinsieme diverso e casuale dei dati di training. Questa suddivisione permette di risolvere un tipico problema: se si creasse un unico albero decisionale per l’intero insieme di dati di training, si andrebbe incontro a un modello di predizione con una scarsa affidabilità dovuta alla presenza di overfitting sul training-set e di una varianza elevata.

Invece, il Random Forest è costituito da n alberi deboli ma che nel complesso costruiscono un insieme di "decision maker" in cui il risultato collettivo può essere

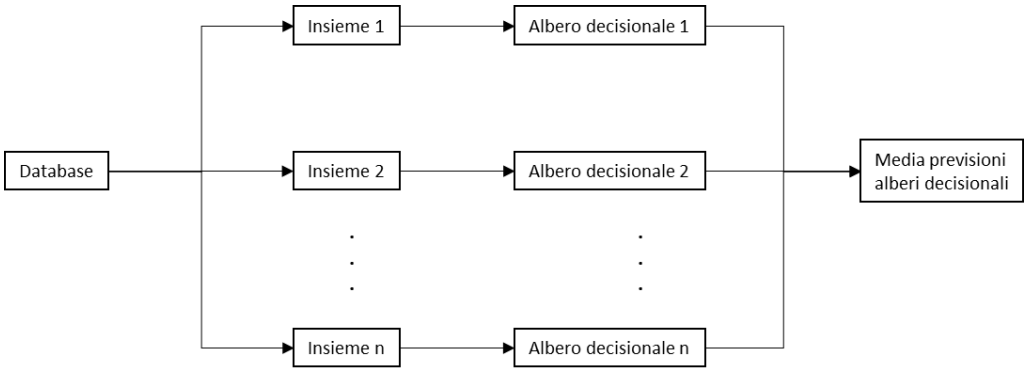

Figura 8: Random Forest

creato con il voto a maggioranza in caso di classificazione o media in caso di regressione (Figura 8).

Ciò compensa i potenziali errori dei singoli alberi in modo che si riduca la varianza e il modello sia più preciso.

1.4 Analisi dello Stato dell'arte scientifico

Le tecniche appena descritte sono state applicate negli anni in diversi contesti dimostrando enormi potenzialità. Per la loro natura, sono in grado di gestire problemi molto diversi tra loro: possono essere utilizzati per stimare i costi di progetti edilizi, autostrade o anche per progetti di componenti meccanici. Inoltre, la variabile dipendente (o target) non sempre è una voce di costo. Avvolte, si preferisce stimare una variabile di processo piuttosto che il costo (come ad esempio il tempo di lavorazione).

Per comprendere meglio quanto detto, in questo paragrafo sono stati descritti sinteticamente i risultati di alcuni studi fatti nel passato, a partire dal settore industriale.

Caputo et al. (2008) [6] confronta tra loro la regressione lineare e le reti neurali, valutando le performance nella stima dei costi del processo di produzione di recipienti a pressione di grandi dimensioni.

Questo studio è stato abbastanza significativo a causa dei relativi costi coinvolti e dell'ampia variabilità delle configurazioni e delle dimensioni dei recipienti. I metodi sviluppati sono stati testati su un periodo di sei mesi nelle strutture di uno dei principali

produttori mondiali con risultati molto incoraggianti: il metodo di funzione parametrica ha permesso di ottenere un errore di stima medio del 12,5%, con valori estremi all'interno dell'intervallo di variabilità del ±33%; la rete neurale si è rilevata ancora più performante e ha permesso di ridurre ulteriormente l'errore medio a meno del 9% con un intervallo di variabilità compreso tra il 33% e il 22%. Inoltre, il modello ANN è stato testato anche su due recipienti completamente nuovi, di diversa configurazione rispetto a quelle incluse nella banca dati, mostrando errori di stima dei costi entro il 5% dei costi effettivi.

Cavalieri et al. (2008) si concentrano nel settore dell'industria automobilistica[4]. Anche in questo caso, l'obiettivo è stato quello di illustrare i risultati di due diversi approcci, rispettivamente tecniche parametriche lineari e rete neurale artificiale, per la stima dei costi di produzione unitari di un nuovo tipo di dischi freno prodotti da un'azienda manifatturiera italiana. I modelli parametrici lineari e ANN sono stati testati e convalidati confrontando i risultati forniti da questi modelli con i costi effettivi dei venti componenti più rilevanti (dischi grezzi) acquistati o fabbricati dall'azienda. Le analisi statistiche hanno mostrato la superiorità delle reti neurali rispetto al modello parametrico lineare: l'errore di stima assume il valore massimo di circa il 15% per il modello parametrico, mentre per la rete neurale raggiunge la soglia del 10%, e solo in sei casi è superiore al 5%. I risultati sembrano confermare risultati più performanti per le reti neurali in questo campo di applicazione, ma non una netta superiorità rispetto all'approccio parametrico più "tradizionale". Il punto più rilevante riguarda la logica intrinseca dei due approcci: mentre l'uso di un modello parametrico richiede la specifica dell'espressione analitica della relazione che collega input e output, ciò non è necessario con una rete neurale. Pertanto, l'ANN è caratterizzata dalla possibilità di determinare autonomamente la forma più appropriata del rapporto, mentre un punto critico, soprattutto nel contesto applicativo specifico, è rappresentato dalla ridotta possibilità di interpretare i dati di output (che è fondamentale per l'"ottimizzazione" delle soluzioni progettuali durante il nuovo processo di sviluppo del prodotto).

Le tecniche di stima dei costi possono essere utilizzate anche nel campo civile.

Puckett si è interessato a stime parametriche dei costi per i progetti di ristrutturazione degli edifici. Sono stati presi come riferimento dati sui costi di cinquanta lavori del

sistema dell'Università dell'Alaska da sette sedi del campus[7].In questo caso il modello è stato sviluppato seguendo un approccio parametrico lineare, senza l'utilizzo e il confronto con tecniche di machine learning.

Il lavoro dimostra l'utilità di questa tecnica di stima dei costi parametrici in quanto permette di facilitare la pianificazione e determinare la fattibilità di progetto.

La rete neurale è stata utilizzata in un secondo studio in cui è stato sviluppato un modello parametrico di stima dei costi per i progetti autostradali. I dati di 18 progetti autostradali costruiti a Terranova negli ultimi cinque anni sono stati utilizzati per formare la ANN. La struttura di un semplice ANN è stata prima simulata utilizzando il programma per fogli di calcolo per fornire una rappresentazione trasparente e semplificata di questa tecnica. Sono stati poi utilizzati diversi approcci per determinare i pesi ottimali del training e quindi la struttura finale della ANN. Hegazy et al. (1998) [8]hanno dimostrato la praticità dell'utilizzo di programmi per fogli di calcolo nello sviluppo di adeguati modelli ANN da utilizzare nel campo delle costruzioni.

Čeh insieme ad altri autori [9] ha come obiettivo di questo studio analizzare le prestazioni predittive della tecnica di apprendimento random forest rispetto ai modelli basati sulla regressione multipla per la previsione dei prezzi degli appartamenti. Un set di dati che include 7407 registrazioni di transazioni di appartamenti che fanno riferimento alle vendite immobiliari dal 2008 al 2013 nella città di Lubiana, la capitale della Slovenia, è stato utilizzato per testare e confrontare le prestazioni predittive di entrambi i modelli.

Nella procedura di modellazione sono state prese in considerazione le variabili esplicative disponibili come ad esempio le caratteristiche strutturali e di età degli appartamenti, le informazioni ambientali e di vicinato, ecc. Tutte le misure di performance hanno rivelato risultati significativamente migliori per le previsioni ottenute con il metodo forestale casuale (random forest), che conferma la prospettiva di questa tecnica di apprendimento sulla previsione del prezzo dell'appartamento.

In alcuni casi, potrebbe essere utile stimare variabili dipendenti diverse dai costi. I metodi che sono stati spiegati in questo capitolo, non si limitano a fare previsioni soltanto dei costi ma possono essere utilizzati anche per altri fini come ad esempio la stima del tempo di lavorazione di un processo meccanico.

L'ultimo caso analizzato, riporta lo sviluppo di modelli di previsione energetica negli edifici con il fine di introdurre una maggiore efficienza energetica[10].

In questo studio sono state confrontate le prestazioni della rete neurale artificiale rispetto al Random Forest, nella previsione del consumo di elettricità di un hotel a Madrid, Spagna. Nel complesso, ANN ha ottenuto risultati marginalmente migliori rispetto alla RF in quanto presenta un errore assoluto più piccolo.

È chiaro quindi che i metodi di stima sono utilizzabili in diversi ambiti e con diversi fini.

Inoltre, non è possibile individuare a priori la migliore tecnica da utilizzare tra regressione, reti neurali e random forest ma queste devono essere valutate caso per caso.

Dagli articoli appena citati, non è stato rilevato un uso frequente delle tecniche di stima di costo parametriche applicate a componenti industriali.

In questa tesi, i metodi di stima dei costi sono stati applicati in un nuovo contesto: si tratta di componenti del settore Oil &Gas caratterizzati da un ciclo di lavorazione molto complesso.

Inoltre, rispetto a precedenti studi, la costruzione dei modelli è stata fatta partendo dall'analisi di Should Cost (realizzata da strumenti di analisi di costo analitici) e sono state confrontate tra loro tecniche di stima parametrica lineare tradizionali e non lineari di machine learning.

Si rimanda al capitolo 6 per una spiegazione più dettagliata del caso studio sviluppato in questo lavoro.

2 Analisi di Should Cost

L 'analisi di Should Cost è un processo che mira a determinare il costo, sulla base di fattori oggettivi quali il costo dei materiali, i costi di produzione, gli overheads ed il guadagno del costruttore[11].

Questa analisi è utile sia quando si acquista una parte da un fornitore sia quando si produce internamente. Infatti, nel primo caso permette di ottenere un vantaggio in fase di contrattazione in quanto si conosce più o meno bene il costo effettivo sostenuto dal fornitore per la produzione della parte. Nel secondo caso, l'analisi permette al team di ingegneri di individuare quali sono le operazioni più gravose a livello di costo e, di conseguenza, concentrare le successive operazioni di re-design, sempre in linea con la fattibilità pratica della modifica e nel rispetto delle tolleranze e performance richieste, esattamente su quelle che risultano essere più impattanti.

L'analisi di Should Cost risulta quindi essere essenziale per la Supply Chain e per il team di ingegneria, specialmente durante la fase di creazione di un nuovo prodotto, ottimizzando il costo dello stesso.

Il livello di dettaglio che si può raggiungere in una analisi di Should Cost dipende dalle capacità e conoscenze di colui che costruisce il modello di costo; una elevata accuratezza richiede la conoscenza del processo produttivo per produrre la parte o le operazioni di assemblaggio, dalle prime operazioni di reperimento delle materie prime fino a requisiti di qualità che devono essere soddisfatti dal prodotto finito.

Generalmente, quest'operazione è svolta dal progettista, in quanto conosce in maniera approfondita il design della parte. Tuttavia, potrebbe non avere conoscenza alcuna di tutta quella serie di operazioni che avvengono in outsourcing, si parla ad esempio di saldature, forgiature, microfusioni etc.

Proprio in quest'area la modellazione Should Cost può aiutare ad ottenere una stima completa dei costi del prodotto.

Gli step di una analisi Should Cost sono riportati in figura 9[2].

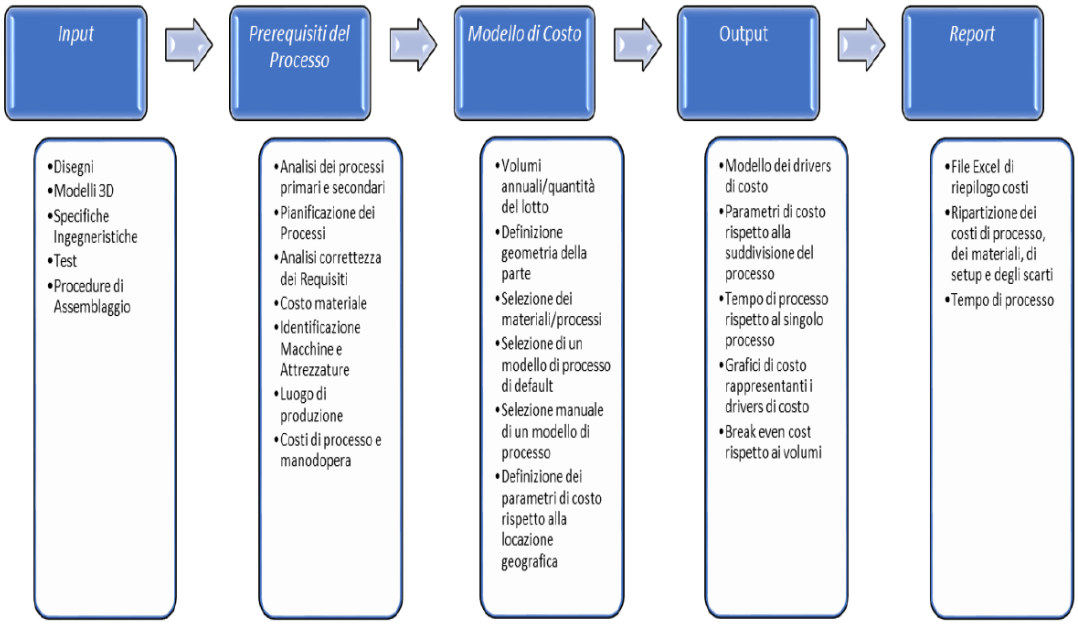

Figura 9: Analisi di should cost

Per effettuare l'analisi, si considera un prodotto come costituito da una serie di caratteristiche, ognuna delle quali ha una funzione particolare da eseguire.
Ogni prodotto è caratterizzato da:

- Caratteristiche fisiche: i materiali da cui può essere ricavato, il processo di fabbricazione, la precisione, la complessità, le dimensioni e la geometria.
- Caratteristiche economiche: è essenziale conoscere i volumi annuali e le dimensioni del lotto che, a loro volta, aiutano a prendere decisioni di "make or buy".

Il modello di costo si costruisce basandosi sulle caratteristiche del prodotto.
Il progettista può fare un trade-off basato sul comportamento dei costi delle singole funzionalità e ottimizzare il design in base alla criticità della funzionalità.
Oggi sono disponibili software di Should Cost utilizzati per analizzare il comportamento dei costi di una parte o di un assieme. La libreria del software è contrassegnata dalle variabili relative alle caratteristiche del prodotto e del processo ed è costruita sulla base di tecnologie, processi, macchinari, costi orari di manodopera in base a geografie distinte, velocità di configurazione, costi generali e tempi.

La compilazione del modello matematico in un algoritmo impostato comporterà l'osservazione delle variabili di processo che contribuiscono in modo significativo a determinare il costo del prodotto.

Questi software consentono inoltre all'utente di personalizzare materiali e parametri di processo esclusivi non inclusi nella libreria.

Esempi di software per l'analisi di Should Cost sono LeanCOST® e DFMA attraverso i quali è possibile studiare il comportamento dei costi di varie iterazioni di progettazione. Se gli utenti desiderano scegliere un materiale o un processo alternativo, il modello di costo può essere duplicato ed espresso in base alla loro area di interesse per la variazione di progettazione. I grafici statistici consentono all'utente di confrontare e analizzare il comportamento dei costi per le varie iterazioni in termini di processo, materiale, costi generali e volumi di produzione. Riassumendo, le potenzialità di un'analisi Should Cost sono le seguenti:

- Fornisce al team del sourcing una visione migliore su quali siano i margini di profitto dei fornitori
- Permette di identificare i principali driver di costo del prodotto aiutando così nella decisione di "make or buy"
- Permette di effettuare un confronto dei prezzi della parte, o dell'assemblaggio, se esternalizzati a diverse posizioni geografiche;
- Per qualsiasi approvvigionamento strategico, l'analisi dovrebbe essere uno strumento ideale per negoziare proficuamente;
- Consente di analizzare possibili alternative di materiali o processo produttivo per studiarne il costo e fare i giusti trade off;

3 Software Utilizzati

3.1 LeanCOST®

Il software LeanCOST® è stato sviluppato da Hyperlean srl.

Lo scopo di questo software è quello di semplificare tutte quelle operazioni da eseguire per determinare il costo di realizzazione di un componente meccanico o di un assieme. Pertanto, una volta che un prototipo CAD 3D è stato sviluppato, il progettista è in grado di definire una stima del costo su cui poi andrà a lavorare il tecnologo. Quest'ultimo ha la possibilità di effettuare modifiche sui parametri tecnologici relativi alle lavorazioni facendo variare le voci di costo (costo materia prima, di set-up, costi accessori e costi delle lavorazioni). L'output di quest'ultima parte è un costo effettivo che verrà poi modificato o no dall'ufficio relazioni esterne.

La struttura del programma può essere sommariamente riassunta in quattro moduli[12]:

1. Modulo di interfaccia CAD: analizza il modello CAD e le relative informazioni non geometriche per identificare le caratteristiche costruttive. Il software esegue un'analisi topologica delle entità geometriche (facce, loop, e bordi), dimensioni, finiture, tolleranze e proprietà fisiche (massa e densità);
2. Modulo di allocazione processo: converte l'insieme delle caratteristiche costruttive individuate, in una serie di operazioni per determinare il processo di fabbricazione;
3. Motore di calcolo: calcola automaticamente i tempi di produzione utilizzando le funzioni di calcolo relative ai processi individuati e le traduce in costi;
4. Modulo generazione di report: gestisce i dati calcolati e permette all'utente di utilizzarli in base alle proprie esigenze.

Un punto di forza di questo software è la capacità di integrarsi con i più noti sistemi CAD 3D commerciali e di ricavare autonomamente le caratteristiche geometriche del prodotto e definire i processi di lavorazione necessari.

LeanCOST® è a tutti gli effetti una piattaforma aziendale per la gestione dei costi di produzione e lo scambio di informazioni tra le diverse figure coinvolte attraverso interfacce utente dedicate. Infatti, ogni figura analizza il costo con un diverso livello di dettaglio e dispone degli strumenti necessari a supportare il proprio lavoro in modo efficiente e completo.

Nonostante le numerose potenzialità, questo software non sostituisce il lavoro eseguito dal tecnologo o dal progettista: esso è uno strumento utile per semplificare e velocizzare il lavoro e quindi per ridurre i costi senza però perdere in termini di qualità e precisione. La conoscenza e l'esperienza del tecnologo e del progettista rimangono sempre delle caratteristiche assolutamente indispensabili per la buona riuscita dell'operazione di costificazione.

La logica di funzionamento è la seguente: a partire dal modello CAD 3D e/o da specifiche progettuali, il software analizza la geometria ed estrae i parametri geometrici significativi per ottenere una definizione completa del prodotto in modo automatico. Il motore di calcolo associa quindi la descrizione geometrica ad una o più tecnologie di produzione grazie a regole che racchiudono processi e procedure aziendali (knowledge), estrae i parametri tecnologici e definisce la tecnologia grazie alle informazioni che caratterizzano il contesto specifico (macchine e materie prime). Il sistema al termine dell'analisi determina per ogni componente il tempo di produzione e il costo, suddiviso in cinque voci di costo che sono: costo della materia prima, costo dell'investimento, di set-up, degli accessori e delle operazioni.

Nonostante la procedura sia automatica, il software lascia ampio spazio alla personalizzazione di qualsiasi particolare della produzione da parte dell'utente. Ciò è reso possibile dall'alto livello di dettaglio per ciascuna voce di costo; è infatti possibile durante la fase di industrializzazione approfondire nel dettaglio tutti i contributi di costo e si può facilmente controllare e, se desiderato, modificare qualsiasi parametro geometrico e tecnologico di qualunque operazione, come ad esempio la velocità di avanzamento dell'utensile di un macchinario o qualsivoglia altra variabile di processo.

3.2 RapidMiner

RapidMiner è una piattaforma di data mining e predictive analytics che permette di effettuare analisi avanzate sui dati in maniera semplice e veloce, grazie a modalità di estrazione, trasformazione e visualizzazione dei dati che non richiedono particolari conoscenze di programmazione.

Si tratta di una piattaforma software sviluppata dalla società con lo stesso nome, che offre un ambiente integrato per l'apprendimento automatico, data mining, text mining, analisi predittiva e analisi di business. Viene utilizzato per le applicazioni aziendali e commerciali, nonché per la ricerca, l'istruzione, la formazione, la prototipazione rapida, e lo sviluppo di applicazioni e supporta tutte le fasi del processo di data mining compresa la preparazione dei dati, visualizzazione dei risultati, la validazione e l'ottimizzazione.

L'interfaccia grafica per la progettazione dei processi è costituita da:

1. Il riquadro centrale per la progettazione del processo dove vengono inseriti e collegati gli operatori per la modellizzazione;
2. Il riquadro dei parametri del processo e di ciascun operatore;
3. Il riquadro dell'elenco di tutti i possibili operatori utilizzabili, da quelli di controllo di processo, a quelli di analisi dei dati, fino ad arrivare ovviamente ad operatori di modellizzazione e validazione; Ogni operatore esegue una singola attività all'interno del processo e la porta di uscita di ciascun operatore costituisce la porta di ingresso del successivo;
4. Il riquadro dei repository dove vengono salvati i dati utilizzati nelle varie sessioni di lavoro.

RapidMiner utilizza il linguaggio XML standardizzato per i processi, modificabile da una apposita finestra.

Le funzionalità RapidMiner possono essere estese con plug-in aggiuntivi messi a disposizione tramite RapidMiner Marketplace. Il RapidMiner Marketplace fornisce una piattaforma per sviluppatori, dove è possibile creare algoritmi di analisi dati e condividerli con la comunità attraverso la pubblicazione in piattaforma.

Con il software è possibile effettuare collegamenti alle principali sorgenti dati tra cui Excel.

3.3 SPSS

SPSS (Statistical Package for Social Science) è uno strumento di analisi di statistica avanzata che offre la possibilità di utilizzare numerose funzioni, dalla pianificazione e raccolta dei dati all'analisi, alla produzione di report e alla distribuzione.
In particolare, è possibile:

- Analizzare un set di dati con il fine di risolvere problemi di business e di ricerca complessi attraverso un'interfaccia intuitiva, simile a quella di Microsoft Excel;
- Comprendere più rapidamente dataset grandi e complessi con procedure statistiche avanzate, che aiutano a garantire un'elevata precisione e un processo decisionale di qualità;
- Utilizzare le estensioni, il codice linguaggio di programmazione Python, R e anche RapidMiner per l'integrazione con software open source.

Il software SPSS offre in particolare la possibilità di realizzare analisi di varianza (ANOVA) di diverso tipo (in base al numero di variabili indipendenti):

- ANOVA a una via: modelli che prevedono una sola variabile indipendente;
- ANOVA fattoriale: modelli che prevedono due o più variabili indipendenti;
- ANOVA univariata: modelli che prevedono una sola variabile dipendente;
- ANOVA multivariata (detta anche MANOVA): modelli che prevedono due o più variabili dipendenti.

In questo lavoro il software è stato utilizzato per realizzare l’analisi MANOVA fattoriale.

4 Metodologia

L’obiettivo di questa tesi è stato definire delle linee guida per la realizzazione di un tool per la stima dei costi. Si è partiti da dei requirements generali e relativi all’interfaccia grafica, entrambi definiti dall’azienda Baker Hughes.

Tali requisiti sono riportati di seguito e sono stati richiamati nei paragrafi successivi:

R1. La stima dei costi fornita dallo strumento deve essere parametrica, basata sui driver (geometrici e non geometrici);

R2. Lo strumento deve essere in grado di estrarre equazioni parametriche da una serie di punti dati memorizzati nel suo database;

R3. Lo strumento deve essere in grado di regolare equazioni e coefficienti parametrici basandosi su dati aggiornati;

R4. Lo strumento deve essere aperto per estendere la copertura ad altre famiglie di parti. Lo strumento fornisce un'interfaccia utente per aggiungere nuove famiglie di parti, definire i loro driver, inviare la serie di riferimenti di punti dati, ecc;

R5. Lo strumento deve consentire l'estensione dei driver. Lo strumento fornisce un'interfaccia utente per aggiungere o modificare i driver per una famiglia di parti esistente inviando un set di riferimento aggiornato (espanso) di punti dati;

R6. Lo strumento deve consentire di eseguire calcoli basati su una serie multipla di punti dati (Actual Cost, Should Cost, Does Cost);

R7. Lo strumento deve eseguire una convalida delle equazioni calcolate rispetto ai punti dati;

R8. Lo strumento fornisce al costo calcolato l'accuratezza applicabile rispetto al set di dati di riferimento;

R9. L'utente può selezionare la famiglia di parti su cui lavorare;

R10. L'utente può selezionare il set di dati (Actual Cost, Should Cost, Does Cost);

R11. L'utente inserisce i driver specifici della famiglia di parti selezionata;

R12. Lo strumento genera costi di materie prime, lavorazioni meccaniche.

La figura 10 riporta lo schema della metodologia designata.

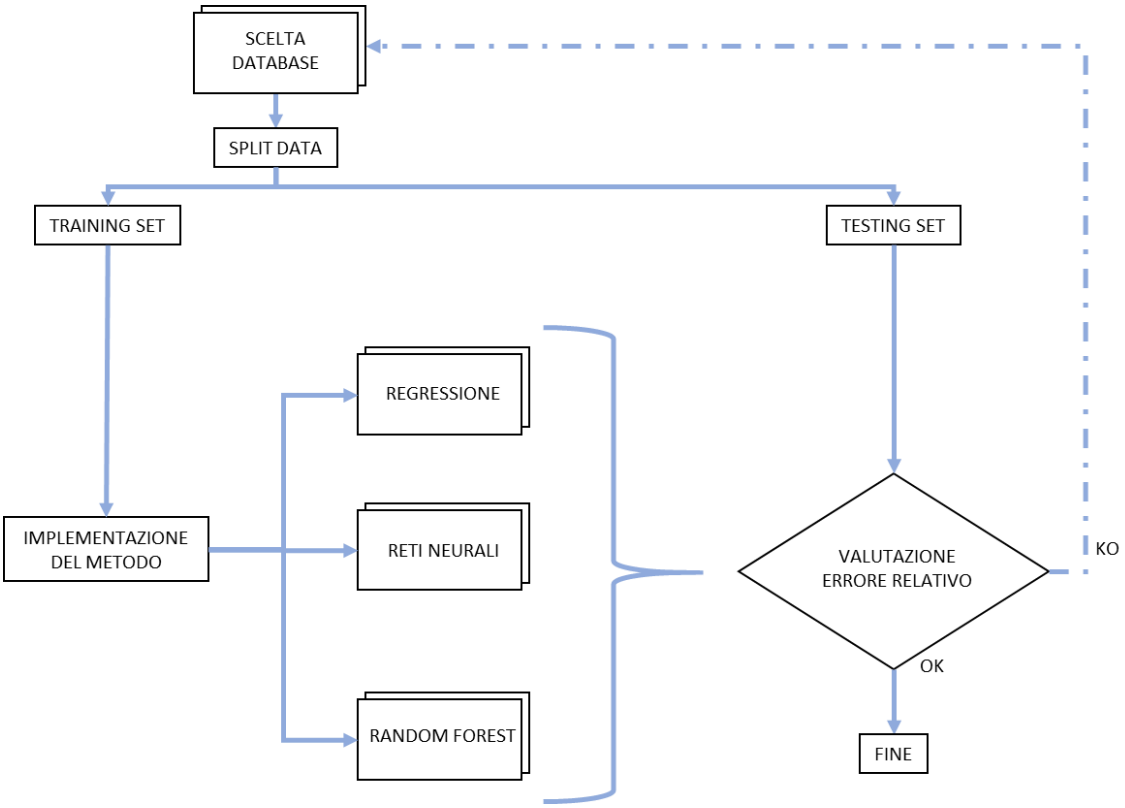

Figura 10: Metodologia

Il primo passo dello schema di figura 10 è la scelta del database: partendo da un set di dati storici, tramite strumenti di analisi si prepara il database in modo da renderlo utilizzabile dai vari metodi. Eventuali valori anomali si escludono e si considerano solo i cost drivers con una forte influenza sull'output. In quest'ultimo caso si utilizza l'analisi MANOVA fattoriale per determinare la significatività e quindi il peso di ciascun input rispetto all'output.

Successivamente, tramite lo spit data si divide il database in due gruppi: training set e testing set. I primi si utilizzando per implementare il metodo di regressione lineare tradizionale e i metodi non lineari di machine learning mentre, i secondi per validare i modelli ottenuti.

In RapidMiner per ogni metodo si genera un process costituito da quattro sub-process relativi a: pre-processing data, training model, testing model, create prediction e output (Figura 11).

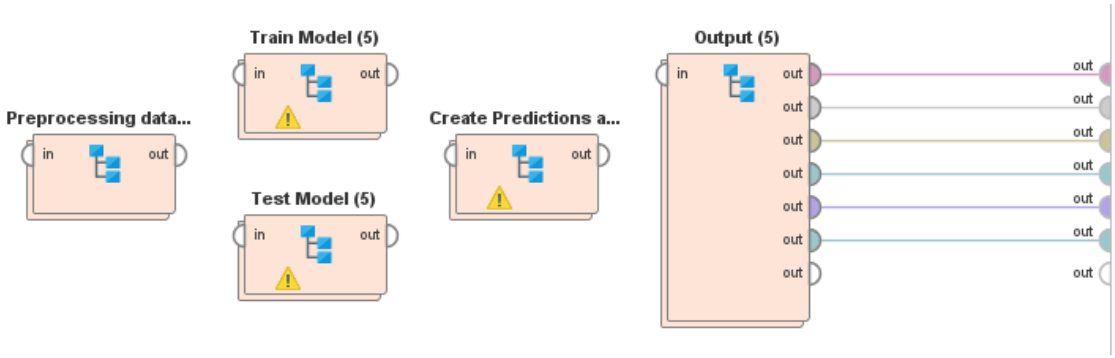

Figura 11: Sottoprocessi operatori

Il "pre-processing data" corrisponde alla scelta del database dello schema di figura 10: come input si inserisce il database iniziale e come output si ottengono i dati di training e di testing. Per collegare tra loro tutti i sotto-processi si utilizzano due operatori:

- "Remember": memorizza l'oggetto specificato nell'archivio oggetti del processo;
- "Recall": richiama l'oggetto dall'archivio.

I dati di training, memorizzati nel primo sotto-processo, si richiamano nel train model dove si generano un modello specifico (linear regression, neural network, deep learning o random forest) [requisito R2].

I dati di test si richiamano, insieme al modello, nel test model dove si valida l'algoritmo [requisito R7].

Nel "create prediction" si richiamano dati di training, di test e il modello per generare il model simulator.

Infine, nel sotto-processo output si richiamano tutti gli oggetti dell'archivio del processo e si trascrivono su un file Excel (oltre che visualizzarli nei risultati di RapidMiner).

4.1 Scelta Database

La precisione di un'analisi di stima dei costi è fortemente dipendente dalla forma del database storico in ingresso. Esso è costituito da dati di input (driver geometrici e non geometrici) [requisito R1] e da relativi dati di output (costi di materie prime e di produzione). Il primo step di questa parte consiste nel generare il database. In alcuni casi, i dati potrebbero essere disponibili in forma tabulata (Actual Cost), in altri potrebbe essere necessario effettuare una Should Cost Analysis con l'aiuto di software come

LeanCOST® o DFMA [requisiti R6 - R10]. Scegliere i giusti drivers di input non è un'operazione facile; bisogna essere in grado di individuare quei parametri che il progettista conosce in fase di progettazione e che hanno un maggiore impatto nel calcolo dei costi. Inoltre, individuare eventuali outliers ed eliminarli permette di ridurre l'errore della predizione dell'output. Quindi, prima di utilizzare i vari algoritmi, bisogna fare delle operazioni di pre-processing, tra cui:

- Analisi di varianza multivariata fattoriale (MANOVA);
- Determinare eventuali valori anomali.

L'analisi di varianza multivariata fattoriale è una tecnica statistica di analisi dati che determina se e quanto fortemente due variabili sono dipendenti tra loro. Il grado di dipendenza viene espresso dal valore di significatività che è un numero compreso tra 0 e 1; in caso di valore molto piccolo (minore di 0,1) si ha forte dipendenza. Se, invece, assume valore 1, le due variabili non hanno nessuna relazione di dipendenza tra loro.
Questa tecnica può aiutare nella scelta dei drivers con maggior "peso" ma anche l'esperienza gioca un ruolo chiave. La conoscenza dei processi di produzione permette di determinare eventuali drivers "nascosti" di processo (che derivano ad esempio da rapporti tra i parametri) ed è indispensabile per capire quali sono quei parametri noti nelle fasi iniziali della progettazione quando viene richiesta la stima dei costi.
Per quanto riguarda invece i valori anomali, in questa fase iniziale di scelta del database, può capitare nello scrivere i dati di Should Costo o di Actual Cost in un foglio elettronico, di commettere errori di battitura e di conseguenza inserire uno o più records errati all'interno del database.
Per ottenere dei modelli in grado di stimare i costi in maniera performante, è importante eliminare tutti quei valori anomali dovuti a records errati o incoerenti.
Per valori anomali non si intendono solo records errati ma anche eventuali dati duplicati o incompleti.
Una volta che il database è stato definito, segue l'operazione di split data. Come mostrato in figura 10, non tutti i dati vengono utilizzati per generare il modello ma ci sarà una restante parte dedicata a una fase di test. Sono proprio i test-set che permettono di capire come risponde un modello e quindi a validare i risultati.

In particolare, una buona percentuale di suddivisione è 60% di dati di training e 40% di dati di testing oppure 70 - 30% (molto dipende dal tipo di database, si potrebbe utilizzare anche 80 - 20%).

4.1.1 RapidMiner

Il sotto-processo “pre-processing data” contiene al suo interno diverse funzioni (Figura 12).

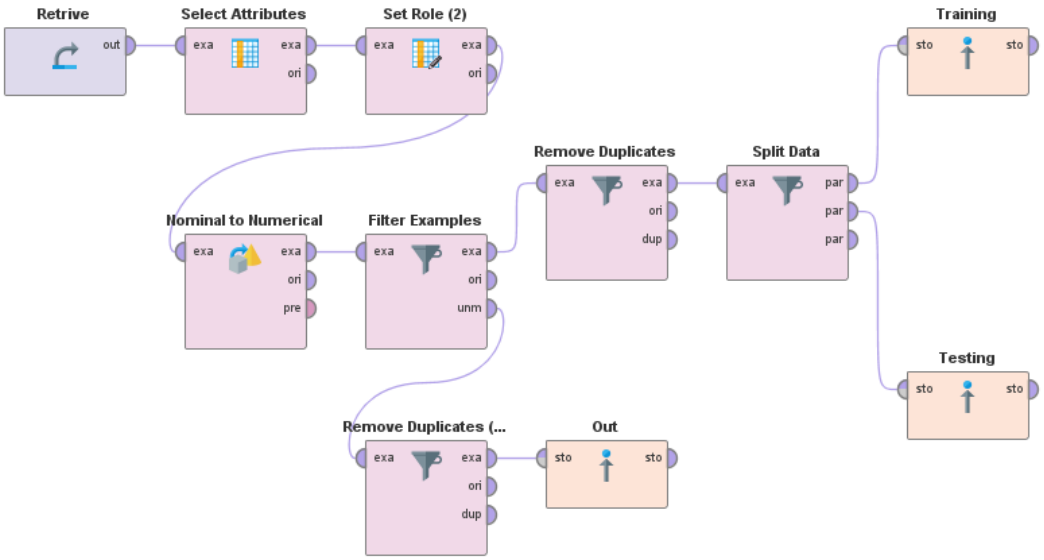

Figura 12:Pre-processing dataset

Per prima cosa, cliccando su “import data”, si può inserire il set di dati iniziali nel repository di RapidMiner: si possono inserire sia variabili numeriche che qualitative (numerical o polynominal). Tramite la funzione “retrive”, si richiama il database dal repository e lo si inserisce sub-process.

Dalla figura 12 è possibile notare che ciascun operato presenta una o più porte di ingresso e di uscita.

Per collegare due operatori tra loro, si collega una porta di uscita del primo operatore con una porta di ingresso del secondo operatore.

I collegamenti non sono casuali ma dipendono dal tipo di informazione che si vuole condividere. RapidMiner, per semplificare i collegamenti, utilizza colori diversi in base al

tipo di informazione condivisa in modo da localizzare velocemente quali sono le porte di ingresso e di uscita.

Se non specificata già nel database, la funzione "set role" consente modificare il ruolo di una variabile da regolare a label (ovvero la variabile target o di risposta).

Un limite di RapidMiner è di non poter effettuare l'analisi MANOVA fattoriale (esiste la funzione "ANOVA" ma è riferita al caso di problemi di classificazione e non di regressione). Pertanto, si può utilizzare un altro software chiamato SPSS per capire quali sono e variabili indipendenti con maggior "peso" sulle variabili dipendenti.

Tornando a RapidMiner, tramite la funzione "select attributes", si selezionano gli input da considerare e si escludono quelli irrilevanti. Dopo questo passaggio è bene utilizzare "remove duplicates" in quanto si potrebbero creare dei records duplicati che andrebbero a falsificare i risultati delle analisi.

Nel caso in cui si volesse considerare un sottoinsieme del database per un secondo test oppure semplicemente per escluderlo dall'analisi successiva, si può inserire la funzione "filter example".

Su RapidMiner ci sono alcuni modelli di predizione che utilizzano database misti (dati quantitativi e qualitativi) e altri che lavorano soltanto con dati numerici o qualitativi. In particolare, per gli algoritmi "linear regression" e "neural network" si può utilizzare la funzione "nominal to numerical" per convertire un database misto in uno numerico tramite la tecnica delle variabili dummy o binarie.

Per comprende il significato delle variabili dummy è stato fatto un esempio in tabella 1:

Materiale	**Conversione**	**Materiale = M1**	**Materiale = M2**
M1	=	1	0
M2	=	0	1
M1	=	1	0
M1	=	1	0

Tabella 1: Variabile dummy

In caso di variabile categorica "materiale" (prima colonna), con due possibili valori M1 e M2, è possibile generare un database fatto di soli dati numerici creando due nuove

variabili binarie (terza e quarta colonna) che assumono valore 1 nel caso in cui una condizione è soddisfatta, viceversa valore nullo.

Quindi, ad esempio, la riga uno corrisponde al materiale M1 in quanto la prima variabile dummy assume valore 1, mentre la seconda valore zero.

Infine, tramite "split data" si divide il database in due insiemi: training set e testing set. L'operatore "remember" consente di memorizzare i due dataset e di riutilizzarli nei successivi sub-process (per richiamarli si utilizza "recall").

4.2 Realizzazione Modelli

4.2.1 Implementazione del metodo: Regressione

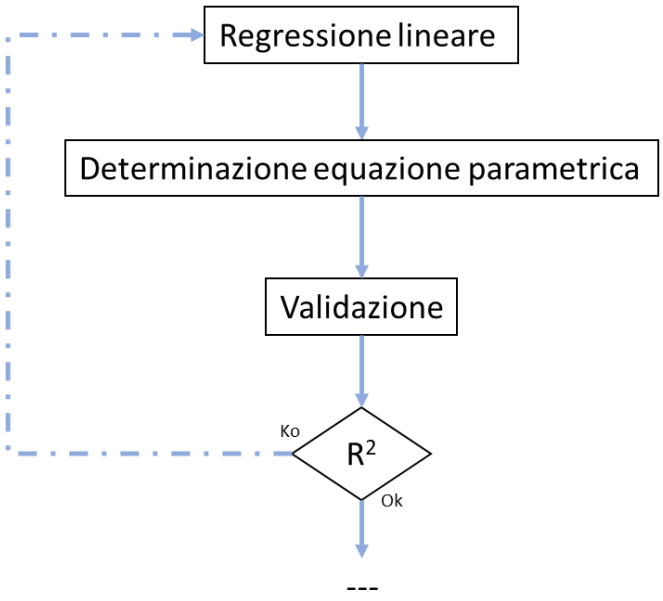

Figura 13:Metodo regressione lineare

La regressione lineare può essere sviluppata solo per variabili numeriche.

La figura 13 riporta i vari step di questo primo metodo.

Il procedimento è iterativo: partendo dai dati di training (definiti nel paragrafo precedente), si realizza la regressione lineare semplice o multipla determinando l'equazione parametrica. Segue la validazione in cui si calcola il coefficiente R^2; in caso di valore accettabile, l'analisi termina, altrimenti, si ripete il procedimento (facendo opportune modifiche al database).

R^2 può assumere i seguenti valori:

- $R^2 \geq 0.9$: l'analisi è soddisfatta;
- $0.6 < R^2 < 0.9$: bisogna fare uno studio più approfondito dei dati e identificare più fattori di costo;
- $R^2 \leq 0.6$: l'analisi non è accettabile e non può essere utilizzata per studi ulteriori.

4.2.1.1 RapidMiner

In figura 14 è riportato il sotto-processo "train model"; i dati di training si richiamano con la funzione "recall".

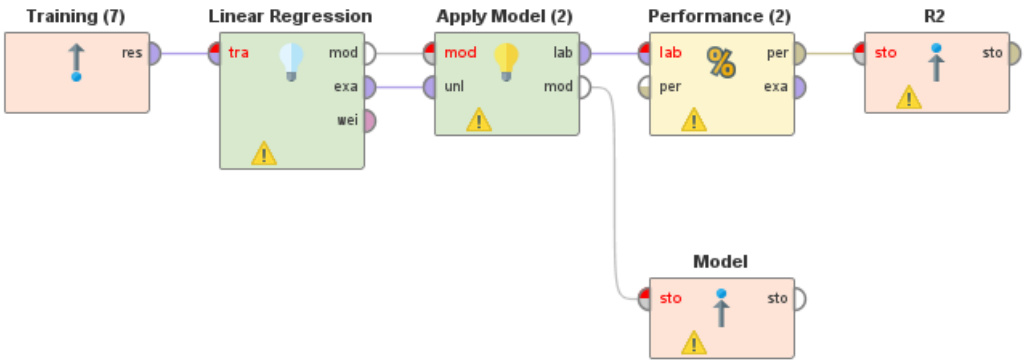

Figura 14: Train model regressione lineare

L'operatore "linear regression" sviluppa la regressione lineare (sia semplice che multipla) e restituisce l'equazione parametrica. Tramite "apply model" si applica l'equazione parametrica al set di dati di training mentre l'operatore "performance" riporta il valore del coefficiente R^2 (nella finestra "parametri" dell'operatore "performance", bisogna selezionare l'errore "squared correlation").

Oltre a questa prima validazione, il modello viene testato con un set di dati nuovi per calcolare l'errore relativo e quindi per confrontarlo con gli altri modelli di machine learning sviluppati parallelamente (vedi paragrafo 4.3). Occorre fare una precisazione: l'operatore "linear regression" (e in generale la regressione) ha il limite di poter essere utilizzato soltanto per variabili numeriche. Questo comporta due diversi modi di procedere:

1. Creare un modello tramite “linear regression” applicato all’intero training set. Eventuali variabili qualitative (polynominal) devono essere convertite in numeriche tramite la tecnica delle variabili dummy. Ad esempio, se ho materiale A e materiale B utilizzo la funzione “nominal to numerical” per generare due variabili dummy, ciascuna riferita a un materiale;
2. Creare n modelli di “linear regression” relativi a n sottoinsiemi del training set. Per determinare i sottoinsiemi si utilizza n volte l’operatore “filter example” che filtra il dataset considerando solo un insieme di dati. Sono le variabili qualitative a guidare la suddivisione dei dati. Ad esempio, se ho materiale A e materiale B, realizzo due insiemi ciascuno riferito a un materiale.

Il modello finale (o i modelli nel caso in cui si creino i sottoinsiemi) si può memorizzare tramite “remember” per poi essere richiamato nel sotto-processo output.

4.2.2 Implementazione del metodo: Rete neurale semplice e deep learning

Le reti neurali, così come altre tecniche di machine learning, possono essere più complicate da realizzare rispetto a metodi parametrici lineari ma spesso risultano più precise. A differenza della regressione sono in grado di gestire variabili numeriche e categoriali (ad eccezione della rete neurale semplice).

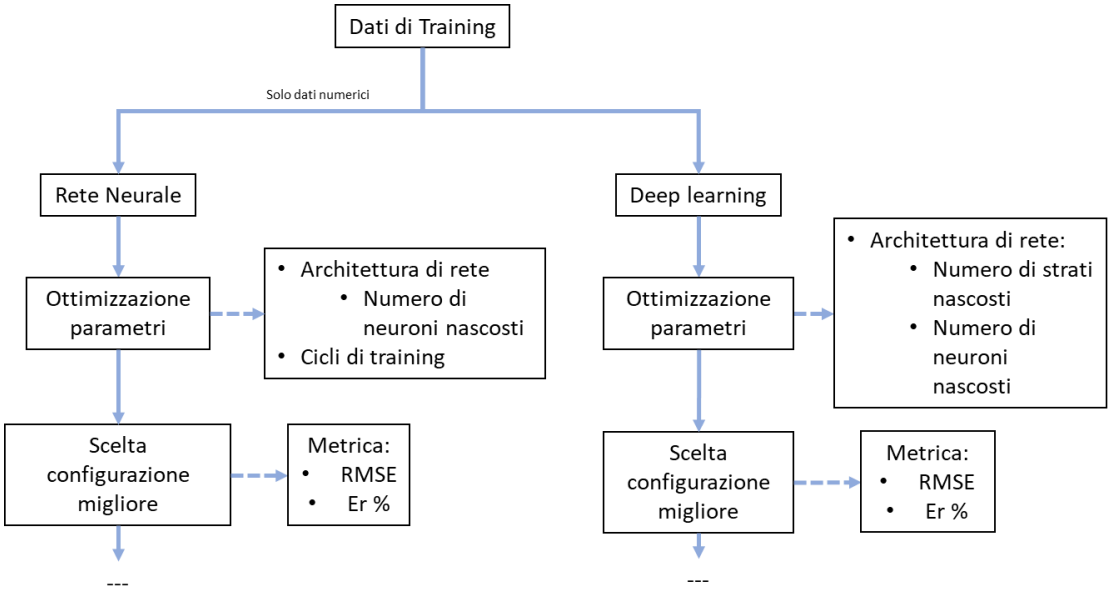

Figura 15:Metodo Reti neurali semplici e deep learning

Nella figura 15 è mostrato il metodo in maniera schematica.

Anche in questo caso il procedimento è iterativo e si ripente fino a quando non si ottiene un modello accettabile. Partendo dai dati di training, si sceglie il tipo di algoritmo da utilizzare: la rete neurale semplice è tipicamente caratterizzata da un solo strato nascosto ed è adeguata a un numero limitato di dati di input; il deep learning si basa sempre sul concetto di reti neurali ma presenta due o più strati nascosti e quindi è preferibile utilizzarlo nel caso di numerosi dati d'ingresso. Scegliere il numero di neuroni di uno strato nascosto così come scegliere il numero di strati nascosti è un'operazione tutt'altro che banale. Quando si utilizza la rete neurale semplice, una regola empirica permette di determinare il numero di neuroni nascosti dell'unico hidden layer (3):

$$n° \, neuroni \, nascosti = \frac{n° \, input + n° \, output}{2} + 1 \qquad (3)$$

Tuttavia, non è detto che l'equazione (3) restituisca l'architettura di rete migliore e inoltre, ci sono anche altri parametri da specificare come ad esempio il numero di cicli di training. È chiaro quindi che individuare il giusto modello di rete neurale semplice o di deep learning non è un'operazione immediata e, non essendoci una regola generale, bisogna fare una sorta di "esplorazione" in cui un'sistema di calcolo avvia automaticamente più esecuzioni simultanee con diversi valori dei parametri e individua la configurazione con le migliori prestazioni. Quindi, una volta individuati quei parametri

fondamentali, si attribuiscono per ciascuno di essi dei valori e si generano più modelli che derivano dalle diverse combinazioni.

Questo processo è detto ottimizzazione parametrica e per valutare quale configurazione della rete neurale semplice o del deep learning risponde meglio si utilizza come metrica l'errore relativo oppure l'RMSE (errore assoluto).

4.2.2.1 RapidMiner

Il primo passo, come mostrato in figura 16, consiste nel richiamare i dati di training.

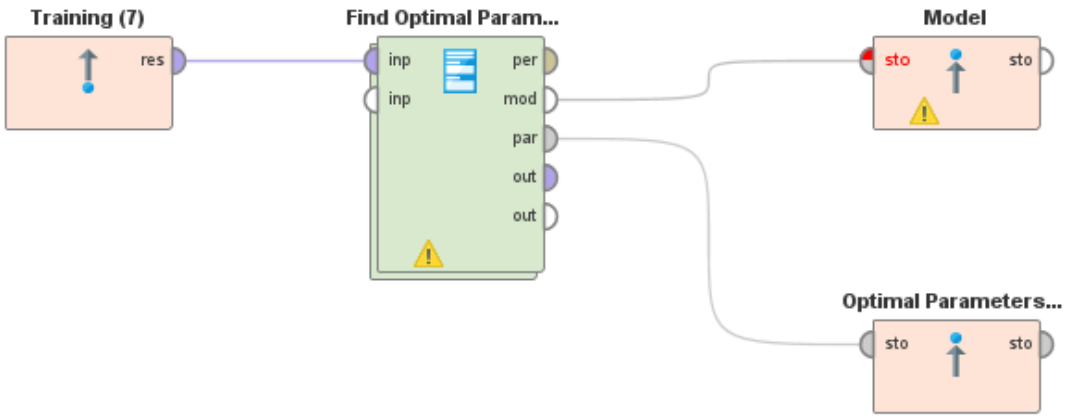

Figura 16: Train model rete neurale

L'ottimizzazione parametrica si fa con "optimize parameters (grid)": cliccando su "edit parameters setting" si selezionano i parametri da ottimizzare e si specificano i valori che possono assumere.

L'operatore "optimize parameters" è un nested operator, ciò significa che al suo interno ci sono ulteriori operatori (Figura 17).

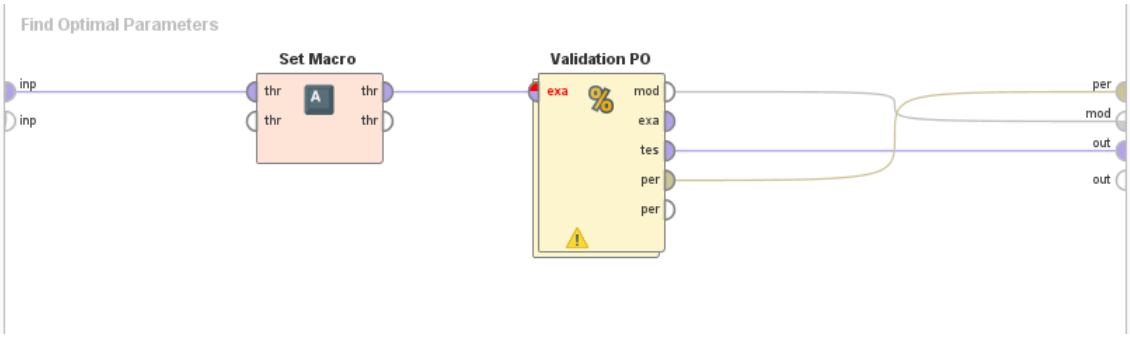

Figura 17: Ottimizzazione parametrica

Il “cross-validation” è anch’esso un nested operator e permette di realizzare la convalida incrociata una volta specificato il numero k-fold.

Pertanto, il set di training viene suddiviso automaticamente in k sottoinsiemi di cui k-1 viene utilizzato per l’addestramento mentre l’ultimo sottoinsieme viene utilizzato per la validazione. Come mostrato in figura 18, nella zona “training” si inserisce l’operatore da addestrare mentre, in quella di “testing”, si valida di volta in volta il modello trovato e si calcola l’errore.

Si noti che la convalida incrociata si utilizza nel sotto-processo “train model” ed è diversa dalla validazione finale fatta con il set di dati di test.

Questa tecnica si utilizza spesso negli algoritmi di machine learning e consente di ottenere modelli con una minore probabilità di problemi di overfitting. In alternativa, per il calcolo delle performance del modello (indispensabile per l’operatore “optimal parameters”) si potrebbe utilizzare “apply model” combinato con “performance” sugli stessi dati di training ma con il rischio di modelli di predizione meno performanti nella fase del “test model”.

Un limite di RapidMiner è di non poter selezionale in modo diretto nell’ottimizzazione l’hidden layers sizes ovvero il numero di neuroni di uno strato nascosto. Per ovviare a questo problema, si utilizza l’operatore “set macro”: deve essere inserito uno ogni hidden layer presente quindi, nel caso del neural network classico ne basta uno mentre per il deep learning due o più (Figura 19).

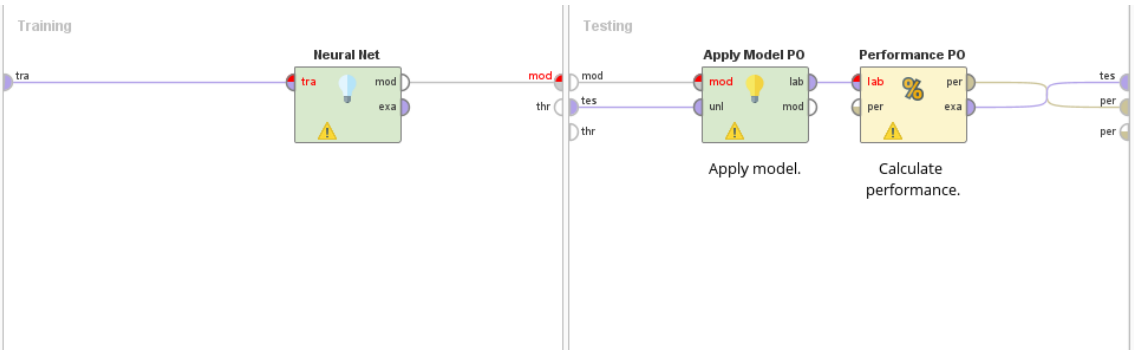

Figura 18: Convalida incrociata

Oltre al “neural network” di figura 18, è possibile utilizzare anche il “deep learning”. Quest’ultimo ha alcuni vantaggi:

- Consente di determinare automaticamente il numero ottimale di epoche di addestramento;

- È in grado di creare un modello misto che considera sia variabili qualitative che quantitative;

E alcuni svantaggi:

- Richiede molto tempo per essere addestrato perché è costituito da più strati nascosti e da tanti neuroni al loro interno;
- Il processo di ottimizzazione parametrica è più complesso.

Nel caso del "neural network" il database deve essere caratterizzato da sole variabili numeriche. Quindi, così come nella regressione, ci sono due diversi modi di agire:

1. Creare un modello tramite "neural network" applicato all'intero training set convertendo le variabili qualitative (polynominal) in numeriche (variabili dummy)
2. Creare n modelli di "neural netwrok" relativi a n sottoinsiemi del training set utilizzando n volte l'operatore "filter example"

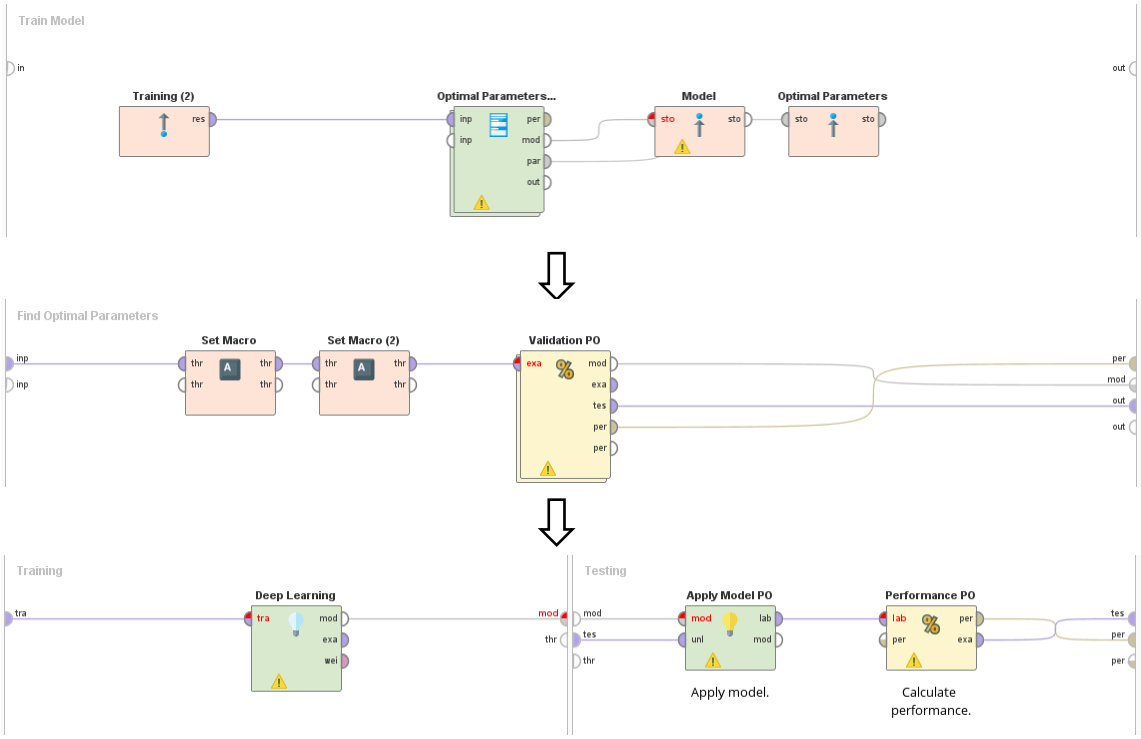

Figura 19:Train model deep learning

4.2.3 Implementazione del metodo: Random Forest

Il Random Forest fa parte delle tecniche di machine learning e il procedimento di modellizzazione è molto simile a quello descritto nel paragrafo precedente nel caso delle reti neurali.

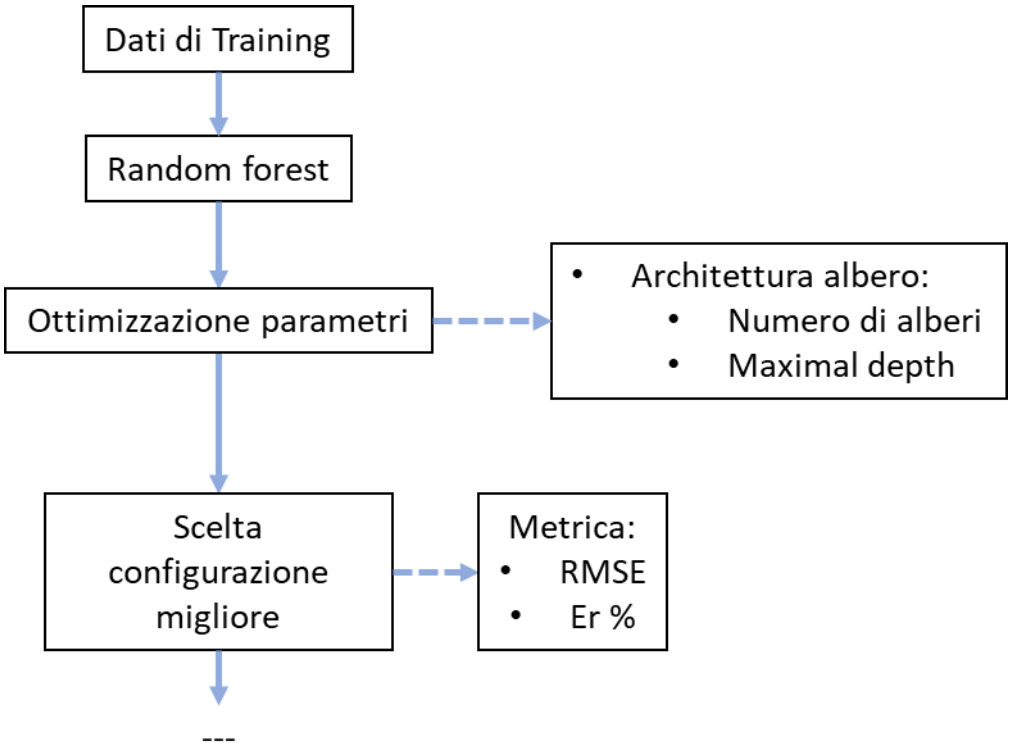

Figura 20: Metodo random forest

Come da figura 20, il primo step consiste nel richiamare i dati di training e procedere con l'implementazione del metodo Random Forest. Si noti che, seppur meno complesso rispetto le reti neurali, anche in questo caso può essere utile avvalersi dell'ottimizzazione parametrica per determinare i valori dei parametri ideali relativi al maximal depth (la profondità dell'albero) e al numero di alberi. Segue un'operazione di selezione del modello con la configurazione ottimale.

4.2.3.1 RapidMiner

Anche dal punto di vista di RapidMiner il processo è molto simile a quello delle reti neurali (Figura 21).

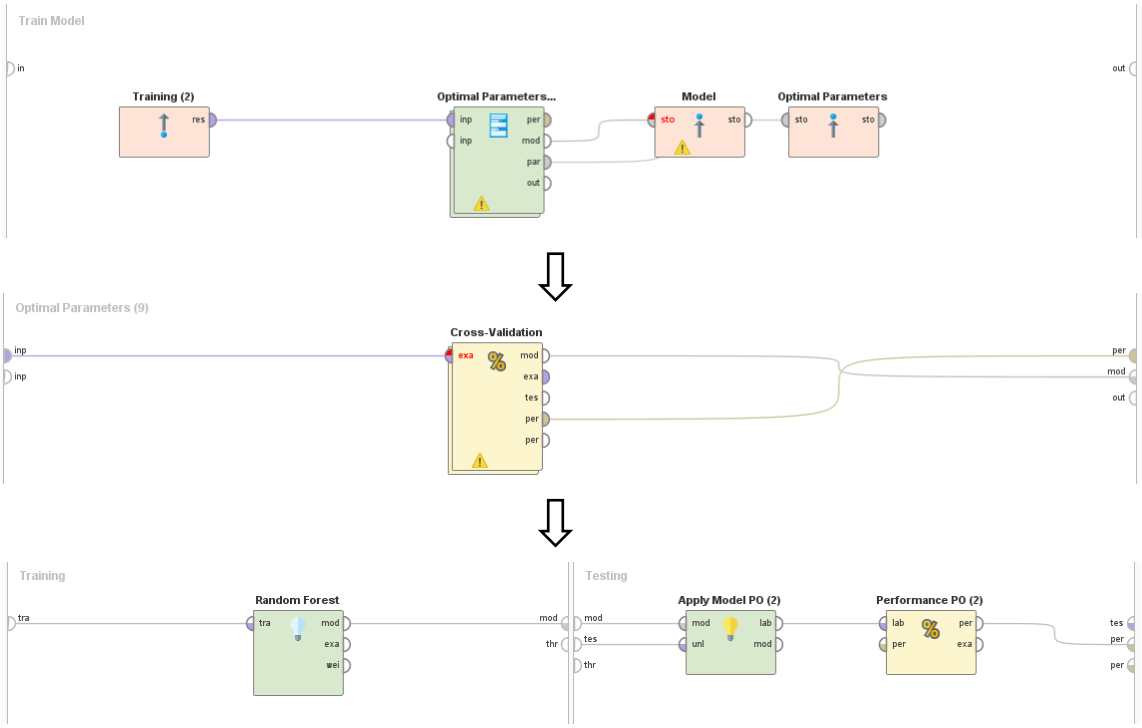

Figura 21:Train model random forest

La procedura è simile a quella vista precedentemente con le reti neurali e presenta le seguenti differenze:

- L'operatore "neural network" è sostituito dal "random forest";
- Non c'è il bisogno di utilizzare il "set macro" perché in questo caso di machine learning RapidMiner non presenta limiti nella scelta dei parametri da ottimizzare;
- Nell'operatore "optimal parameters" cambiano i parametri da ottimizzare (Figura 21).

4.3 Validazione Modelli

Una volta determinati tutti i possibili modelli segue la fase di test. Come detto precedentemente, il database iniziale si divide in due grandi gruppi: dati di training e dati di testing. I primi, si utilizzano per generare i modelli mentre i secondi per validarli. Questa fase è essenziale nella stima dei costi perché consente di capire quali dei modelli è più idoneo al tipo di studio. Per confrontare tra loro la regressione, la rete neurale

semplice, il deep learning e il random forest si utilizzano come metriche gli errori relativi. Ne esistono diversi, sono riportati di seguito (il più utilizzato è detto MAPE):

- MAPE: Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{y_i} \times 100 \right) \quad (4)$$

- Relative Error

$$RE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{\hat{y}_i} \times 100 \right) \quad (5)$$

- Relative Error Lenient

$$REl = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{\max\{y_i, \hat{y}_i\}} \times 100 \right) \quad (6)$$

Inoltre, anche gli errori assoluti possono essere utilizzati:

- MSE:

$$MSE = \frac{1}{n}\ \sum_{i=1}^{n}(\ y_i - \hat{y}_i\ )^2 \quad (7)$$

- RMSE:

$$RMSE = \sqrt{\frac{1}{n}\ \sum_{i=1}^{n}(\ y_i - \hat{y}_i\ )^2} \quad (8)$$

Dove:

- y: Valore di actual cost o should cost stimato
- $\hat{y}$: Valore di actual cost o should cost noto
- n: numero totale di coppie

4.3.1 RapidMiner

Su RapidMiner, riprendendo la figura 10, la validazione si realizza nel sotto-processo "Test Model" (Figura 22). L'idea è quella di applicare il modello in esame a un set di dati nuovi pertanto, come da figura, si richiamano i dati di testing e si applicano al modello tramite "apply model". L'operatore "performance" permette di calcolare gli errori

descritti nel paragrafo precedente ad esclusione del MAPE (che si può calcolare sul software Excel).

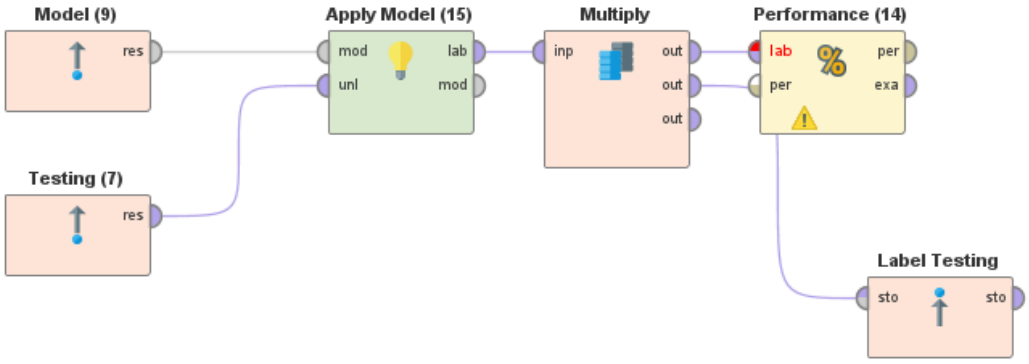

Figura 22: Test model

L'operatore "multiply" di figura 22 permette di creare una copia dei dati "label testing" per memorizzarli tramite "remember". Essi riportano tutti i dati di test con in aggiunta una colonna finale contenente i valori di output predetti da un modello.

In figura 11, dopo la fase di "Test Model" c'è il "Create Prediction" (Figura 23) dove richiamando il modello, i dati di training e i dati di test si realizza un simulatore del modello.

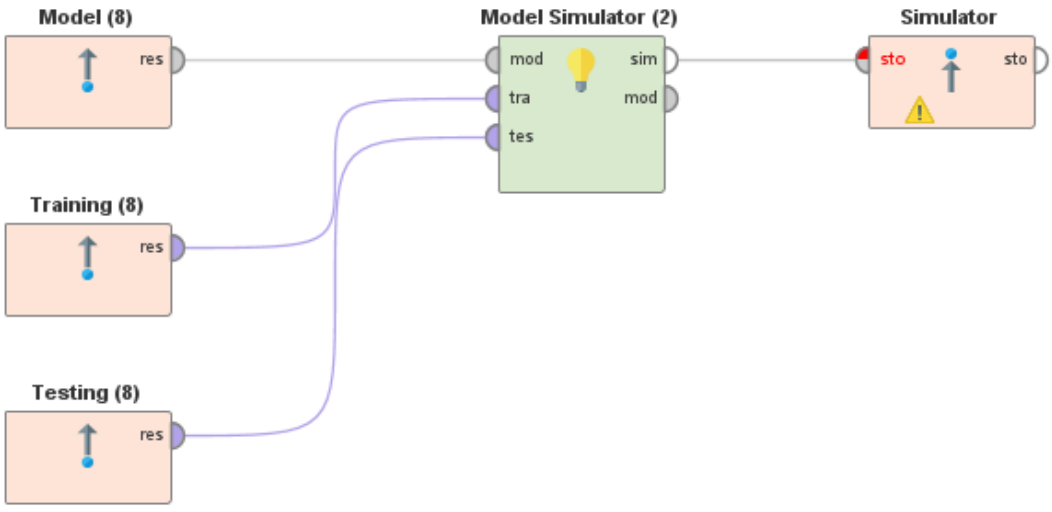

Figura 23. Create prediction

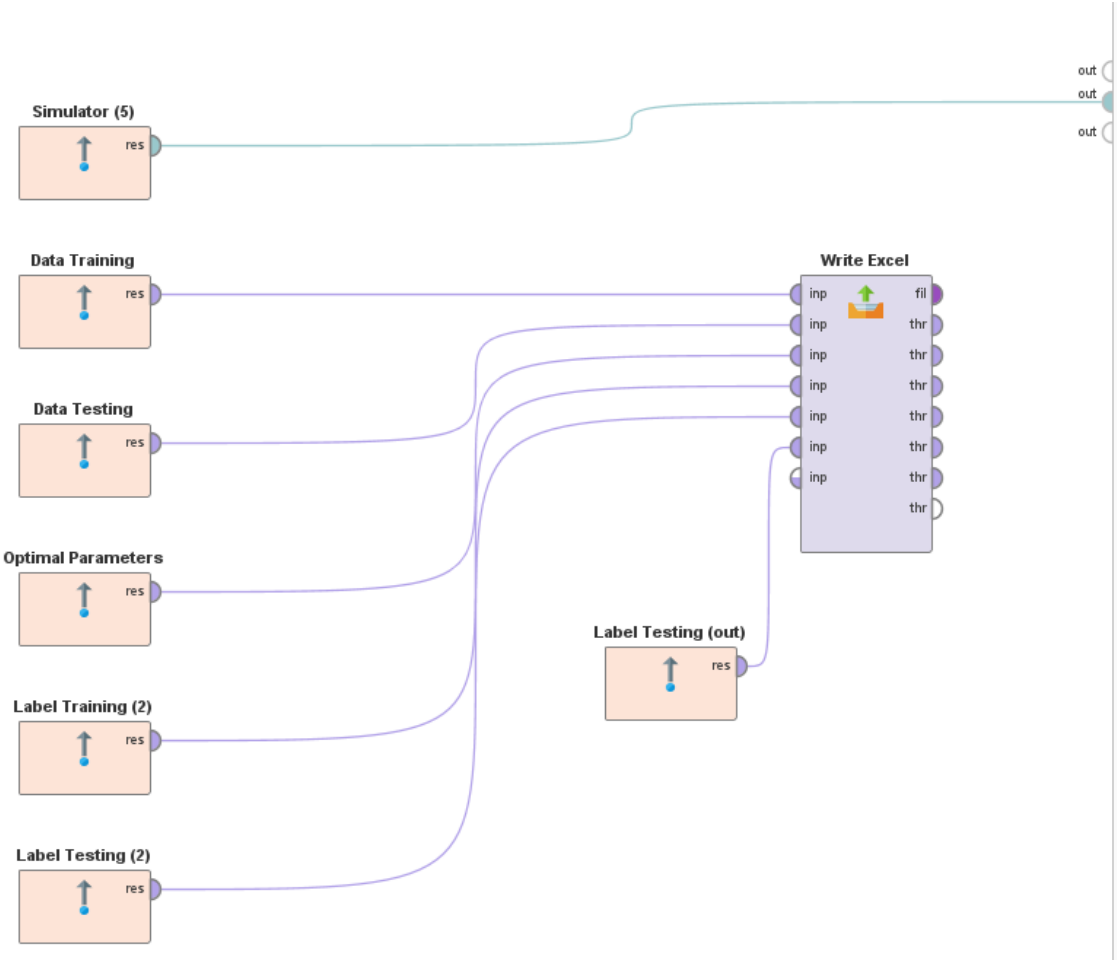

Figura 24: Output

Infine, l'ultimo sotto-processo "output", contiene al suo interno tutti i risultati ottenuti nei precedenti sub-process (Figura 24) richiamati con la funzione "recall" e trascritti in un foglio file excel con "write excel" (ad eccezione del "model simulator" che si può visualizzare nei risultati di RapidMiner).

5 Caso Studio: Dischi compressore assiale

5.1 Il contesto aziendale

Il lavoro di questa tesi è stato svolto in collaborazione con l'azienda BH.
Baker Hughes (BH) è il primo e unico fornitore globale fullstream
in grado di offrire congiuntamente attrezzature all'avanguardia, servizi e soluzioni digitali per l'intera catena del valore del petrolio e del gas. È l'unica società ad offrire contemporaneamente questo numero di competenze nell'ambito dell'intera filiera dell'Oil & Gas, dal segmento upstream (a monte) fino al downstream (a valle), passando per il midstream (al centro):

- Upstream: Il settore upstream estrae e lavora gli idrocarburi e viene spesso denominato "Exploration and Production" (E&P) dove per exploration (esplorazione) si intende la ricerca degli idrocarburi nel sottosuolo o sott'acqua mediante l'analisi dei risultati di ispezioni geologiche e indagini sismiche mentre per production (produzione) si intende l'estrazione di petrolio e gas naturale dai giacimenti tramite l'uso di pozzi di trivellazione e le conseguenti fasi di lavorazione e di trasporto.
- Midstream: Il settore midstream gestisce la lavorazione, il trasporto e l'immagazzinamento del petrolio e del gas; fornisce un importante collegamento tra i settori upstream e downstream, perché raccoglie le risorse estratte dal settore upstream e le immagazzina negli impianti di raffinazione del settore downstream (mediante reti di condutture, autocarri, chiatte, navi cisterna e carri cisterna), dove tali risorse vengono trasformate in vari prodotti.
- Downstream: è la fase finale del processo dell'industria del gas e del petrolio, incentrata sulla raffinazione del petrolio greggio e sulla lavorazione/purificazione del gas naturale, oltre che sulla commercializzazione e distribuzione dei prodotti derivati da petrolio e gas.

Questo ricco portafoglio mette Baker Hughes in condizioni di creare nuove fonti di valore, migliorando la produttività e l'economicità dei progetti attraverso un'offerta integrata di attrezzature e servizi.

Con operazioni in oltre 120 paesi e circa 70.000 dipendenti, l'azienda garantisce ai clienti di rispondere prontamente alla volatilità che caratterizza il mercato dell'Oil & Gas, di lavorare in modo più efficiente, nonché di fornire energia a un maggior numero di persone, aumentando la produttività e riducendo al minimo i costi.

In Italia, sono presenti due dei quattro business del gruppo Baker Hughes: Turbomachinery & Process Solutions - il cui centro decisionale mondiale ha sede a Firenze, rappresentato dalla storica azienda italiana Nuovo Pignone - e quello di Oilfield Services. Baker Hughes conta in Italia sette siti produttivi in sei regioni (Firenze, Massa, Bari, Vibo Valentia, Talamona (SO), Casavatore (NA) e Cepegatti (PE) e un cantiere per l'assemblaggio di moduli industriali ad Avenza (Carrara). Complessivamente, i dipendenti in Italia sono oltre 5.000.

Nuovo Pignone rappresenta per il gruppo BH il centro di eccellenza mondiale per lo sviluppo e la produzione di turbine a gas, compressori e pompe ed è un punto di riferimento per le applicazioni dell'Industria 4.0.

In Italia, infatti, l'azienda svolge gran parte delle attività di Ricerca & Sviluppo legate alle turbomacchine e a innovative applicazioni digitali, dalla stampa 3D, all'automazione, al monitoraggio e alla diagnostica da remoto sulle macchine installate presso i propri clienti in tutto il mondo.

5.1.1 Le origini

L'attuale azienda, localizzata a Firenze, nasce nel 1842 come fonderia di ghisa "Il Pignone" e inizia l'attività meccanica nei primi del '900, specializzandosi nel settore dei compressori alternativi.

Rilevata da ENI nel gennaio del 1954, la società inizia a produrre attrezzature per il settore energetico con il nuovo nome "Nuovo Pignone – Officine
meccaniche e fonderia".

Negli anni l'azienda cresce, specializzandosi nella realizzazione di impianti nei luoghi più remoti e mantenendo sempre una costante redditività.

ENI mantiene la proprietà fino al 1994, quando Nuovo Pignone viene acquisita dalla società americana General Electric, che decide di investire sulle competenze manifatturiere e sulle capacità tecnologiche e innovative della società Nuovo Pignone divenendone socio di maggioranza. Gli anni dal 1994 a oggi rappresentano per GE anche i primi 20 anni di crescita nel settore petrolifero e del gas e la nascita di GE Oil & Gas.

Dalla fusione tra GE Oil & Gas e Baker Hughes nasce nel 2017 una nuova società: "Baker Hughes, a GE company"; BHGE sceglie di stabilire a Firenze il centro decisionale globale del business Turbomachinery & Process Solutions (TPS).

A settembre 2019 il gruppo General Electric annuncia di essere scesa al 38,4% del controllo di Baker Hughes (dal 50,06% già ridotto dall'originario 62,5%) destinando il ricavato della vendita delle azioni al rilancio del gruppo. Baker Hughes, a GE Company (BHGE) torna ad essere semplicemente Baker Hughes.

Ad oggi, BH è una società quotata presso la borsa di New York e il Chairman e CEO di BH è Lorenzo Simonelli, precedentemente Presidente e CEO di GE Oil & Gas.

Attualmente, il Presidente di Nuovo Pignone è Michele Stangarone, in carica dal 1° aprile 2018.

5.2 Caso studio

Come detto in precedenza, l'obiettivo dell'azienda BH è quello di sviluppare un tool di supporto al progettista che consenta di effettuare valutazioni di carattere economico su componenti di turbomacchine in fase di progettazione concettuale.

In questa tesi è stato fatto uno studio su dati reali per valutare le potenzialità dei tre metodi: linear regression, neural network (e deep learning) e random forest. I componenti che sono stati studiati sono i dischi dei compressori assiali e fanno riferimento a due macchine che chiameremo macchina 1 e macchina 2.

Prima di entrare nello specifico, occorre fare un richiamo dello stato dell'arte in quanto la stima dei costi dei dischi del compressore assiale è stata già trattata in precedenti lavori.

In[13], sono state utilizzate tecniche di regressione lineare per stimare il costo totale dei dischi finiti.

Lo studio è stato fatto in maniera molto approfondita attribuendo una voce di costo ad ogni operazione.

Pertanto, il costo totale è stato ottenuto da due voci:

1. Il costo della parte forgiata, compreso il materiale;
2. Il costo della lavorazione per ottenere il disco finito.

Per comprendere meglio di che componente si tratta, di seguito sono state riportate le figure inerenti alla parte forgiata e al disco finito.

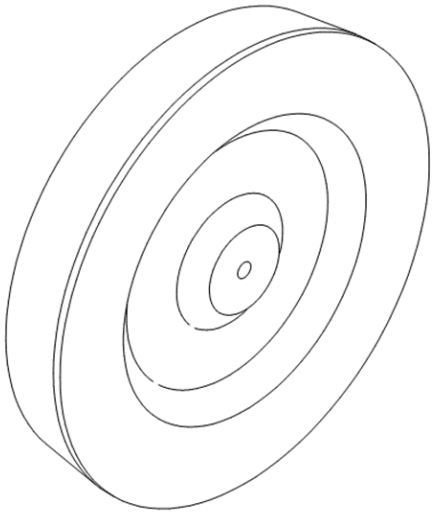

Figura 25: Forgiato

Il disco in figura 25 è il risultato del processo di forgiatura: la forma è stata ottenuta a partire da una bielletta in metallo, tramite successive azioni di ricalcatura eseguite mediante una pressa.

Quindi, Il costo della parte forgiata comprende:

- Il costo del materiale: funzione del volume del materiale, della densità del materiale e del costo unitario del materiale;
- Il costo del processo di forgiatura.

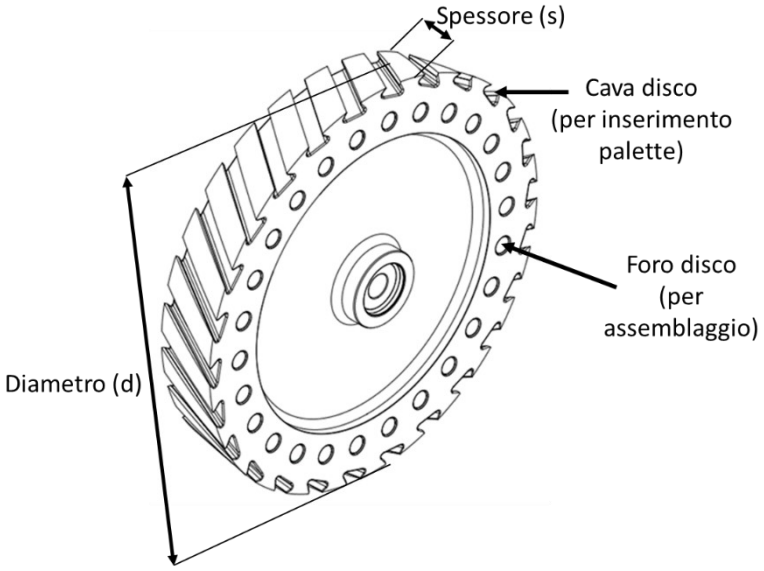

Figura 26: Disco finito

A partire dal pezzo forgiato, tramite la tornitura ed eventuali trattamenti termici è stato ottenuto il disco semilavorato.

Il disco finito, invece, è stato rappresentato in figura 26, rispetto al semilavorato, presenta delle cave alle estremità per il posizionamento delle palette del compressore assiale e dei fori intermedi per l'assemblaggio del disco con gli altri componenti della turbina a gas.

Per quanto riguarda quindi il costo di lavorazione per ottenere il finito, sono state considerate le seguenti voci di costo:

- Il costo di tornitura;
- Il costo della perforazione dei fori;
- Il costo di brocciatura;
- I costi di ispezione dimensionale e test.

Sommando tutti i costi, è stata ottenuta l'equazione di regressione risultante in grado di stimare il costo totale del disco finito.

Realizzare un'analisi molto dettagliata, permette di individuare quelle operazione più gravose in termini di costo.

Testando l'equazione parametrica su due dischi nuovi, il costo della forgiatura è stato rilevato come il più impattante in quanto questa operazione deve essere eseguita in un impianto specifico e richiede tempi di configurazione e posizionamento molto lunghi.

I risultati ottenuti mostrano come le incertezze più elevate sono state rilevate nell'operazione di forgiatura (circa il 10% del valore deterministico). Tutte le altre voci di costo presentano incertezze inferiori sul costo finale (meno del 2% del valore deterministico).

Le imprecisioni e le incertezze del funzionamento della forgiatura e dei relativi costi sono strettamente legate alla complessità del processo, caratterizzata da un basso livello di automazione e da un elevato grado di manodopera.

È chiaro quindi che il precedente studio eseguito sui dischi del compressore assiale si basa su un approccio empirico di regressione lineare molto dettagliato e quindi è fortemente dipendente dalle conoscenze dell'analista dei processi di lavorazione dei componenti analizzati.

In questa tesi, invece, è stato definito un nuovo approccio "sistematico" in cui l'obiettivo è stato quello di stimare due voci di costo (costo del materiale e costo del processo) senza scendere nel dettaglio di tutte le operazioni.
Inoltre, nello studio precedente, è stato considerato sempre lo stesso materiale e la stessa tecnica di forgiatura (a stampo aperto).
In questo caso, sono stati considerati nuovi materiali (5 in totale) ed è stata introdotta anche la tecnica di forgiatura a stampo chiuso, diversa da quella a stampo aperto, in particolare:

- Forgiatura a stampo chiuso: il pezzo è compresso tra due stampi sagomati ed assume la forma della cavità tra essi compresa (permette di risparmiare materiale);
- Forgiatura a stampo aperto: stampi non sagomati e di conseguenza geometrie finali semplici e maggiore spreco di materiale.

Quindi, lo studio è stato generalizzato rispetto al precedente lavoro, in quanto sono state introdotte nuove variabili qualitative (o categoriche) ed è stato realizzato, oltre che con la tecnica tradizionale di regressione, con tecniche innovative non lineari di intelligenza artificiale.
Più nel dettaglio, inizialmente sono state condotte delle analisi di Should Cost tramite l'uso del software LeanCOST®.
I risultati delle analisi sono stati inserirti nell'appendice e costituiscono il database iniziale o storico. Una parte della tabella rappresentata nell'appendice finale è riportata in tabella 2. Successivamente, seguendo le linee guida descritte nel capitolo 4, tramite software RapidMiner sono stati implementati i metodi. Nelle successive tabelle, i dati sensibili sono stati oscurati per motivi di riservatezza (in particolare i costi).

Macchina	Cod	d	s	M	N° cave	R	Tratt	Lotto	Grezzo di partenza	Peq [Kg]	Psl	Pg	Veq [m3]	Tot
Disco Macchina 1	A	AA	AAA	M1	X	I	si	1	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	5	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	10	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	20	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	50	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	1	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	5	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	10	Forgiato a stampo aperto	L	M	N	O	€ -

Tabella 2: Database iniziale

Il database è costituito dai seguenti dati di input:

- Codice finito (cod);
- Diametro finito (d);
- Spessore finito (s);
- Materiale (M);
- Numero cave;
- Rugosità (R);
- Presenza trattamenti (Tratt);
- Lotto;
- Grezzo di partenza;
- Peso disco equivalente (Peq);
- Peso semilavorato (Psl);

- Peso grezzo (Pg);
- Volume disco equivalente (Veq);

L'ultima colonna, il costo totale (Tot), è relativa ai dati di output e deriva dalla somma del costo del materiale e del costo del processo. A sua volta il costo del processo è dato da: forming process cost, machining cost, non destructive tests and inspections (se presenti) e altro.

Pertanto, è stata considerata un'unica voce di costo relativa al processo che racchiude tutte le operazioni effettuate per ottenere il disco finito.

Trattandosi dello stesso componente, i processi di produzione sono uguali a quelli descritti a inizio paragrafo ma, in questo caso, la stima dei costi è stata sviluppata senza entrare nel dettaglio nelle singole operazioni.

Ad ogni codice finito corrisponde un tipo particolare di disco caratterizzato da un valore specifico di diametro finito, spessore finito e numero di cave (Figura 26);

Il volume del disco equivalente è stato calcolato con i dati relativi al diametro e allo spessore del disco finito.

In particolare, è stato considerato un disco fittizio (o equivalente) in cui a differenza del disco finito di figura 27, non sono presenti le cave per le palette e i fori (come se fosse un disco pieno).

Il database presenta circa 90 records che sono stati generati facendo variare, per uno stesso disco, i valori dei dati non geometrici (come ad esempio il lotto o il materiale). Nella tabella 2 sono stati studiati nove dischi, dalla A alla I; a questi se ne aggiunge un altro L che, come vedremo nei paragrafi successivi, è stato utilizzato per una seconda prova di test.

5.3 Scelta database

Il database iniziale, così come presentato in parte nella tabella 2, non è ancora idoneo per essere utilizzato per implementare i metodi di stima dei costi ma necessita di una fase di pre-processing.

Questa operazione iniziale è stata fatta per individuare il peso di ciascuna variabile indipendente e per eliminare eventuali valori anomali o dati duplicati.

Se si considerasse il database per intero, così come riportato nell'appendice, la regressione lineare presenterebbe errori troppo elevati sulla stima finale mentre, i metodi di machine learning sarebbero meno precisi in quanto ci potrebbero essere degli input ininfluenti (sul costo) che causerebbero un rumore bianco sulla stima finale.
In generale, nelle tecniche tradizionali di regressione lineare è importante scegliere solo i dati di input necessari ai fini dell'analisi; nelle tecniche innovative di machine learning, è possibile considerare anche input meno "importanti" nella stima dei costi ma bisogna evitare di considerare quelli con dipendenza nulla.
A tal proposito, è stato fatto uno studio di analisi di varianza fattoriale multivariata per determinare le variabili indipendenti maggiormente correlate con quelle dipendenti. Come si nota dalla tabella 3, i valori con significatività minore di 0,1 (evidenziati in verde) sono stati presi in considerazione.

Input	Output	Significatività
Volume disco equivalente	Costo Processo	0,000000000000007
	Costo Materiale	0,000000000034661
Materiale	Costo Processo	0,000000000000000
	Costo Materiale	0,000000000000000
Rugosità finito	Costo Processo	1,000000000000000
	Costo Materiale	1,000000000000000
Presenza trattamenti	Costo Processo	0,287756351108742
	Costo Materiale	0,880266970706251
Lotto	Costo Processo	0,000000000000023
	Costo Materiale	0,767459104124280
Grezzo di partenza	Costo Processo	0,002319938911480
	Costo Materiale	0,072432694138500

Tabella 3: MANOVA fattoriale

Come dato di input geometrico è stato considerato solo il volume disco equivalente in quanto deriva da altri dati geometrici caratterizzanti i dischi (il diametro e lo spessore finale). Si noti come non è stata considerata una sola variabile dipendente, il costo totale, ma sono stati presi separatamente il costo del materiale e il costo del processo.

Pertanto, come da requisito di BH [R12], i vari modelli dovranno essere in grado di prevedere le singole voci di costo (da cui comunque è possibile ricavare il totale).

Dai risultati dell'analisi MANOVA fattoriale è evidente come il volume disco equivalente (e quindi i rispettivi dati geometrici), il materiale e il grezzo di partenza siano fortemente correlati con le variabili dipendenti. La rugosità e la presenza di trattamenti hanno un impatto bassissimo e pertanto sono trascurabili mentre il lotto ha un comportamento dipendente dal tipo di output.

In tabella 4 sono riportati i drivers definitivi scelti per implementare i metodi.

Input	Costo Materiale	Costo Processo
Diametro	X	
Spessore	X	
Materiale	X	X
N° cave		
Lotto		X
Grezzo di partenza	X	X
Peso disco equivalente	X	X
Volume disco equivalente		X

Tabella 4: Input

Oltre all'analisi MANOVA fattoriale, nella fase iniziale di pre-processing è importante individuare eventuali valori anomali e dati duplicati.

Non sono stati rilevati dati anomali, ovvero dati iniziali incoerenti con il resto del database ma, per quanto riguarda lo studio relativo al costo del materiale, sono stati eliminati tutti i dati duplicati che si sono venuti a creare escludendo l'input "lotto".

I duplicati si sono generati perché nel costruire il database con LeanCOST®, spesso le righe sono state create mantenendo tutti i valori di input costanti e modificando solo il lotto.

Dai risultati del MANOVA fattoriale, non c'è alcuna dipendenza tra "lotto" e "costo materiale" e di conseguenza, eliminando tutti i dati del lotto si generano i duplicati.

Discorso diverso invece per il "costo processo" in cui non sono state rilevate situazioni similare a quella appena descritta.

Pertanto, nel caso del costo del processo il database presenta lo stesso numero di records iniziali mentre, per la voce di costo relative al materiale, il numero di righe si è notevolmente ridotto (da 90 a 33 righe).

Segue l'operazione di split data dove si suddividono i dati in training e testing: nel caso del costo del materiale, l'80% dei dati è stato utilizzato come training e il restante 20% come testing; nel caso del costo del processo le percentuali sono 70% training e 30% testing.

La suddivisione non è stata fatta in maniera casuale: nel caso del "costo del materiale" sono stati considerati più dati di training in quanto il database è notevolmente ridotto rispetto al "costo processo" e i materiali considerati presentano costi unitari molto diversi tra loro. L'idea è stata quella di eseguire un addestramento con un numero maggiore di dati di training per ovviare a queste problematiche.

Come vedremo in seguito, nonostante questa distinzione nella suddivisione dei dati a favore del "costo materiale", sono stati ottenuti risultati più precisi con il "costo processo" a dimostrazione del fatto che le performance delle tecniche di stima dei costi sono estremamente dipendenti dalle forme del database iniziale.

Ricapitolando, partendo dai dati iniziali sono stati ottenuti due database: uno relativo al costo del materiale e l'altro relativo al costo dei processi. Per entrambi, sono stati individuati i drivers più importanti (maggiormente correlati con l'output) ed eliminati eventuali dati anomali o duplicati. Dalla tabella 2 si nota come non tutti le variabili indipendenti sono di tipo numerico; in particolare, il materiale e il grezzo di partenza sono variabili qualitative. Su RapidMiner, gli algoritmi "random forest" e "deep learning" sono in grado di lavorare con variabili miste ma ciò non vale per il "linear regression" e per il "neural network".

Pertanto, è stata utilizzata la tecnica della variabile dummy per generare dei database costituiti da soli dati numerici. Nella tabella 4 sono stati inserite le uniche due variabili qualitative considerate nei due database con le relative conversioni. Una variabile binaria, o dummy, assume valore 0 o 1, a seconda che sia soddisfatta o meno una data condizione. Prendendo in considerazione la variabile qualitativa "materiale", sono state

utilizzate cinque variabili dummy, ciascuna riferita a un tipo di materiale. Per quanto riguarda il grezzo di partenza, ci sono due variabili binarie di cui una è relativa al caso forgiato a stampo aperto, l'altra al caso forgiato a stampo chiuso. Nella prima riga sono soddisfatte le condizioni materiale = M1 e grezzo di partenza = Forgiato a stampo chiuso; ovviamente non ci saranno mai due condizioni di una stessa variabile di riferimento soddisfatte per una stessa riga.

Materiale = M1	Materiale = M2	Materiale = M3	Materiale = M4	Materiale = M5	Grezzo di partenza = Forgiato a stampo aperto	Grezzo di partenza = Forgiato a stampo chiuso
1	0	0	0	0	0	1
0	1	0	0	0	1	0
0	0	0	1	0	1	0
1	0	0	0	0	1	0
1	0	0	0	0	0	1

Tabella 5: Variabile Dummy

Una volta definiti i sottoinsiemi di dati dai training e di testing, si può procedere con l'implementazione dei metodi.

5.4 Implementazione Metodi

Partendo dai dati di training sono stati implementati i vari metodi sul software RapidMiner. Il primo metodo analizzato è la regressione lineare, i database di partenza sono quelli numerici (Tabella 6 e tabella 7):

Costo Materiale	
Attributo	**Coefficiente**
Materiale = M1	-4621,9
Materiale = M2	-3640,7
Materiale = M3	7858,0
Materiale = M4	4101,6
Materiale = M5	-3697,0
Grezzo di partenza = Forgiato a stampo aperto	167,6
Grezzo di partenza = Forgiato a stampo chiuso	-167,6
Diametro	-35,9
Spessore	-157,6
Peso finito [Kg]	145,9
Intercetta	25329,6

Tabella 6: Modello costo materiale

Costo Processo	
Attributo	**Coefficiente**
Materiale = M1	-689,8
Materiale = M2	-475,7
Materiale = M3	906,1
Materiale = M4	696,4
Materiale = M5	-436,9
Grezzo di partenza = Forgiato a stampo aperto	121,1
Grezzo di partenza = Forgiato a stampo chiuso	-121,1
Lotto	-15,5
Peso disco equivalente [Kg]	-1,3
Volume disco equivalente [m3]	35476,1
Intercetta	1195,0

Tabella 7: Modello costo processo

Le equazioni parametriche delle regressioni multiple lineari hanno la stessa forma dell'equazione 2 e si ottengono dalla sommatoria dei prodotti tra gli attributi delle tabelle 6 e 7 e i relativi coefficienti (Equazione 9 ed equazione 10).

$$Cmateriale = 25329{,}6 - 4621{,}9 * M1 - 3640{,}7 * M2 + 7858{,}0 * M3 + 4104{,}6 * M4 - 3697 * M5 + 167{,}6 * forgiato\ aperto - 167{,}6 * forgiato\ chiuso - d * 35{,}9 - s * 157{,}6 + Peq * 145{,}9 \quad (9)$$

$$Cprocesso = 1195{,}0 - 689{,}8 * M1 - 475{,}7 * M2 + 906{,}1 * M3 + 696{,}4 * M4 - 436{,}9 * M5 + 121{,}1 * forgiato\ aperto - 121{,}1 * forgiato\ chiuso - lotto * 15{,}5 - 35476 * Veq - 1{,}3 * Peq \quad (10)$$

La bontà delle equazioni ottenute è data dal coefficiente R^2 che nel modello costo materiale assume valore 0,90 mentre nel modello costo processo è pari a 0,84. Essendo valori accettabili, i due modelli sono stati confermati e utilizzati nelle fasi successive di test.

Per quanto riguarda le tecniche di stima dei costi innovative di machine learning, sappiamo che determinare il valore ottimo dei parametri iniziali (ad esempio numero di neuroni nascosti nel caso delle reti neurali o numero di alberi nel caso del random forest) non è un'operazione semplice e richiede un processo di ottimizzazione parametrica.

Di seguito sono stati analizzati i vari casi.

5.4.1 Rete Neurale Semplice

Su RapidMiner, l'uso della funzione "neural network" richiede in ingresso l'utilizzo del database numerico con variabili binarie. Come detto nel capitolo 4.2.2, per determinare il numero di neuroni nascosti è stata utilizzata l'equazione 3 empirica.

In figura 27 sono state raffigurate le architetture delle reti neurali semplici di entrambi i casi studiati.

I segmenti che collegano i neuroni di due strati successivi presentano delle tonalità differenti a seconda del valore del peso del collegamento.

Nella figura 27, a sinistra è stata rappresentata l'architettura della rete neurale relativa al costo materiale mentre a destra quella relativa al costo del processo.

Il numero di neuroni contenuto nei vari strati è identico nei due casi mentre, il valore dei parametri caratteristici è diverso.

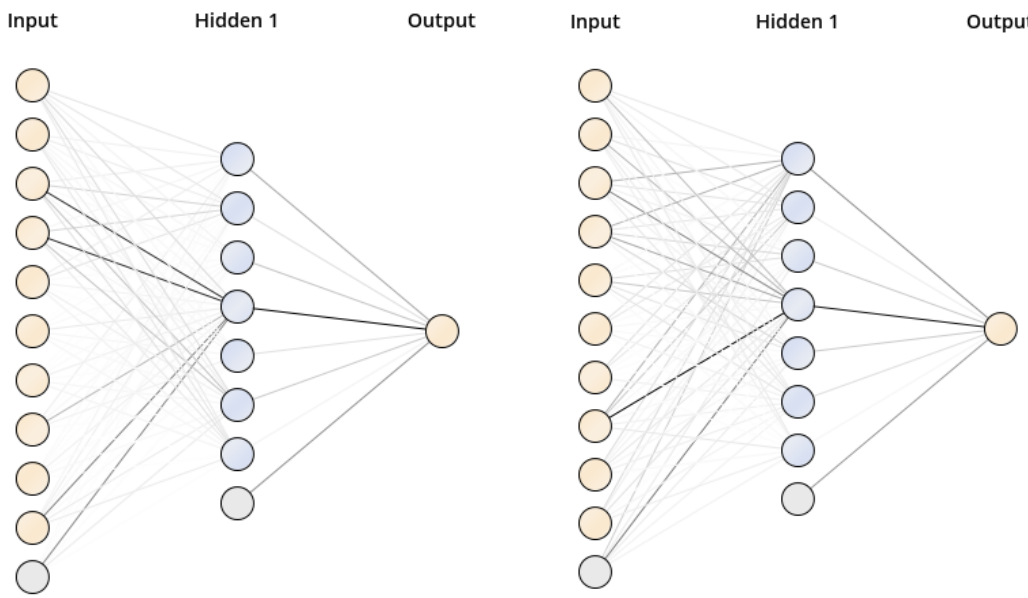

Figura 27: Architettture reti neurali

Per quanto riguarda il parametro “numero di cicli di training”, è stata utilizzata la funzione di ottimizzazione parametrica per determinare il valore ottimale: sono stati inseriti in ingresso dei possibili valori ed è stata calcolata la corrispondente voce di costo “relative error leninet”. RapidMiner calcola in automatico le diverse iterazioni in modo da determinare il valore ottimo di cicli di training.

I risultati dell’ottimizzazione parametrica sono stati riportati su due grafici in figura 28 e figura 29.

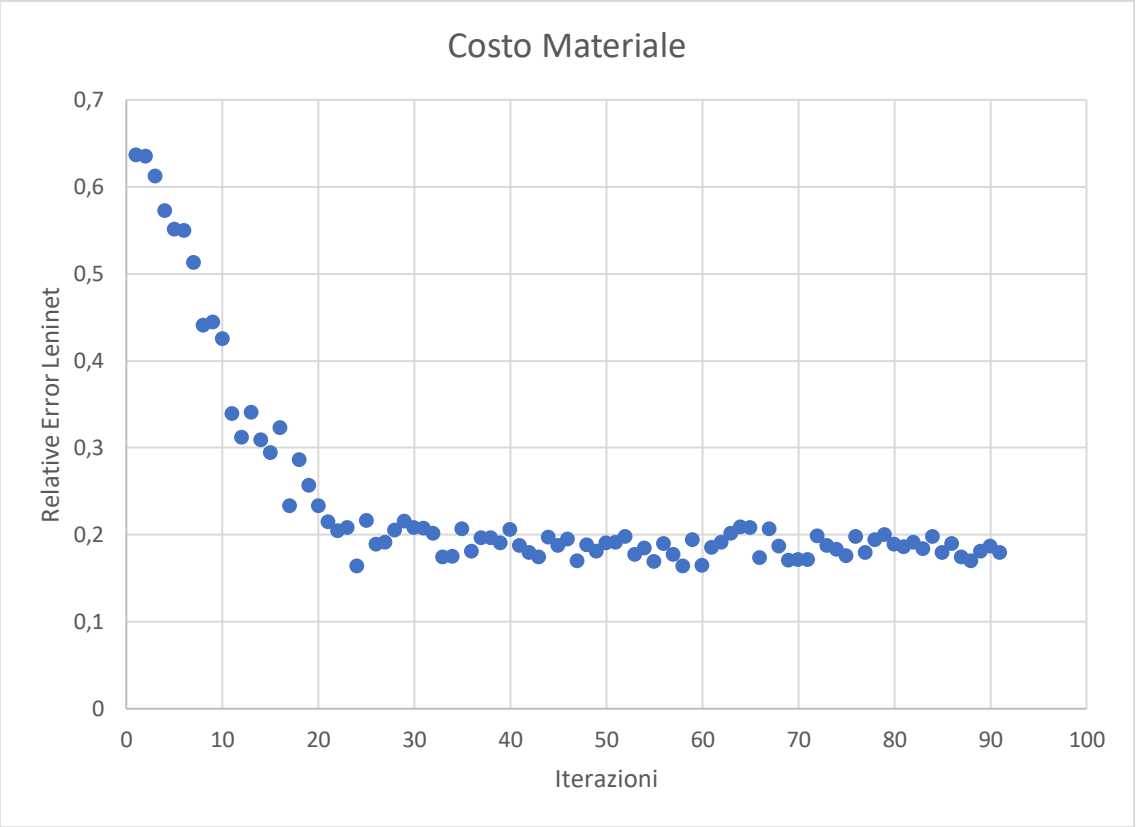

Figura 28: Ottimizzazione parametrica rete neurale (costo materiale)

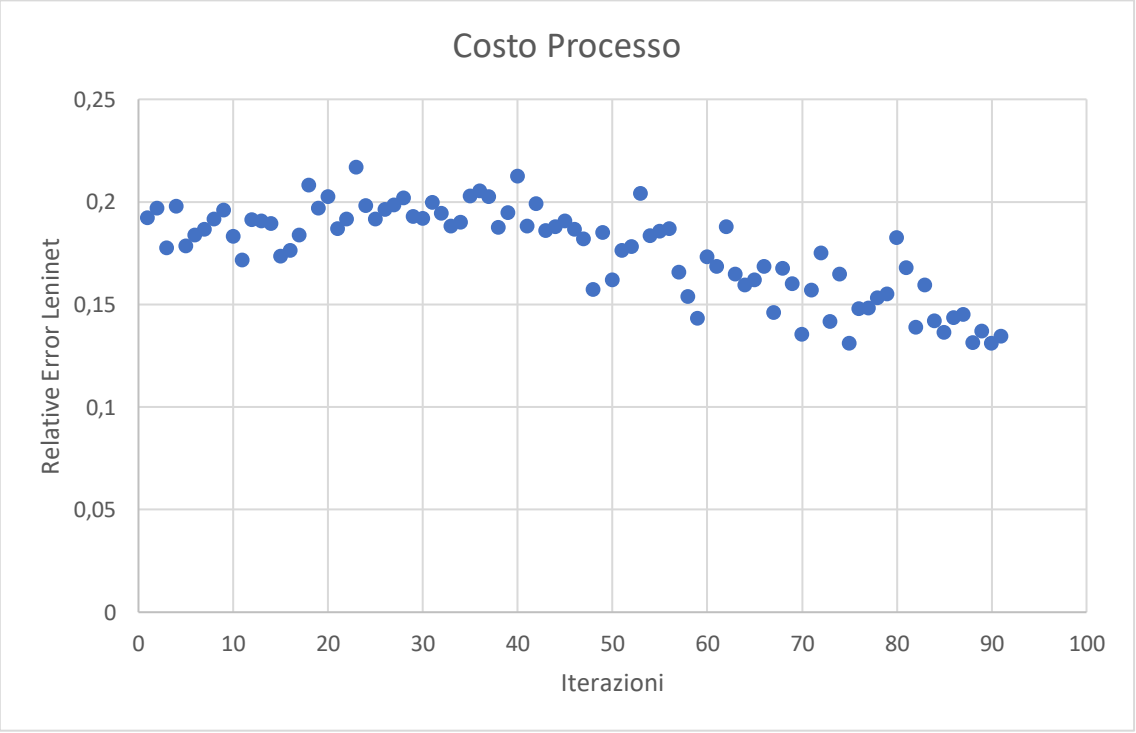

Figura 29:Ottimizzazione parametrica rete neurale (costo processo)

Da questi ultimi due grafici si nota come un aumento del numero di cicli di addestramento non sempre porta alla riduzione dell'errore. L'obiettivo di questo studio

di ottimizzazione parametrica è stato individuare il numero di cicli ottimo per addestrare la rete neurale sia nel caso del costo del materiale che nel caso del costo del processo. Dalla figura 23 si nota come l'iterazione 58, caratterizzata da 670 cicli di training, presenta l'errore più basso pari a 0,164 ovvero il 16,4%.
Invece, dalla figura 24 il caso migliore è dato dall'iterazione numero 90 che presenta 990 cicli di addestramento e il 13,1% di relative error leninet.
Questi risultati sono stati ottenuti da dati di training; in particolare, è stata utilizzata la tecnica del cross-validation con l'obiettivo di generalizzare il modello di machine learning evitando il problema dell'overfitting.
Nei paragrafi successivi i modelli sono stati testati con dati nuovi di test.

5.4.2 Deep learning

L'algoritmo "deep learning" è in grado di lavorare con un database misto, costituito sia da dati qualitativi che quantitativi.
In questo caso, il numero di neuroni nascosti non può essere determinato con una regola empirica ma necessita di un processo di ottimizzazione parametrica. Anche se il deep learning può avere più di due starti nascosti, in questo studio ne sono stati considerati soltanto due con un numero minimo di neuroni pari a 5 e un numero massimo pari a 50 (per strato).
L'utilizzo dell'algoritmo "deep learning" su RapidMiner presenta alcuni vantaggi rispetto al semplice "neural network":

- Può lavorare con un database misto;
- È in grado di determinare autonomamente il numero di epoche di training ottimale (tramite la discesa stocastica del gradiente). Quindi non c'è bisogno di fare l'ottimizzazione parametrica per trovare il numero di cicli di training ottimale;
- È in grado di determinare autonomamente alcuni parametri secondari caratterizzanti le reti neurali. In particolare, mentre nella semplice rete neurale questi parametri sono stati lasciati con il loro valore di default, nel caso del deep learning sono stati calcolati di caso in caso abilitando "adaptive rate".

Nonostante questi vantaggi, l'utilizzo di un unico strato nascosto e quindi della semplice rete neurale potrebbe dare risultati migliori (dipende dal tipo di problema).

In questo caso l'ottimizzazione parametrica è stata fatta prendendo come parametri di riferimento il numero di neuroni nascosti nel primo e nel secondo strato nascosto. Sono state calcolate diverse iterazioni con corrispondente valore di relative error leninet e anche in questo caso i risultati sono stati graficati (Figure 30-31).

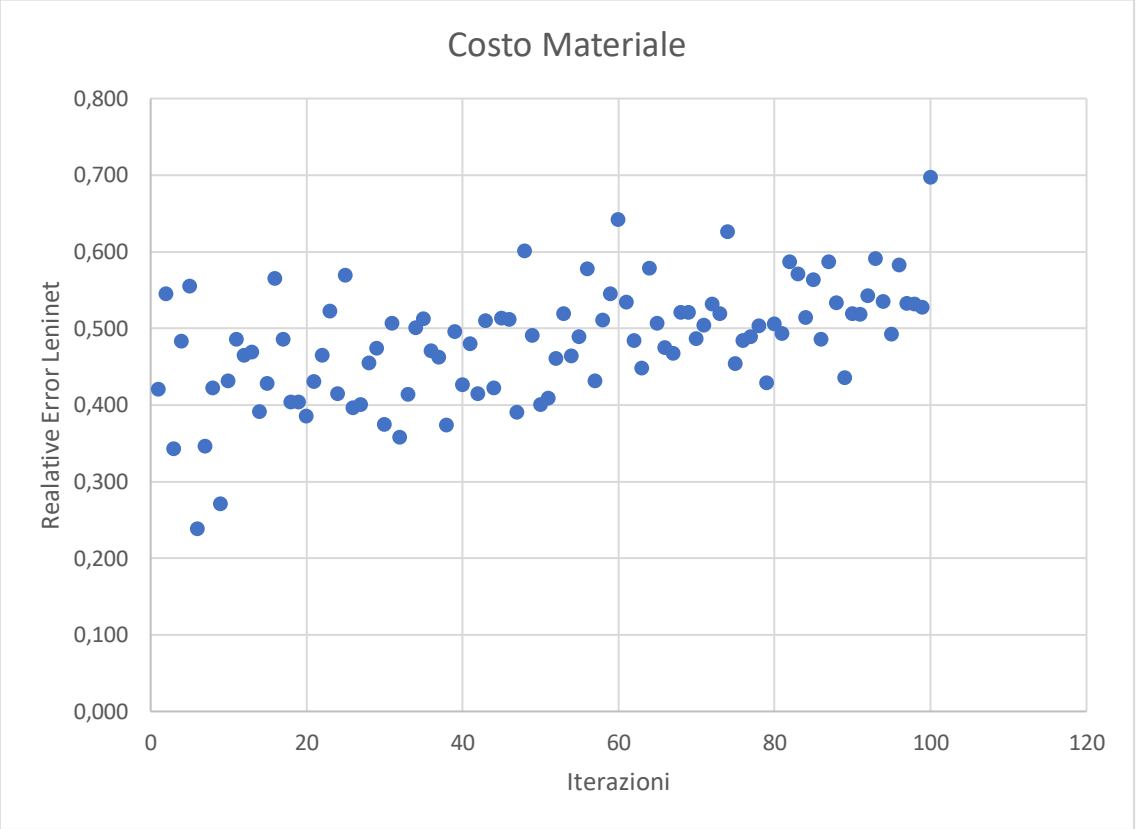

Figura 30: Ottimizzazione parametrica deep learning (costo materiale)

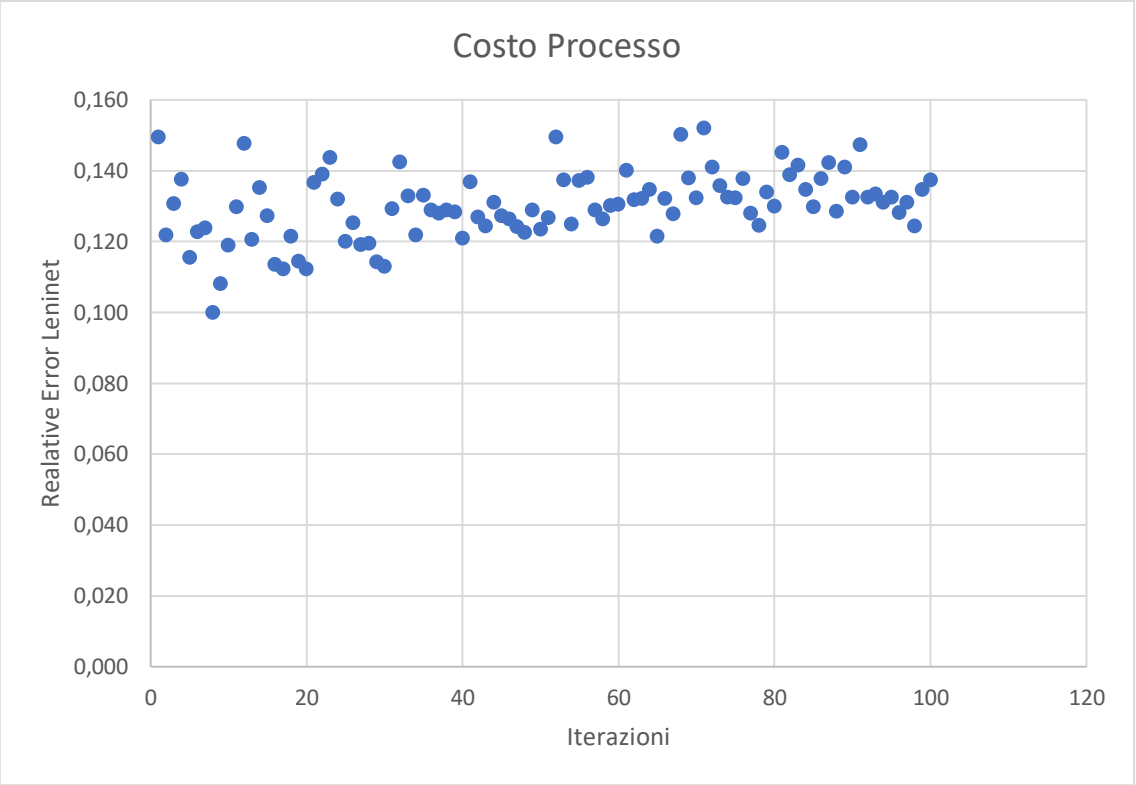

Figura 31: Ottimizzazione parametrica deep learning (costo processo)

In questo caso è difficile interpretare i risultati ottenuti: al cambiare del numero di neuroni nascosti, si ottengono risultati diversi e non è possibile prevedere a priori la configurazione migliore.

Per questo, l'ottimizzazione parametrica è di grande aiuto per individuare i valori ottimi di neuroni nascosti che nel caso del costo del materiale sono 5 per il layer 1 e 30 per il layer 2 (iterazione 6). L'errore è 0,239 ovvero circa 24%.

Mentre, per quanto riguarda il costo del processo, il caso ottimo si ha con l'iterazione numero 8: 40 neuroni nel layer 1 e 5 neuroni nel layer 2. L'errore è pari a circa il 10%.

Anche per il deep learning, questi risultati sono stati trovati applicando il cross-validation ai dati di training.

Nel capitolo successivo sono stati fatti altri test per valutare il modello con dati nuovi.

5.4.3 Random Forest

L'ultimo modello che è stato implementato è il random forest.

Il database iniziale è misto, quindi è in grado di analizzare sia variabili qualitative che quantitative.

I parametri caratteristici che sono stati studiati per l'ottimizzazione parametrica cambiano rispetto alla rete neurale e al deep learning e sono il numero di alberi (deboli) e la profondità dell'albero o maximal depth.

Il primo indica il numero di alberi deboli generati dalla foresta casuale per prevedere il valore di output: il risultato finale è dato dalla media della previsione di ciascun albero debole.

Questo parametro è direttamente proporzionale al numero di sottoinsiemi generati nel set di training in quanto ad ognuno di essi corrisponde un albero decisionale debole.

Per quanto riguarda il maximal depth, RapidMiner permette di utilizzare un valore "speciale": inserendo "-1", tutti gli alberi vengono costruiti fino a quando non vengono soddisfatti altri criteri di arresto.

Di seguito sono state riportati i grafici relativi ai due casi (Figura 32-33)

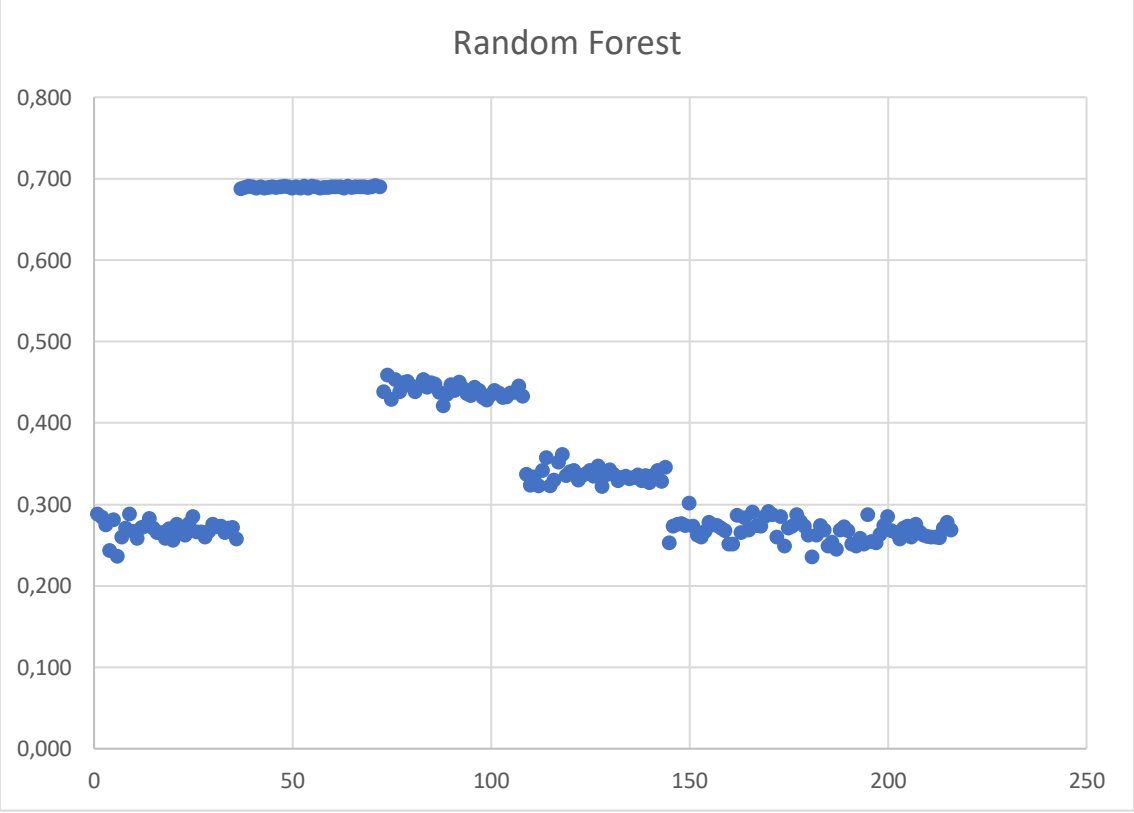

Figura 32: Ottimizzazione parametrica random forest (costo materiale)

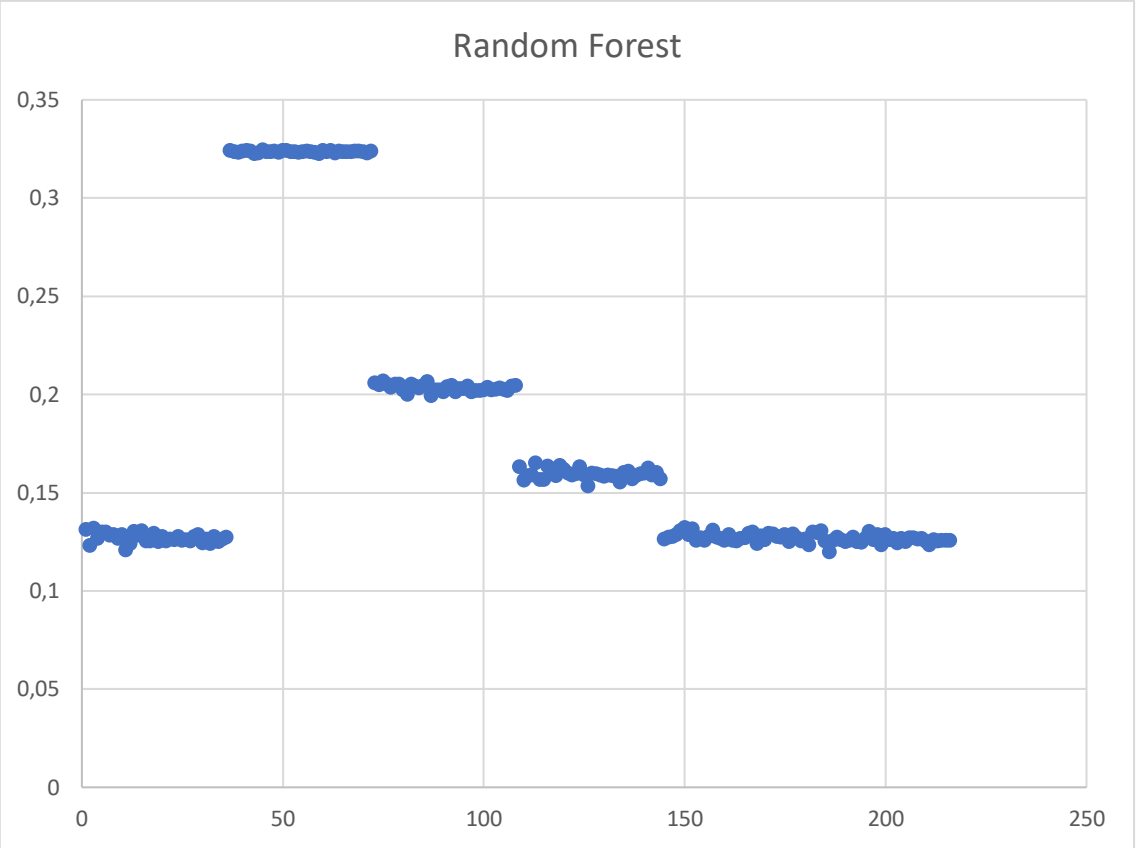

Figura 33: Ottimizzazione parametrica random forest (costo processo)

L'andamento dei punti nei due casi di "costo materiale" e "costo processo" è molto simile.

In entrambi i grafici si possono individuare delle nuvole di punti: ciascuna di esse corrisponde a un valore del maximal depth quindi, aumentanto il valore di questo parametro si ottiene una maggiore precisione nella stima (diminuisce l'errore relativo).

La prima nuvola di punti è relativa al valore della profondità dell'albero di "-1" e restituisce dei buoni risultati.

Per quanto riguarda il parametro numero di alberi, è stato utilizzato un range di valori compreso tra 50 e 400 e si nota come i risultati sono poco dipendenti da questo parametro (il valore dell'relative error leninet è all'incirca lo stesso).

Si potrebbe concludere nel caso del random forest che l'ottimizzazione parametrica non è estremamente necessaria come nel caso della rete neurale semplice e del deep learning in quanto prendendo un valore del maximal depth elevato (ad esempio 10 è un buon valore) si ottengono risultati accettabili indipendentemente dal numero di alberi.

I risultati relativi alle performance sono stati ottenuti con la tecnica cross-validation.

Per quanto riguarda il costo del materiale, l'errore più basso si ottiene con l'iterazione numero 181 caratterizzata da 50 alberi con profondità 10 (l'errore relativo è circa il 23,5%). La stima più precisa nel caso del costo del processo si ottiene con l'iterazione numero 186 con 100 alberi e con profondità 10 (circa 12% di relative error leninet).

5.5 Test e risultati

Terminata la fase di addestramento, i vari modelli sono stati testati con dati nuovi, diversi da quelli di training.

In particolare, sono stati fatti due tipi di prove:

1. Test con dati interni al range di addestramento;
2. Test con dati esterni al range di addestramento.

I dati del primo test derivano dallo split data fatto inizialmente per suddividere il database storico in dati di training e di test.

Nel secondo test, invece, è stato considerato un nuovo disco (codice finito "L") che presenta dati fuori range rispetto a quelli utilizzati per addestrare i modelli.

Iniziando dal primo caso, nelle seguenti tabelle non è stata riportata la colonna "prediction" (che rappresenta il valore di output dei diversi metodi) per motivi di riservatezza aziendale.

Non è stata oscurata, invece, la colonna relativa al MAPE che rappresenta la metrica utilizzata per valutare le performance (capitolo 3.3, equazione 4).

Sia nel caso del costo del materiale che nel costo del processo, anche per i dati di test sono stati costruiti due database (uno numerico e uno misto).

Inoltre, escludendo il dato di input "lotto" nel caso del costo del materiale, così come accadeva nel training, anche i dati di test sono notevolmente ridotti rispetto al costo del processo.

Di seguito i risultati relativi al costo materiale:

Regressione lineare										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**d**	**s**	**Peq [Kg]**	**MAPE**
0	1	0	0	0	0	1	A	AA	LL	45,0%
0	0	1	0	0	1	0	A	AA	MM	22,4%
0	0	1	0	0	0	1	A	AA	MM	42,0%
0	0	1	0	0	0	1	F	GG	LLL	11,0%
0	0	0	1	0	0	1	F	GG	LLL	21,6%
0	0	0	0	1	0	1	H	II	Q	10,0%
0	0	0	1	0	0	1	H	II	M	16,4%
0	0	0	1	0	0	1	H	EE	PP	38,1%
MAPE										**25,8%**
Deviazione Standard										**13,0%**

Tabella 8: Risultati regressione lineare (costo materiale)

Rete neurale semplice										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**d**	**s**	**Peq [Kg]**	**MAPE**
0	1	0	0	0	0	1	A	AA	LL	5,77%
0	0	1	0	0	1	0	A	AA	MM	2,18%
0	0	1	0	0	0	1	A	AA	MM	8,45%
0	0	1	0	0	0	1	F	GG	LLL	16,96%
0	0	0	1	0	0	1	F	GG	LLL	12,75%
0	0	0	0	1	0	1	H	II	Q	40,83%
0	0	0	1	0	0	1	H	II	M	13,69%
0	0	0	1	0	0	1	H	EE	PP	5,88%
MAPE										**13,31%**
Deviazione Standard										**11,35%**

Tabella 9: Risultati rete neurale semplice (costo materiale)

Deep Learning					
d	**s**	**M**	**Grezzo di partenza**	**Peq [Kg]**	**MAPE**
A	AA	M2	Forgiato a stampo chiuso	LL	18,33%
A	AA	M3	Forgiato a stampo aperto	MM	4,85%
A	AA	M3	Forgiato a stampo chiuso	MM	13,23%
F	GG	M3	Forgiato a stampo chiuso	LLL	15,70%
F	GG	M4	Forgiato a stampo chiuso	LLL	3,69%
H	II	M5	Forgiato a stampo chiuso	Q	4,36%
H	II	M4	Forgiato a stampo chiuso	M	8,36%
H	EE	M4	Forgiato a stampo chiuso	PP	29,37%
MAPE					**12,24%**
Deviazione Standard					**8,28%**

Tabella 10: Risultati deep learning (costo materiale)

Random Forest					
d	**s**	**M**	**Grezzo di partenza**	**Peq [Kg]**	**MAPE**
A	AA	M2	Forgiato a stampo chiuso	LL	21,63%
A	AA	M3	Forgiato a stampo aperto	MM	14,87%
A	AA	M3	Forgiato a stampo chiuso	MM	10,16%
F	GG	M3	Forgiato a stampo chiuso	LLL	15,95%
F	GG	M4	Forgiato a stampo chiuso	LLL	2,70%
H	II	M5	Forgiato a stampo chiuso	Q	20,49%
H	II	M4	Forgiato a stampo chiuso	M	38,65%
H	EE	M4	Forgiato a stampo chiuso	PP	24,88%
MAPE					**18,67%**
Deviazione Standard					**10,00%**

Tabella 11: Risultati Random Forest (costo materiale)

Per quanto riguarda i risultati del costo del processo, i dati di test sono più numerosi data l'assenza di duplicati.

Questa differenza tra i due database comporta una maggiore affidabilità dei risultati ottenuti con il "costo processo" rispetto al "costo materiale" in quanto il numero di records che sono stati utilizzati per effettuare il test è nettamente superiore.

Di seguito sono riportate le tabelle relative al "costo processo" e, come nel caso precedente, la colonna relativa ai costi è stata oscurata mentre quella relativa al MAPE è visibile e può essere utilizzata per le considerazioni finali.

Regressione lineare										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**Lotto**	**Peq [Kg]**	**Veq [m3]**	**MAPE**
1	0	0	0	0	1	0	1	L	O	2,55%
1	0	0	0	0	1	0	10	L	O	33,26%
1	0	0	0	0	0	1	5	L	O	21,85%
1	0	0	0	0	0	1	20	L	O	11,25%
0	1	0	0	0	1	0	1	L	O	11,44%
0	1	0	0	0	0	1	50	L	O	94,93%
0	0	1	0	0	1	0	1	MM	O	2,11%
1	0	0	0	0	1	0	1	P	SS	9,20%
0	0	0	0	1	1	0	1	T	Z	24,33%
0	0	0	0	1	0	1	5	T	Z	2,00%
0	0	1	0	0	1	0	5	LL	OO	29,02%
0	1	0	0	0	0	1	1	PP	SS	46,76%
0	0	1	0	0	1	0	5	LLL	ZZ	0,51%
0	0	1	0	0	0	1	10	LLL	ZZ	12,70%
0	0	0	0	1	1	0	5	Q	ZZ	26,33%
0	1	0	0	0	0	1	1	Q	ZZ	6,06%
0	0	0	1	0	1	0	1	LLL	ZZZ	33,48%
1	0	0	0	0	1	0	5	MM	ZZZ	32,96%
0	0	1	0	0	1	0	5	LLL	ZZZ	6,72%
0	1	0	0	0	0	1	10	Q	ZZZZ	18,96%
1	0	0	0	0	0	1	10	Q	ZZZZ	29,37%
0	0	0	0	1	0	1	5	Q	ZZZZ	19,85%
0	0	0	1	0	0	1	5	RR	ZZZZ	15,05%
0	0	0	0	1	1	0	10	T	SS	19,08%
MAPE										**21,24%**
Deviazione Standard										**19,47%**

Tabella 12: Risultati Regressione lineare (costo processo)

Rete neurale semplice										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**Lotto**	**Peq [Kg]**	**Veq [m3]**	**MAPE**
1	0	0	0	0	1	0	1	L	O	12,05%
1	0	0	0	0	1	0	10	L	O	4,61%
1	0	0	0	0	0	1	5	L	O	17,68%
1	0	0	0	0	0	1	20	L	O	11,02%
0	1	0	0	0	1	0	1	L	O	14,22%
0	1	0	0	0	0	1	50	L	O	7,47%
0	0	1	0	0	1	0	1	MM	O	2,54%
1	0	0	0	0	1	0	1	P	SS	24,19%
0	0	0	0	1	1	0	1	T	Z	13,82%
0	0	0	0	1	0	1	5	T	Z	22,69%
0	0	1	0	0	1	0	5	LL	OO	16,61%
0	1	0	0	0	0	1	1	PP	SS	5,61%
0	0	1	0	0	1	0	5	LLL	ZZ	9,57%
0	0	1	0	0	0	1	10	LLL	ZZ	7,90%
0	0	0	0	1	1	0	5	Q	ZZ	17,20%
0	1	0	0	0	0	1	1	Q	ZZ	10,80%
0	0	0	1	0	1	0	1	LLL	ZZZ	6,44%
1	0	0	0	0	1	0	5	MM	ZZZ	18,90%
0	0	1	0	0	1	0	5	LLL	ZZZ	0,31%
0	1	0	0	0	0	1	10	Q	ZZZZ	3,07%
1	0	0	0	0	0	1	10	Q	ZZZZ	9,35%
0	0	0	0	1	0	1	5	Q	ZZZZ	11,96%
0	0	0	1	0	0	1	5	RR	ZZZZ	7,48%
0	0	0	0	1	1	0	10	T	SS	12,74%
MAPE										**11,18%**
Deviazione Standard										**6,09%**

Tabella 13: Risultati rete neurale semplice (costo processo)

Deep learning					
Materiale	**Lotto**	**Grezzo di partenza**	**Peq [Kg]**	**Veq [m3]**	**MAPE**
M1	1	Forgiato a stampo aperto	L	O	15,66%
M1	10	Forgiato a stampo aperto	L	O	2,49%
M1	5	Forgiato a stampo chiuso	L	O	14,47%
M1	20	Forgiato a stampo chiuso	L	O	3,53%
M2	1	Forgiato a stampo aperto	L	O	8,62%
M2	50	Forgiato a stampo chiuso	L	O	8,32%
M3	1	Forgiato a stampo aperto	MM	O	1,66%
M1	1	Forgiato a stampo aperto	P	SS	27,04%
M5	1	Forgiato a stampo aperto	T	Z	7,81%
M5	5	Forgiato a stampo chiuso	T	Z	14,46%
M3	5	Forgiato a stampo aperto	LL	OO	10,60%
M2	1	Forgiato a stampo chiuso	PP	SS	8,85%
M3	5	Forgiato a stampo aperto	LLL	ZZ	11,42%
M3	10	Forgiato a stampo chiuso	LLL	ZZ	4,21%
M5	5	Forgiato a stampo aperto	Q	ZZ	12,17%
M2	1	Forgiato a stampo chiuso	Q	ZZ	4,51%
M4	1	Forgiato a stampo aperto	LLL	ZZZ	4,68%
M1	5	Forgiato a stampo aperto	MM	ZZZ	23,21%
M3	5	Forgiato a stampo aperto	LLL	ZZZ	6,84%
M2	10	Forgiato a stampo chiuso	Q	ZZZZ	7,79%
M1	10	Forgiato a stampo chiuso	Q	ZZZZ	5,29%
M5	5	Forgiato a stampo chiuso	Q	ZZZZ	12,50%
M4	5	Forgiato a stampo chiuso	RR	ZZZZ	12,12%
M5	10	Forgiato a stampo aperto	T	SS	15,54%
MAPE					**10,16%**
Deviazione Standard					**6,10%**

Tabella 14: Risultati deep learning (costo processo)

Random Forest					
Materiale	**Lotto**	**Grezzo di partenza**	**Peq [Kg]**	**Veq [m3]**	**MAPE**
M1	1	Forgiato a stampo aperto	L	O	2,69%
M1	10	Forgiato a stampo aperto	L	O	2,08%
M1	5	Forgiato a stampo chiuso	L	O	13,17%
M1	20	Forgiato a stampo chiuso	L	O	3,03%
M2	1	Forgiato a stampo aperto	L	O	2,80%
M2	50	Forgiato a stampo chiuso	L	O	6,81%
M3	1	Forgiato a stampo aperto	MM	O	13,66%
M1	1	Forgiato a stampo aperto	P	SS	1,11%
M5	1	Forgiato a stampo aperto	T	Z	0,40%
M5	5	Forgiato a stampo chiuso	T	Z	20,26%
M3	5	Forgiato a stampo aperto	LL	OO	26,19%
M2	1	Forgiato a stampo chiuso	PP	SS	5,09%
M3	5	Forgiato a stampo aperto	LLL	ZZ	3,49%
M3	10	Forgiato a stampo chiuso	LLL	ZZ	8,89%
M5	5	Forgiato a stampo aperto	Q	ZZ	11,80%
M2	1	Forgiato a stampo chiuso	Q	ZZ	8,42%
M4	1	Forgiato a stampo aperto	LLL	ZZZ	20,16%
M1	5	Forgiato a stampo aperto	MM	ZZZ	5,97%
M3	5	Forgiato a stampo aperto	LLL	ZZZ	3,99%
M2	10	Forgiato a stampo chiuso	Q	ZZZZ	8,04%
M1	10	Forgiato a stampo chiuso	Q	ZZZZ	1,07%
M5	5	Forgiato a stampo chiuso	Q	ZZZZ	9,26%
M4	5	Forgiato a stampo chiuso	RR	ZZZZ	14,62%
M5	10	Forgiato a stampo aperto	T	SS	18,03%
MAPE					**8,79%**
Deviazione Standard					**6,93%**

Tabella 15: Risultati random forest (costo processo)

Riguardo al MAPE, nelle ultime due righe delle tabelle appena presentate, sono stati riportati due valori:

- Il valore in alto indica l'errore risultante dato dalla media dei contributi dei singoli dati di test;
- Il valore in basso indica la deviazione standard.

I risultati ottenuti nel caso del costo materiale sono meno precisi rispetto al costo processo. I motivi di questa differenza sono attribuibili a:

- Diverso numero di records dei due database iniziali;

- Natura dei dati: sono stati considerati materiali molto diversi tra loro (in termini di costo unitario).

In merito a quest'ultimo punto, in figura 34 è stata rappresentata la regressione lineare semplice relativa ai materiali presi singolarmente.

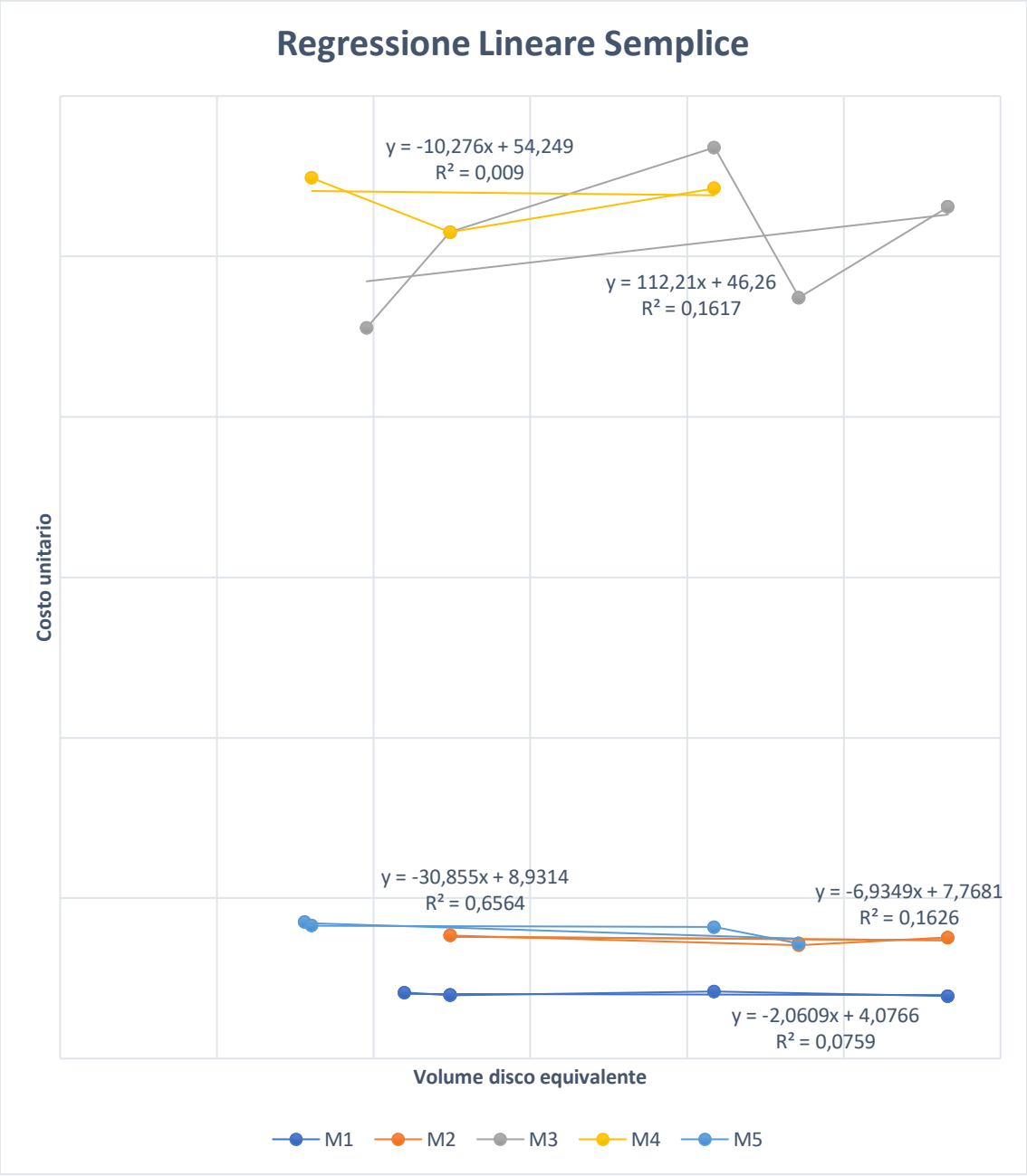

Figura 34: Regressione lineare semplice

Come si può notare, sono state tracciate cinque curve, ciascuna inerente a un tipo di materiale.

Inoltre, è stata realizzata una regressione lineare semplice caratterizzata da una relazione tra una variabile dipendente (costo unitario del materiale) e una variabile indipendente (volume disco equivalente).

Per motivi di riservatezza non sono stati riportati i valori esatti dei costi unitari ma, si riesce comunque a vedere chiaramente come i costi dei materiali M3 e M4 siano più elevati rispetto a M1, M2 e M5.

Quindi, avendo a disposizione pochi dati e relativi a materiali completamente diversi tra loro e inevitabile avere una stima dei costi meno accurata.

Sebbene anche nel costo del processo il materiale ha un peso importante, in questo caso l'elevato numero di records permette di ottenere stime di costo più precise.

I risultati ottenuti dai diversi metodi sono stati confrontati tra loro per individuare quale di questi ha le migliori performance (Figure 35-36).

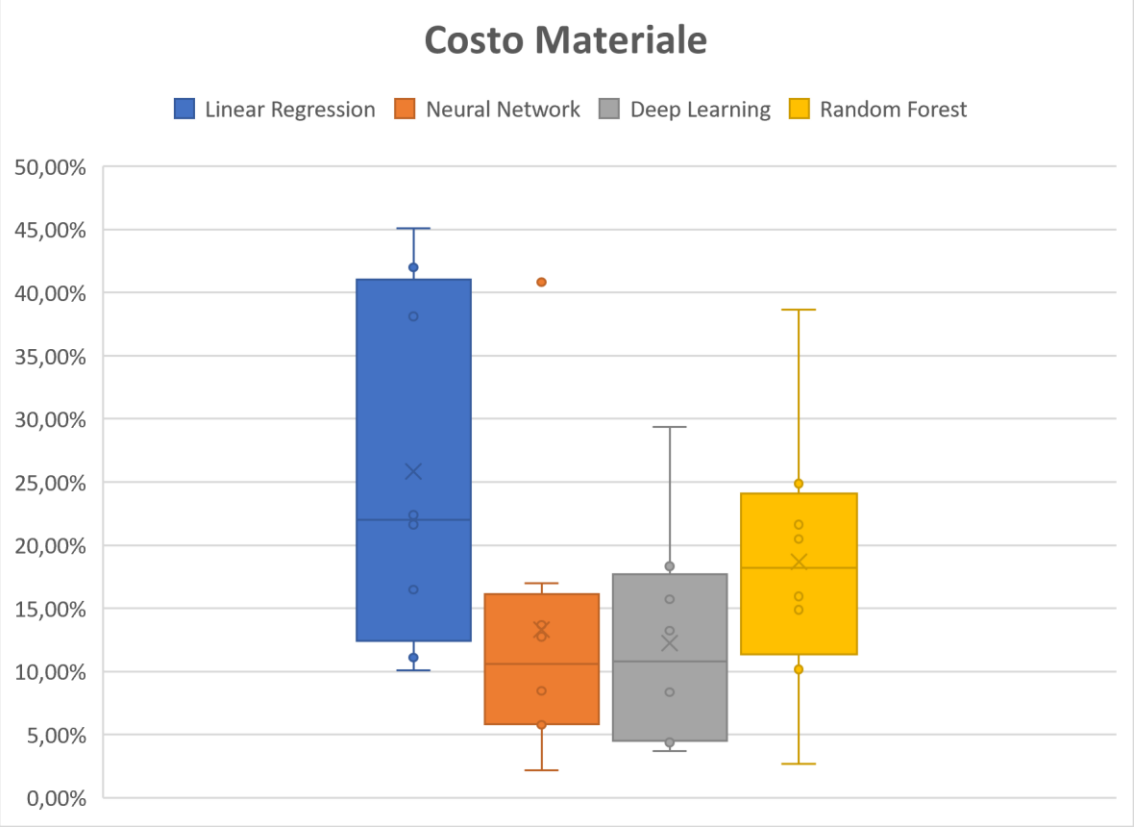

Figura 35: Confronto costo materiale

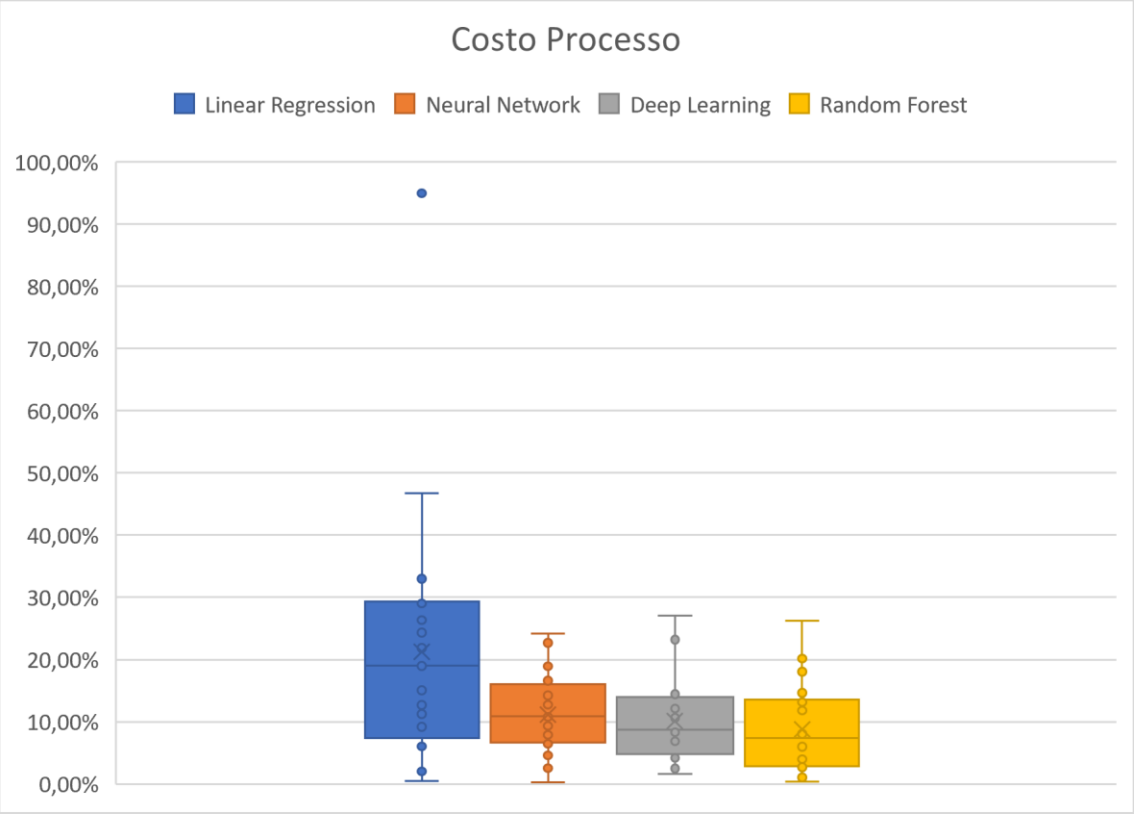

Figura 36: Confronto costo processo

Le figure 35 e 36 rappresentano due grafici a scatola e baffi e mostrano la distribuzione dei dati relativi ai componenti del MAPE di ciascun modello in quartili, evidenziando la mediana e i valori aberranti (outliers):

- I quartili sono quei valori che ripartiscono la popolazione in quattro parti di uguale numerosità:
- I valori aberranti o outliers sono dei valori anomali numericamente distanti dal resto dei dati raccolti. In statistica vengono definiti come valori al di fuori di un certo intervallo e, dato che possono essere fuorvianti, conviene non considerarli.

Nei grafici di figura 35 e 36, i rettangoli sono detti “scatole” mentre le linee che si estendono in verticale sono chiamate “baffi”.

Le scatole rappresentano il range interquartile, ovvero la differenza tra il terzo quartile (bordo superiore) e il primo quartile (bordo inferiore).

All’interno dell’range interquartile, cadono il 50% delle osservazioni (quindi i valori più frequenti sono contenuti in questo range).

La linea interna alle scatole corrisponde alla mediana o al secondo quartile.

Viene considerato valore aberrante (o in inglese outliers) un valore con scostamento positivo dal terzo quartile superiore a 1,5 volte il range interquartile o, simmetricamente, un valore con scostamento negativo dal primo quartile superiore (in valore assoluto) a 1,5 volte il range interquartile.

Invece, i baffi (whiskers) corrispondono al valore minimo (baffo inferiore) e al valore massimo (baffo superiore) osservati dopo avere escluso gli outliers.

Sia nel caso del costo del materiale che nel costo del processo, le tecniche di stima innovative di machine learning sono risultate più precise rispetto alla tecnica tradizionale di regressione lineare.

Nel caso del costo del materiale, le reti neurali risultano più precise del random forest (in particolare il deep learning). Per quanto riguarda il costo del processo, invece, il random forest è leggermente più preciso del deep learning (i due hanno performance molto simili).

Nel complesso si potrebbe sviluppare un tool che utilizzi il deep learning nel caso del costo del materiale e il random forest nel costo del processo.

Oppure, considerando che il deep learning ha comunque ottime prestazioni anche nel secondo caso, si potrebbe sviluppare un tool che utilizzi solamente questo algoritmo.

I grafici appena descritti permettono di testare gli algoritmi su un insieme di dati di test.

Oltre a questa prima verifica, è stato fatto un secondo test su un nuovo set di dati relativi a un nuovo disco (codice finito “L”).

Il nuovo disco fa parte della macchina 2 e nella tabella successiva è stato mostrato il database:

Cod	d	s	M	N° cave	Lotto	Grezzo di partenza	Peq [Kg]	Veq [m3]	Tot
L	I	FF	M2	Y	1	Forgiato a stampo aperto	1P	OOO	€
L	I	FF	M2	Y	5	Forgiato a stampo aperto	1P	OOO	€
L	I	FF	M2	Y	10	Forgiato a stampo aperto	1P	OOO	€
L	I	FF	M3	Y	1	Forgiato a stampo aperto	1Q	OOO	€
L	I	FF	M3	Y	5	Forgiato a stampo aperto	1Q	OOO	€
L	I	FF	M3	Y	10	Forgiato a stampo aperto	1Q	OOO	€
L	I	FF	M1	Y	1	Forgiato a stampo aperto	1R	OOO	€
L	I	FF	M1	Y	5	Forgiato a stampo aperto	1R	OOO	€
L	I	FF	M1	Y	10	Forgiato a stampo aperto	1R	OOO	€
L	I	FF	M2	Y	1	Forgiato a stampo chiuso	1P	OOO	€
L	I	FF	M2	Y	5	Forgiato a stampo chiuso	1P	OOO	€
L	I	FF	M2	Y	10	Forgiato a stampo chiuso	1P	OOO	€
L	I	FF	M3	Y	1	Forgiato a stampo chiuso	1Q	OOO	€
L	I	FF	M3	Y	5	Forgiato a stampo chiuso	1Q	OOO	€
L	I	FF	M3	Y	10	Forgiato a stampo chiuso	1Q	OOO	€
L	I	FF	M1	Y	1	Forgiato a stampo chiuso	1R	OOO	€
L	I	FF	M1	Y	5	Forgiato a stampo chiuso	1R	OOO	€
L	I	FF	M1	Y	10	Forgiato a stampo chiuso	1R	OOO	€

Tabella 16: Database disco "L"

Come si nota nell'appendice, nel database iniziale sono stati inseriti 9 dischi che presentano uno spessore compreso tra un valore massimo pari a AA e un valore minimo EE. Può capitare nella realtà che ci sia la necessità di stimare un costo di un componente caratterizzato da uno o più dati di input esterni al range di addestramento. Ad esempio,

in questo caso, il disco "L" presenta uno spessore pari a FF, valore superiore al range del database iniziale di tabella 2 [EE; AA].

Per questo motivo, è stato fatto un secondo test, diverso dal precedente, in cui è stato utilizzato un disco che presenta valori nuovi e in particolare uno spessore molto elevato ed esterno al range dei dati di training.

Anche in questo caso i dati sensibili sono stati oscurati. Per valutare i risultati, le tabelle di seguito presentano un'ultima colonna con i valori degli scostamenti, ovvero la differenza tra il costo generato dagli algoritmi e il costo reale.

Regressione Lineare										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**d**	**s**	**Peq [Kg]**	**Scostamento**
0	1	0	0	0	1	0	I	FF	1P	232,32%
0	0	1	0	0	1	0	I	FF	1R	-0,32%
1	0	0	0	0	1	0	I	FF	1Q	512,78%
0	1	0	0	0	0	1	I	FF	1P	340,99%
0	0	1	0	0	0	1	I	FF	1R	33,23%
1	0	0	0	0	0	1	I	FF	1Q	723,90%

Tabella 17: Risultati test 2 regressione lineare (costo materiale)

Rete neurale semplice										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**d**	**s**	**Peq [Kg]**	**Scostamento**
0	1	0	0	0	1	0	I	FF	1P	109,76%
0	0	1	0	0	1	0	I	FF	1R	-6,73%
1	0	0	0	0	1	0	I	FF	1Q	212,97%
0	1	0	0	0	0	1	I	FF	1P	113,64%
0	0	1	0	0	0	1	I	FF	1R	19,75%
1	0	0	0	0	0	1	I	FF	1Q	207,48%

Tabella 18: Risultati test 2 rete neurale semplice (costo materiale)

Deep Learning					
d	**s**	**M**	**Grezzo di partenza**	**Peq [Kg]**	**Scostamento**
I	FF	M2	Forgiato a stampo aperto	1P	77,32%
I	FF	M3	Forgiato a stampo aperto	1R	-26,45%
I	FF	M1	Forgiato a stampo aperto	1Q	248,53%
I	FF	M2	Forgiato a stampo chiuso	1P	79,81%
I	FF	M3	Forgiato a stampo chiuso	1R	-4,81%
I	FF	M1	Forgiato a stampo chiuso	1Q	313,61%

Tabella 19: Risultati test 2 deep learning (costo materiale)

Random forest					
d	**s**	**M**	**Grezzo di partenza**	**Peq [Kg]**	**Scostamento**
I	FF	M2	Forgiato a stampo aperto	1P	192,00%
I	FF	M3	Forgiato a stampo aperto	1R	-30,13%
I	FF	M1	Forgiato a stampo aperto	1Q	431,04%
I	FF	M2	Forgiato a stampo chiuso	1P	280,77%
I	FF	M3	Forgiato a stampo chiuso	1R	-9,43%
I	FF	M1	Forgiato a stampo chiuso	1Q	603,89%

Tabella 20: Risultati test 2 random forest (costo materiale)

Anche in questo caso, per quanto riguarda il costo del processo i dati di test presentano più records e i risultati sono più precisi:

Regressione lineare										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**Lotto**	**Peq [Kg]**	**Veq [m3]**	**Scostamento**
0	1	0	0	0	1	0	1	1P	OOO	-2,13%
0	1	0	0	0	1	0	5	1P	OOO	-27,31%
0	1	0	0	0	1	0	10	1P	OOO	-27,72%
0	0	1	0	0	1	0	1	1Q	OOO	19,77%
0	0	1	0	0	1	0	5	1Q	OOO	5,23%
0	0	1	0	0	1	0	10	1Q	OOO	4,77%
1	0	0	0	0	1	0	1	1R	OOO	-25,32%
1	0	0	0	0	1	0	5	1R	OOO	-72,70%
1	0	0	0	0	1	0	10	1R	OOO	-75,85%
0	1	0	0	0	0	1	1	1P	OOO	6,18%
0	1	0	0	0	0	1	5	1P	OOO	-22,89%
0	1	0	0	0	0	1	10	1P	OOO	-23,94%
0	0	1	0	0	0	1	1	1Q	OOO	13,19%
0	0	1	0	0	0	1	5	1Q	OOO	-7,55%
0	0	1	0	0	0	1	10	1Q	OOO	-8,76%
1	0	0	0	0	0	1	1	1R	OOO	-9,30%
1	0	0	0	0	0	1	5	1R	OOO	-59,01%
1	0	0	0	0	0	1	10	1R	OOO	-63,00%

Tabella 21:Risultati test 2 regressione lineare (costo processo)

Rete neurale semplice										
M = M1	**M = M2**	**M = M3**	**M = M4**	**M = M5**	**Grezzo di partenza = Forgiato a stampo aperto**	**Grezzo di partenza = Forgiato a stampo chiuso**	**Lotto**	**Peq [Kg]**	**Veq [m3]**	**Scostamento**
0	1	0	0	0	1	0	1	1P	OOO	-16,43%
0	1	0	0	0	1	0	5	1P	OOO	-26,34%
0	1	0	0	0	1	0	10	1P	OOO	-6,57%
0	0	1	0	0	1	0	1	1Q	OOO	0,17%
0	0	1	0	0	1	0	5	1Q	OOO	-11,67%
0	0	1	0	0	1	0	10	1Q	OOO	-1,97%
1	0	0	0	0	1	0	1	1R	OOO	-55,36%
1	0	0	0	0	1	0	5	1R	OOO	-85,84%
1	0	0	0	0	1	0	10	1R	OOO	-56,84%
0	1	0	0	0	0	1	1	1P	OOO	4,55%
0	1	0	0	0	0	1	5	1P	OOO	-12,65%
0	1	0	0	0	0	1	10	1P	OOO	-1,74%
0	0	1	0	0	0	1	1	1Q	OOO	-2,17%
0	0	1	0	0	0	1	5	1Q	OOO	-16,43%
0	0	1	0	0	0	1	10	1Q	OOO	-3,29%
1	0	0	0	0	0	1	1	1R	OOO	-21,53%
1	0	0	0	0	0	1	5	1R	OOO	-57,96%
1	0	0	0	0	0	1	10	1R	OOO	-43,15%

Tabella 22: Tabella 19: Risultati test 2 rete neurale semplice (costo processo)

Deep learning					
Materiale	**Lotto**	**Grezzo di partenza**	**Peq [Kg]**	**Veq [m3]**	**Scostamento**
M2	1	Forgiato a stampo aperto	1P	OOO	-6,60%
M2	5	Forgiato a stampo aperto	1P	OOO	-14,41%
M2	10	Forgiato a stampo aperto	1P	OOO	-2,02%
M3	1	Forgiato a stampo aperto	1Q	OOO	6,35%
M3	5	Forgiato a stampo aperto	1Q	OOO	-3,55%
M3	10	Forgiato a stampo aperto	1Q	OOO	7,62%
M1	1	Forgiato a stampo aperto	1R	OOO	-38,26%
M1	5	Forgiato a stampo aperto	1R	OOO	-69,96%
M1	10	Forgiato a stampo aperto	1R	OOO	-60,21%
M2	1	Forgiato a stampo chiuso	1P	OOO	-3,61%
M2	5	Forgiato a stampo chiuso	1P	OOO	-14,23%
M2	10	Forgiato a stampo chiuso	1P	OOO	2,03%
M3	1	Forgiato a stampo chiuso	1Q	OOO	-1,79%
M3	5	Forgiato a stampo chiuso	1Q	OOO	-18,53%
M3	10	Forgiato a stampo chiuso	1Q	OOO	-6,85%
M1	1	Forgiato a stampo chiuso	1R	OOO	-26,58%
M1	5	Forgiato a stampo chiuso	1R	OOO	-58,95%
M1	10	Forgiato a stampo chiuso	1R	OOO	-43,75%

Tabella 23: Risultati test 2 deep learning (costo processo)

Random forest					
Materiale	**Lotto**	**Grezzo di partenza**	**Peq [Kg]**	**Veq [m3]**	**Scostamento**
M2	1	Forgiato a stampo aperto	1P	OOO	-12,30%
M2	5	Forgiato a stampo aperto	1P	OOO	-24,85%
M2	10	Forgiato a stampo aperto	1P	OOO	-18,78%
M3	1	Forgiato a stampo aperto	1Q	OOO	15,21%
M3	5	Forgiato a stampo aperto	1Q	OOO	9,80%
M3	10	Forgiato a stampo aperto	1Q	OOO	14,48%
M1	1	Forgiato a stampo aperto	1R	OOO	-37,37%
M1	5	Forgiato a stampo aperto	1R	OOO	-64,10%
M1	10	Forgiato a stampo aperto	1R	OOO	-60,76%
M2	1	Forgiato a stampo chiuso	1P	OOO	-5,81%
M2	5	Forgiato a stampo chiuso	1P	OOO	-22,82%
M2	10	Forgiato a stampo chiuso	1P	OOO	-18,77%
M3	1	Forgiato a stampo chiuso	1Q	OOO	10,73%
M3	5	Forgiato a stampo chiuso	1Q	OOO	0,91%
M3	10	Forgiato a stampo chiuso	1Q	OOO	4,09%
M1	1	Forgiato a stampo chiuso	1R	OOO	-28,02%
M1	5	Forgiato a stampo chiuso	1R	OOO	-62,65%
M1	10	Forgiato a stampo chiuso	1R	OOO	-60,23%

Tabella 24: Risultati test 2 random forest (costo processo)

Nel secondo test non è stato calcolato il MAPE ma è stato tracciato un grafico (per il costo del materiale e per il costo del processo) in cui si confrontano i valori reali e predetti dagli algoritmi (Figura 37-38).

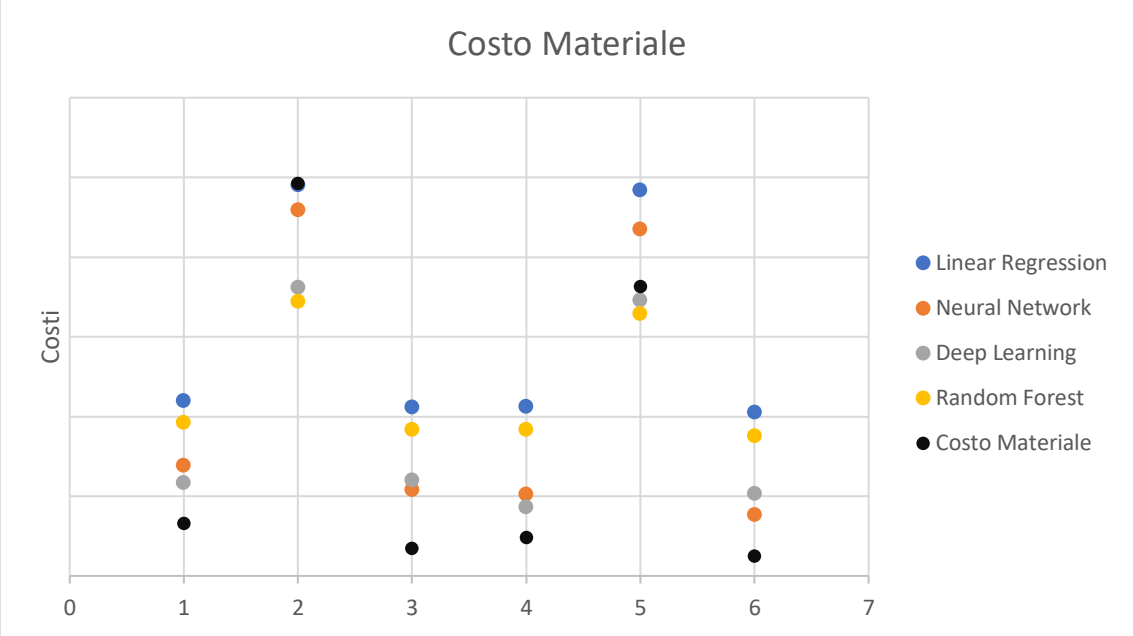

Figura 37: Test 2 costi materiale

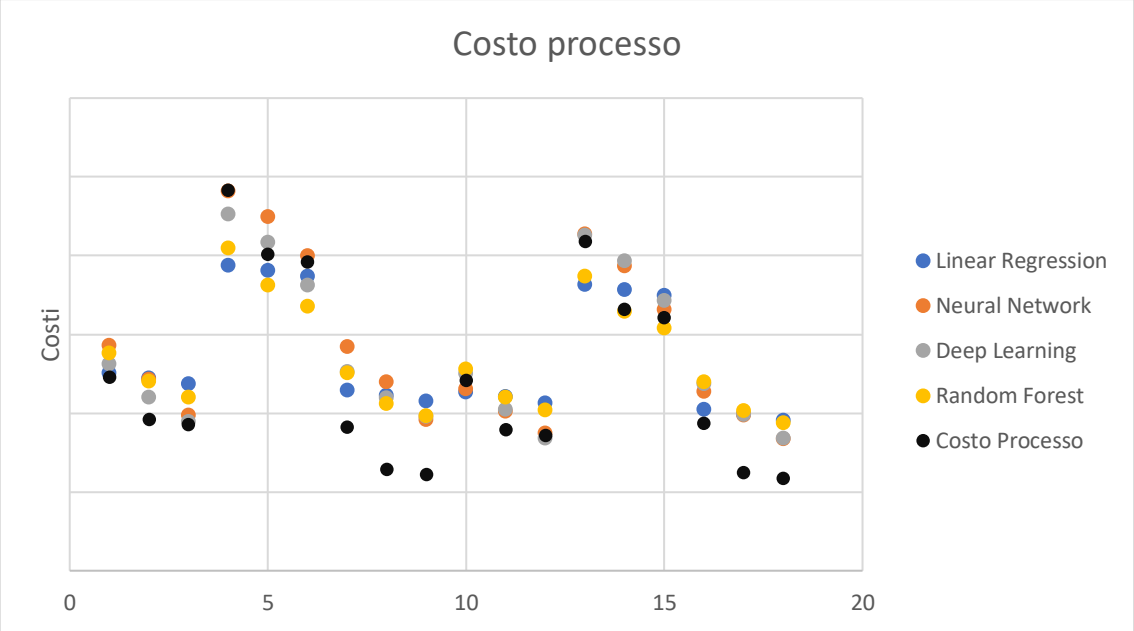

Figura 38: Test 2 costi processo

Nelle figure 38 e 39 sono stati utilizzati dei pallini colorati per rappresentare i valori di costo predetto dagli algoritmi e il costo reale (in nero). Si nota come le reti neurali (colore arancio e grigio) forniscono buone performance in caso di dati esterni al range di partenza.

Nel caso del costo del materiale, il risultato è coerente con quello del primo test. Infatti, le reti neurali riesco a stimare meglio i costi.

I valori degli scostamenti possono essere sia positivi che negativi: nel primo caso si ha una sovrastima in quanto il costo generato dall'algoritmo è maggiore del costo reale. Viceversa, nel secondo caso si ha una sottostima in quanto il costo generato dall'algoritmo è minore del costo reale.

Le tecniche di stima dei costi, applicate al caso del costo del materiale, presentano scostamenti molto elevati.

Nel caso del costo del processo, invece, anche se il random forest nel primo test risultava leggermente più preciso, in questa variante che considera dati di test esterni al range di partenza le reti neurali riescono ad essere più precise rispetto al random forest (non sempre ma nella maggior parte dei casi).

Pertanto, questo caso particolare potrà essere approfondito in successivi lavori per verificare se effettivamente i metodi rete neurale semplice e deep learning risultano sempre più precisi rispetto agli altri algoritmi di machine learning nel caso di dati di test esterni al range di partenza.

Concludendo, le tecniche di stima dei costi di machine learning applicate ai dischi dei compressori assiali si sono rilevate una valida alternativa alla tecnica tradizionale di regressione lineare.

In particolare, nel caso del costo del materiale i risultati sono stati meno performanti data la ridotta quantità di dati di training e la natura dei materiali considerati (costi unitari molto diversi).

Nel caso del costo dei processi, i risultati sono stati soddisfacenti in quanto sono stati rilevati errori relativi molto bassi nel caso di dati di test interni al range di partenza (circa 10%) e anche nel secondo test gli scostamenti sono stati accettabili.

In questo lavoro di tesi si è risposto in parte al requisito R6, dato che l'analisi è stata realizzata soltanto su dati di Should Cost.

Conclusioni

Dalle analisi effettuate risulta evidente che le tecniche innovative di machine learning, oltre ad essere una valida alternativa alla tecnica tradizionale di regressione lineare, presentano performance migliori.

Nonostante questo vantaggio, non sempre si decide di utilizzarle perché hanno il limite di essere considerate delle “black box”: non è possibile dare un'interpretazione teorica dei risultati, soprattutto nel caso di valori imprevisti o ingiustificati.

La regressione lineare, invece, si contraddistingue per essere deducibile da considerazioni tecniche e, di conseguenza, facilmente interpretabile (è una relazione tra parametri dipendenti e indipendenti).

Le tecniche di machine learning, data la loro non linearità, sono in grado di gestire problemi molto complessi in quanto:

- Non è necessario controllare attentamente ogni potenziale cost driver per verificarne la correlazione con l’output
- Non c’è bisogno di fare delle assunzioni iniziali sulla forma delle funzioni di approssimazione.

Per la regressione si riesce ad ottenere una maggiore precisione nei risultati se si semplifica il problema (ad esempio considerando solo un materiale invece di cinque).

Per quanto riguarda le differenze tra il random forest e le reti neurali (semplici e deep learning) è emerso che le prime si realizzano in minor tempo (processo di ottimizzazione parametrica più immediato) e con più facilità.

Tuttavia, nei casi di stima di dati esterni al range di addestramento, le reti neurali si sono dimostrate nella maggior parte dei casi più precise rispetto al random forest.

È evidente quindi che l’algoritmo da utilizzare per stimare i costi in fase di progettazione concettuale è fortemente dipendente dal tipo di problema o più in particolare dalla forma del database in ingresso.

Come sviluppi futuri si potrebbero implementare i metodi lineari e non lineari per altri componenti di turbomacchine (come ad esempio le palette) e si potrebbe fare uno studio su dati di Actual Cost.

Questi ultimi, essendo più casuali e imprevedibili rispetto a dati di Should Cost, andrebbero a testare gli algoritmi per problemi ancora più complessi.

Inoltre, l'utilizzo del metodo Montecarlo potrebbe essere da supporto per generare un database più robusto di dati di actual cost.

Concludendo, il presente lavoro ha permesso di confermare le potenzialità delle tecniche di regressione non lineare di machine learning. Pertanto, lo studio potrà essere approfondito con il fine di sviluppare in tempi brevi un tool in grado di stimare i costi di un nuovo prodotto nella fase di progettazione concettuale e di conseguenza scartare o modificare i progetti non idonei prima che siano state investite significative risorse economiche per la loro realizzazione.

Appendice

cod = codice finito

d = diametro finito

s = spessore finito

M = materiale

R = rugosità

Tratt = presenza trattamenti

Peq = peso equivalente

Psl = peso semilavorato

Pg = peso grezzo

Veq = volume disco equivalente

Macchina	Cod	d	s	M	N° cave	R	Trattamenti	Lotto	Grezzo di partenza	Peq [Kg]	Psl	Pg	V eq [m3]	Tot
Disco Macchina 1	A	AA	AAA	M1	X	I	si	1	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	5	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	10	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	20	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	si	50	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	1	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	5	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	10	Forgiata stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	J	si	20	Forgiata stampo aperto	L	M	N	O	€ -

Disco Macchina 1	A	AA	AAA	M1	X	0,8	si	50	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	1	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	5	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	10	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	20	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	50	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	1	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	5	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	10	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	20	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M1	X	I	no	50	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M2	X	I	no	1	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M2	X	I	no	20	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M2	X	I	no	50	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M2	X	I	no	1	Forgiato a stampo chiuso	0L	0M	0N	O	€ -

Disco Macchina 1	A	AA	AAA	M2	X	I	no	20	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M2	X	I	no	50	Forgiato a stampo chiuso	0L	0M	0N	O	€ -
Disco Macchina 1	A	AA	AAA	M3	X	I	no	1	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M3	X	I	no	10	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M3	X	I	no	1	Forgiato a stampo chiuso	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M3	X	I	no	10	Forgiato a stampo chiuso	L	M	N	O	€ -
Disco Macchina 1	A	AA	AAA	M4	X	I	no	10	Forgiato a stampo aperto	L	M	N	O	€ -
Disco Macchina 1	B	BB	BBB	M1	Y	I	no	1	Forgiato a stampo aperto	P	Q	R	S	€ -
Disco Macchina 1	B	BB	BBB	M1	Y	I	no	20	Forgiato a stampo aperto	P	Q	R	S	€ -
Disco Macchina 1	B	BB	BBB	M1	Y	I	no	1	Forgiato a stampo chiuso	P	Q	R	S	€ -
Disco Macchina 1	B	BB	BBB	M1	Y	I	no	20	Forgiato a stampo chiuso	P	Q	R	S	€ -
Disco Macchina 1	B	BB	BBB	M4	Y	I	no	1	Forgiato a stampo chiuso	P	Q	R	S	€ -
Disco Macchina 1	B	BB	BBB	M4	Y	I	no	20	Forgiato a stampo chiuso	P	Q	R	S	€ -
Disco Macchina 1	C	CC	CCC	M5	Z	I	no	1	Forgiato a stampo aperto	T	U	V	Z	€ -
Disco Macchina 1	C	CC	CCC	M5	Z	I	no	5	Forgiato a stampo aperto	T	U	V	Z	€ -

Disco Macchina 1	C	CC	CCC	M5	Z	I	no	1	Forgiato a stampo chiuso	T	U	V	Z	€ -
Disco Macchina 1	C	CC	CCC	M5	Z	I	no	5	Forgiato a stampo chiuso	T	U	V	Z	€ -
Disco Macchina 1	C	CC	CCC	M4	Z	I	no	1	Forgiato a stampo aperto	T	U	V	Z	€ -
Disco Macchina 1	C	CC	CCC	M4	Z	I	no	5	Forgiato a stampo aperto	T	U	V	Z	€ -
Disco Macchina 1	D	DD	DDD	M3	Z	I	no	1	Forgiato a stampo aperto	LL	M M	NN	OO	€ -
Disco Macchina 1	D	DD	DDD	M3	Z	I	no	5	Forgiato a stampo aperto	LL	M M	NN	OO	€ -
Disco Macchina 1	E	CC	EEE	M2	W	I	no	1	Forgiato a stampo chiuso	PP	QQ	RR	SS	€ -
Disco Macchina 1	E	CC	EEE	M2	W	I	no	10	Forgiato a stampo chiuso	PP	QQ	RR	SS	€ -
Disco Macchina 1	F	EE	FFF	M5	J	I	no	1	Forgiato a stampo aperto	TT	UU	VV	Z	€ -
Disco Macchina 1	F	EE	FFF	M5	J	I	no	10	Forgiato a stampo aperto	TT	UU	VV	Z	€ -
Disco Macchina 1	F	EE	FFF	M1	J	I	no	1	Forgiato a stampo chiuso	TT	UU	VV	Z	€ -
Disco Macchina 1	F	EE	FFF	M1	J	I	no	10	Forgiato a stampo chiuso	TT	UU	VV	Z	€ -
Disco Macchina 2	G	FF	GGG	M3	Z	I	no	1	Forgiato a stampo aperto	LLL	M M M	NN N	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M3	Z	I	no	5	Forgiato a stampo aperto	LLL	M M M	NN N	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M3	Z	I	no	10	Forgiato a stampo aperto	LLL	M M M	NN N	ZZ	€ -

Disco Macchina 2	G	FF	GGG	M3	Z	I	no	1	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M3	Z	I	no	5	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M3	Z	I	no	10	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M5	Z	I	no	1	Forgiata stampo aperto	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M5	Z	I	no	5	Forgiata stampo aperto	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M5	Z	I	no	10	Forgiata stampo aperto	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M2	Z	I	no	1	Forgiata stampo aperto	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M2	Z	I	no	10	Forgiata stampo aperto	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M2	Z	I	no	1	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M2	Z	I	no	10	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M2	Z	I	no	50	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M4	Z	I	no	1	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M4	Z	I	no	10	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	G	FF	GGG	M4	Z	I	no	50	Forgiata stampo chiuso	LLL	MMM	NNN	ZZ	€ -
Disco Macchina 2	H	GG	HHH	M4	Z	I	no	1	Forgiata stampo aperto	LLL	QQQ	RRR	ZZZ	€ -

Disco Macchina 2	H	GG	HHH	M4	Z	I	no	5	Forgiatoa stampo aperto	LLL	QQQ	RRR	ZZZ	€ -
Disco Macchina 2	H	GG	HHH	M1	Z	I	no	1	Forgiatoa stampo aperto	TTT	UUU	VVV	ZZZ	€ -
Disco Macchina 2	H	GG	HHH	M1	Z	I	no	5	Forgiatoa stampo aperto	TTT	UUU	VVV	ZZZ	€ -
Disco Macchina 2	H	GG	HHH	M5	Z	I	no	1	Forgiatoa stampo aperto	TTT	UUU	VVV	ZZZ	€ -
Disco Macchina 2	H	GG	HHH	M5	Z	I	no	5	Forgiatoa stampo aperto	TTT	UUU	VVV	ZZZ	€ -
Disco Macchina 2	H	GG	HHH	M3	Z	I	no	1	Forgiatoa stampo aperto	LLLL	MMMM	NNNN	ZZZ	€ -
Disco Macchina 2	H	GG	HHH	M3	Z	I	no	5	Forgiatoa stampo aperto	LLLL	MMMM	NNNN	ZZZ	€ -
Disco Macchina 2	I	HH	III	M2	W	I	no	1	Forgiatoa stampo chiuso	PPPP	QQQQ	RRRR	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M2	W	I	no	5	Forgiatoa stampo chiuso	PPPP	QQQQ	RRRR	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M2	W	I	no	10	Forgiatoa stampo chiuso	PPPP	QQQQ	RRRR	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M1	W	I	no	1	Forgiatoa stampo chiuso	TTTT	UUUU	VVVV	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M1	W	I	no	5	Forgiatoa stampo chiuso	TTTT	UUUU	VVVV	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M1	W	I	no	10	Forgiatoa stampo chiuso	TTTT	UUUU	VVVV	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M5	W	I	no	1	Forgiatoa stampo chiuso	TTTT	UUUU	VVVV	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M5	W	I	no	5	Forgiatoa stampo chiuso	TTTT	UUUU	VVVV	ZZZZ	€ -

Disco Macchina 2	I	HH	III	M5	W	I	no	10	Forgiato a stampo chiuso	TTTT	UUUU	VVVV	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M4	W	I	no	1	Forgiato a stampo chiuso	1L	1M	1N	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M4	W	I	no	5	Forgiato a stampo chiuso	1L	1M	1N	ZZZZ	€ -
Disco Macchina 2	I	HH	III	M4	W	I	no	10	Forgiato a stampo chiuso	1L	1M	1N	ZZZZ	€ -

Bibliografia

[1] Irene Martinelli, Federico Campi, Emanuele Checcacci, Giulio Marcello Lo Presti, Francesco Pescatori, Antonio Pumo, Michele Germani (2019). «Cost Estimation Method for Gas Turbine in Conceptual Design Phase».

[2] Davide Durastanti, «Metodi e strumenti per il Design to Cost di gruppi».

[3] NASA Cost Estimating Handbook Version 4.0, Appendix C: Cost Estimating Methodologies.

[4] Sergio Cavalieri, Paolo Maccarrone, Roberto Pinto (2004). «Parametric vs. neural networkmodels for the estimation of production costs: A case study in the automotive industry».

[5] Lorenzo Govoni. «L'Overfitting e l'Underfitting nel machine learning». Available Online: https://lorenzogovoni.com/overfitting-e-underfitting-machine-learning/.

[6] Antonio C. Caputo, Pacifico M.Pelagagge (2008)., «Parametric and neural methods for cost estimation of process vessels».

[7] Gregory Puckett (2016). «Development of a parametric cost estimating model for University of Alaska system renovation construction projects».

[8] Tarek Hegazy, Amr Ayed (1998). «Neural network model for parametric cost estimation of highway projects».

[9] Marjan Čeh,Milan Kilibarda, Anka Lisec, Branislav Bajat (2018). «Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments».

[10] Muhammad Waseem Ahmad, MonjurMourshed, YacineRezgui (2017). «Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy».

[11] Kumar Varadarajan, Koushal (2015). «Should-cost analysis a key tool for sourcing and product designers».

[12] Maura Mengoni, Marco Mandolini, Marco Matteucci, Michele Germani (2016). «A scalable Design for Costing platform: a practical case in ball valves industry».

[13] Claudio Favi, Federico Campi, Marco Mandolini, Irene Martinelli, Michele Germani (2019). «Conceptual cost estimation of multistage Axial Compressor modules».