

DIPARTIMENTO DI INGEGNERIA DELL’INFORMAZIONE

Corso di

Machine Learning & Data Mining

Relazione di progetto:

AI4I Predictive Maintenance: Machine Failure Classification

Esaminandi:

Edoardo Coppola - 719599

Andrea Fiori - 719219

Anno Accademico 2020/2021

Introduzione

La presente relazione ha lo scopo di documentare il lavoro svolto durante il progetto previsto dal corso di *Machine Learning & Data Mining*. Sono illustrate le osservazioni, le analisi e le procedure seguite a partire dalle fasi iniziali e preliminari di *data analysis* fino alla valutazione delle prestazioni dei modelli individuati ed addestrati.

Il *dataset* utilizzato [1] contiene record caratterizzati da misurazioni di grandezze fisiche in ambito industriale e delle etichette che specificano per un guasto o per il corretto funzionamento del macchinario in esame. Il compito previsto dal progetto consiste nell’individuazione di un modello capace di classificare nuove istanze, la cui forma è la medesima dei record sopracitati, come “ad alto rischio di guasto” o come “normale funzionamento”. Tale compito è quindi un compito di classificazione binaria. Maggiori dettagli sui record, gli attributi e le etichette relative saranno illustrati in dettaglio nelle sezioni successive.

1. Il dataset: struttura e proprietà

Il *dataset* consiste in diecimila istanze, ciascuna caratterizzata da otto attributi riassunti in tabella 1.

Tabella 1 – *Attributi delle istanze all’interno del dataset*

|  |  |
| --- | --- |
| Nome attributo | Descrizione |
| *Product ID* | Identificativo di un prodotto della lavorazione. È caratterizzato da una lettera tra L, M, H che ne indica la qualità (bassa, media o alta) seguita da un numero di serie |
| *Type* | Il tipo del prodotto realizzato. Coincide con la lettera presente nel *Product ID* |
| *Air Temperature [K]* | Temperatura dell’aria, in Kelvin, realizzata tramite generazione di numeri casuali normalizzati attorno ad una media di 300 K e con una deviazione standard di 2 K |
| *Process Temperature [K]* | Temperatura di processo generata casualmente, normalizzata con una deviazione standard di 1 K e sommata alla temperatura dell’aria più 10 |
| *Rotational Speed [rpm]* | Velocità di rotazione calcolata a partire da una potenza di 2860 W e sovrapposta ad un rumore gaussiano |
| *Torque [Nm]* | Momento torcente i cui valori sono distribuiti in modo normale attorno ad una media di 40 Nm e con una deviazione standard di 10 Nm. Tutti i valori sono non negativi |
| *Tool wear [min]* | Usura del macchinario espressa in minuti |
| *Machine Failure [binary]* | Etichetta che mostra se si è verificato un guasto o meno. ‘1’ se si è verificato, ‘0’ altrimenti |

Il *dataset* è estremamente sbilanciato sicché solamente 339 istanze sono etichettate con ‘1’ (guasti). Questo ha reso molto complicato l’addestramento dei modelli e il raggiungimento di prestazioni dignitose e ha quindi richiesto che venissero prese in considerazioni sia tecniche a livello algoritmico che a livello dei dati stessi. Tali tecniche verranno illustrate nelle sezioni successive.

Per quanto riguarda le proprietà intrinseche dei dati in nostro possesso, possiamo innanzitutto osservare la presenza di alcune forti correlazioni tra coppie di attributi. Ad esempio si nota una correlazione di 0.88 tra la temperatura dell’aria e quella di processo, mentre esiste una correlazione di -0.88 tra la velocità di rotazione e il momento torcente. Tuttavia, questi legami non destano troppo scalpore dato che i valori delle due temperature, seppure ottenuti in modo casuale, sono legati matematicamente mentre dalla fisica si sa che il momento torcente è inversamente proporzionale alla velocità rotazionale (in perfetto accordo col significato proprio di correlazione negativa). Possiamo osservare la figura 1 e la figura 2 per avere una visione d’insieme circa le correlazioni tra tutte le coppie di *feature.* Le matrici illustrate nelle figure successive chiaramente sono simmetriche. Dalla figura 1 è poi possibile vedere anche una distribuzione dei valori della temperatura dell’aria, di processo e del momento torcente *simil-gaussiana* proprio come ci si aspettava.

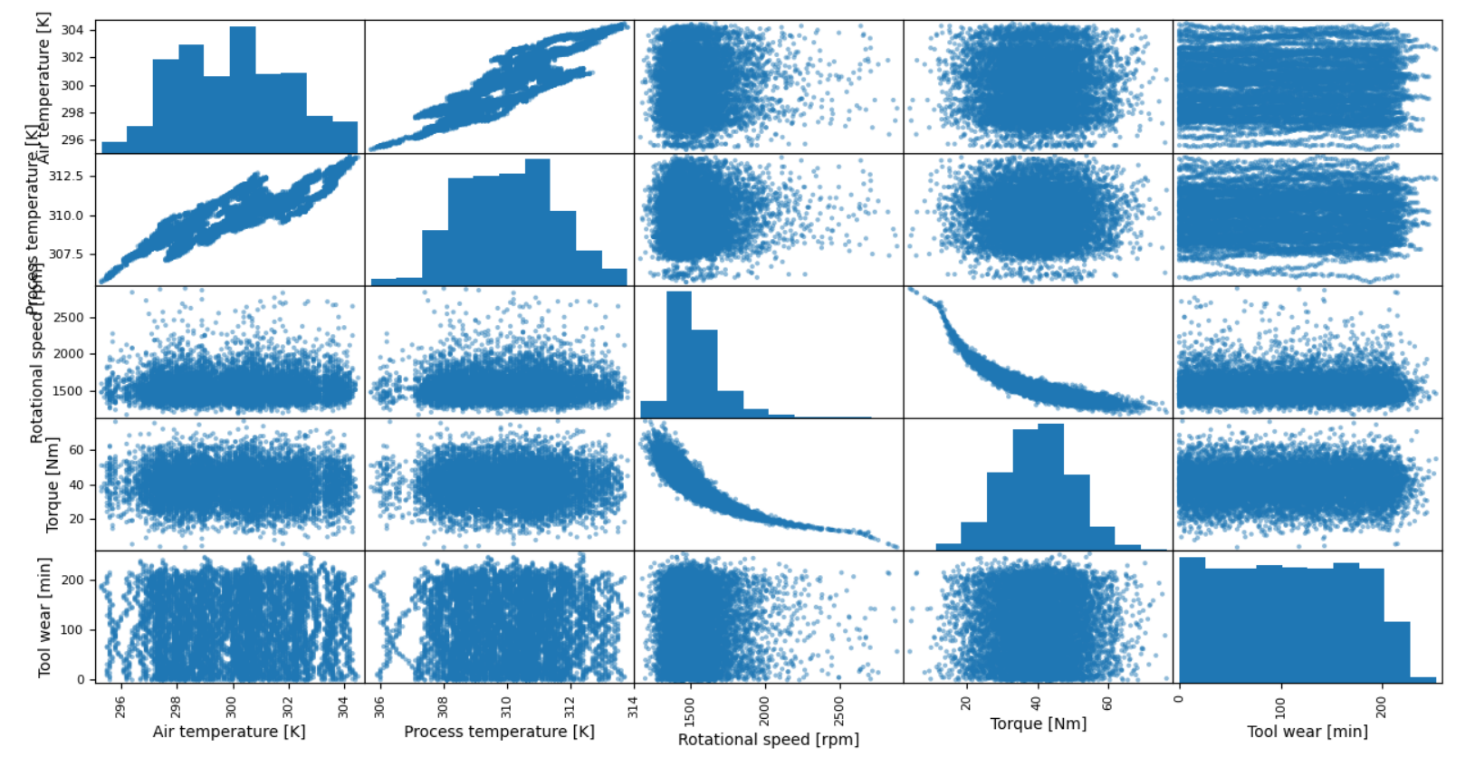
2. Fase preliminare di Data Analysis

Figura 2 – Heatmap delle correlazioni fra tutte le possibili coppie di attributi

Figura 1 – Matrice delle correlazioni in formato scatter fra tutte le possibili coppie di attributi

In questa fase preliminare ci siamo preoccupati della qualità dei dati, della presenza di record duplicati, di eventuali valori non definiti o mancanti, dell’eventuale presenza di *outliers* e di una prima selezione degli attributi rilevanti ai fini del compito da svolgere.

Per prima cosa è stata eliminato l’attributo *Product ID* perché si compone di una parte irrilevante (*serial number*) e di una ridondante (*type*). Difatti, la qualità di un dato prodotto, espressa tramite una lettera fra L, M o H, è già specificata dall’attributo *Type* dello stesso record. Il numero di serie del prodotto realizzato, invece, non è rilevante nel caratterizzare un eventuale guasto del macchinario perché non si conosce il prodotto fisico dietro tale numero.

Il *dataset* originale annoverava anche altri cinque attributi (non riportati in tabella 1) caratterizzanti la tipologia di guasto. Per esempio, l’attributo *HDF* (*Heat Dissipation Failure*) specificava per un guasto dovuto al raggiungimento di elevate temperature ed era stato aggiunto, così come per gli altri quattro, per compiti di *Causal Discovery* che esulano dallo scopo di questa relazione. Per questi motivi, tali attributi, e le relative colonne, sono state eliminate.

Stando a quanto riporta l’articolo [2] associato al *dataset*, reperire sufficienti quantità di dati per compiti di manutenzione predittiva è un’impresa ardua. Infatti, spesso simili informazioni non sono note oppure non sono rese pubbliche. Per questo motivo i record utilizzati dall’autore originale del progetto, e dagli scriventi, hanno una natura spesso artificiale o semi-artificiale. Essi, infatti, provengono dalla condensazione di informazioni ed esperienze accumulate in anni di ricerche su questa tematica. Posta quindi la natura dei dati, la ricerca da noi effettuata per trovare record duplicati non ha prodotto alcun risultato. In modo del tutto analogo, e probabilmente per le stesse ragioni, anche la ricerca di valori non definiti o mancanti non ha prodotto alcun risultato. Questi ultimi due esiti sono quindi rassicuranti circa la completezza dal *dataset.*

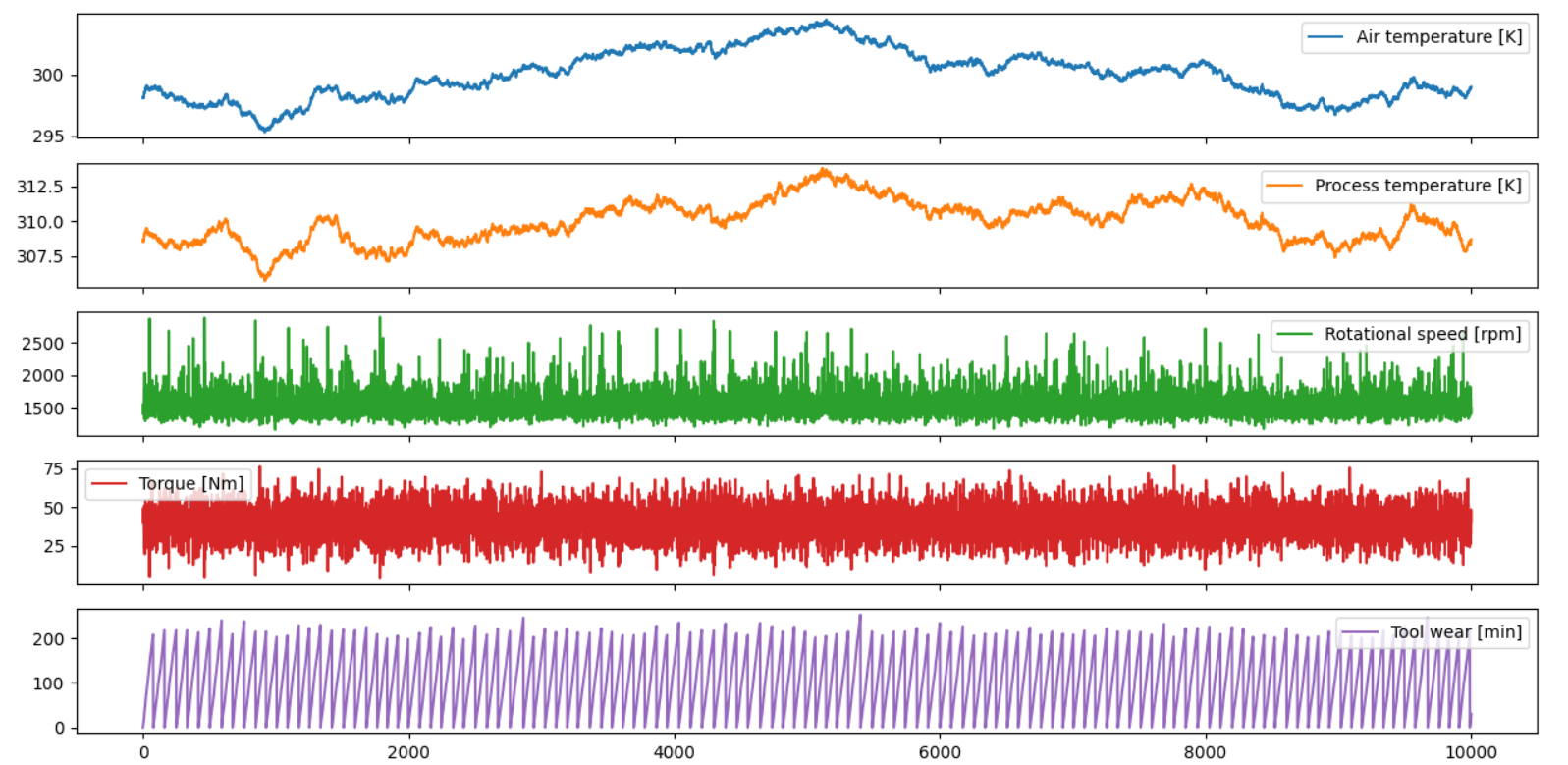
Per verificare la presenza di valori anomali, o *outliers*, sarebbe necessario disegnare un grafico e riportarvi all’interno le istanze del *dataset*. In questo modo sarebbe possibile individuare quali di loro occupano posizioni “anomale” e quindi scartarle. Tuttavia, la dimensionalità dei record non consente una facile rappresentazione pittorica. Di conseguenza, si è optato per una ricerca grafica di valori anomali nei singoli attributi e, come possono confermare i grafici in figura 3, non ne sono stati trovati. L’attributo di categoria *Type* è stato escluso da questa ricerca perché il grafico avrebbe semplicemente riportato una successione di punti etichettati con L, M, o H. Dai grafici sulle temperature dell’aria e di processo non si notano discontinuità marcate né punti di massimo o minimo eccessivamente rilevanti. Questo può essere dovuto alla trasformazione dei dati casuali originali tramite la distribuzione gaussiana accennata in precedenza. Per quanto riguarda la velocità di rotazione, si nota subito che la presenza di punti di massimo o di minimo è in perfetta contrapposizione con la presenza di punti di minimo e di massimo, rispettivamente, nei valori del momento torcente (come ci si aspetta dato che sono grandezza inversamente proporzionali e hanno una forte correlazione negativa). La presenza di valori al di fuori della fascia [1500, 2100] rpm e [20, 60] Nm non è da considerarsi negativa, anzi, è proprio ciò che garantisce rappresentatività all’intero *dataset*. Infatti, spesso elevate velocità di rotazione, combinate anche un’elevata usura del macchinario, sono la causa della rottura dello stesso. Quindi, eliminare simili valori sarebbe un grave errore. L’andamento assunto dai valori di usura (*Tool Wear)* è invece molto regolare.

Figura 3 – Andamento dei valori assunti dagli attributi numerici di tutti i record presenti

3. Pre-processing dei dati

Per quanto concerne le misure adottate nel *pre-processing*, abbiamo optato per una codifica *OneHot* dell’attributo *Type* per via dei modelli che avremmo addestrato in seguito. Infatti, sebbene alberi di decisioni o *random forest* ben si prestano a lavorare anche con attributi non numerici, non si può dire lo stesso per modelli come *support vector machines*. È stata scelta questa codifica e non una *integer encode* per migliorare l’efficienza e le prestazioni dell’addestramento.

Un altro accorgimento è stato normalizzare le scale degli attributi per motivi di ….