Università degli Studi di Milano-Bicocca

**Dipartimento di Informatica, Sistemistica e Comunicazione**

**Corso di Laurea Magistrale in Data Science**

**Esame di Industry Lab:  
Caso di studio Bosch – GP5 Data analysis and leakage coefficient prediction**

**Studenti**

*Mauro Samarelli, 834196*

*Ivan Mera, XXXXXX*

**Anno Accademico 2019-2020**

SOMMARIO

1. Descrizione del processo produttivo 2
2. Import dei dati 4
3. Esplorazione dei dati 5
4. Novelty detection 8
5. Regressione lineare della portata in funzione della pressione 10
6. Conclusioni: risposte ai task xx

**1. Descrizione del processo produttivo**

Il test della linea di assemblaggio della *gear pump GP5* è diviso in due fasi:

* *test a regime*, in cui si rilevano pressione e portata impostando una velocità di 2300 rpm;
* *test di controllo*, in cui si rilevano pressione e portata impostando una velocità di 140 rpm.

Le variabili di interesse per l’analisi dei dati sono le seguenti:

* *programma*: descrizione del programma associato ad una specifica classe di prodotto GP5;
* *velocità a regime*: valore della velocità impostata per il test a regime [rpm];
* *picco e media pressione velocità a regime*: valore del picco e medio della pressione rilevata per il test a regime [bar];
* *picco e media portata velocità a regime*: valore del picco e medio della portata rilevata per il test a regime [L/h];
* *velocità 1*: valore della velocità impostata per il test di controllo [rpm];
* *picco e media pressione velocità 1*: valore del picco e medio della pressione rilevata per il test di controllo [bar];
* *picco e media portata velocità 1*: valore del picco e medio della portata rilevata per il test di controllo [L/h];
* *picco e media coppia finale*: valore del picco e medio della coppia finale rilevati a 100 rpm;
* *temperatura*: temperatura di prova del liquido;
* *esito*: descrizione esito finale della prova.

Di seguito, viene riportata l’equazione lineare che descrive le performance, verificata sperimentalmente da Bosch, che mette in relazione le grandezze di pressione e portata durante il test di controllo (@140rpm, T=40°C).

Immagine che contiene screenshot

Descrizione generata automaticamente

Immagine che contiene mappa, testo, uomo

Descrizione generata automaticamente

Si tratta di un’equazione lineare, descritta dalle seguenti variabili:

* **QGP**è la *portata attuale* del prodotto misurata tramite apposito trasduttore [L/h];
* **QGP, theor** è la *portata teorica* del prodotto, nota a priori [L/h]: essa rappresenta il valore della portata del prodotto, escludendo le perdite dovute alla sua conformazione;
* **α** è il coefficiente di perdita (*leakage coefficient*) espresso in [(L/h)/bar]: si tratta di un indice di performance del prodotto, dipende dalla conformazione del prodotto e rappresenta la perdita di portata in *L/h* per ogni *bar* di pressione;
* **p** è la *pressione outlet* misurata tramite apposito trasduttore [bar].

Di seguito, viene riportata l’associazione delle grandezze descritte con le variabili a disposizione nel database messo a disposizione da Bosch.

|  |  |  |  |
| --- | --- | --- | --- |
| *Grandezza fisica* | *Descrizione* | *Variabile/i database* | *Unità di misura* |
| **Speed** | velocità fase di controllo | velocita\_1 | Rpm |
| **T** | temperatura del liquido | Temperatura | °C |
| **QGP** | portata attuale | picco\_portata\_velocita\_1  media\_portata\_velocita\_1 | L/h |
| **QGP, theor** | portata teorica | *valore noto R&D*: **108,36** | L/h |
| **P** | pressione outlet | picco\_pressione\_velocita\_1  media\_pressione\_velocita\_1 | Bar |
| **α** | coefficiente di perdita | NA \* | (L/h) / bar |

\* da stimare (obiettivo del caso di studio)  
  
Dall’equazione lineare descritta, è possibile notare che il *coefficiente di perdita* **α** è la differenza tra *portata teorica* e *portata attuale* **ΔQ = QGP, theor - QGP**, divisa per la *pressione outlet* **p**.

Indagando su una possibile relazione che consente di esprimere la *portata attuale* **QGP** in funzione della *pressione* *outlet* **p**, è possibile predire il valore del *coefficiente di perdita* **α**, dato che la *portata teorica* **QGP, theor** è un valore costante noto.

Inoltre, la regressione lineare tra la *portata attuale* **QGP** e la *pressione outlet* **p** permette di valutare il cambiamento della portata in funzione del cambiamento della pressione.

**2. Import dei dati**

Il database messo a disposizione da Bosch consiste in 5 file con estensione *.mdb*, riferiti ai dati raccolti dal 2016 al 2020, ognuno dei quali contiene una tabella ESITI con le stesse variabili.

L’import dei dati richiede l’utilizzo del driver ODBC per Microsoft Access, con lo scopo di stabilire la connessione tra l’ambiente di sviluppo Python e i database in formato proprietario.

Dopo aver importato correttamente i dati contenuti nei file, essi sono stati concatenati in unico dataframe le cui dimensioni sono: **296605** righe, **33** colonne.

**3. Esplorazione dei dati**

In modo preliminare all’analisi delle distribuzioni delle variabili, il dataframe è stato filtrato sia per colonne, mantenendo le variabili di interesse descritte nel paragrafo 1, sia per righe settando i seguenti filtri:

* esito = “PEZZO BUONO”
* velocita\_1 = 140
* velocita\_a\_regime = 2300
* pressione\_1 > 0 e portata\_1 > 0
* pressione\_regime > 0 e portata\_regime > 0

Successivamente, è stata effettuata la ricerca di missing values, rilevandone l’assenza.

Sapendo che la variabile “*Programma*” raggruppa una specifica classe di prodotti GP5, è stata ricavata la numerosità delle osservazioni per ogni gruppo:

Immagine che contiene testo, ricevuta

Descrizione generata automaticamente

Si è scelto di indagare sulle relazioni tra le variabili per quanto riguarda le classi di prodotti GP5 aventi più di 1000 osservazioni all’interno del dataframe.

Si tratta dei seguenti 7 gruppi:   
*18\_GP5\_910\_CW.DBF 17\_GP5\_430\_CCW.DBF 16\_GP5\_500\_CCW.DBF 15\_GP5\_430\_CCW.DBF 14\_GP5\_340\_D1.DBF 13\_GP5\_500\_CW.DBF 12\_GP5\_430B\_D1.DBF*

Tale scelta può essere comunque estesa anche a classi con numerosità inferiore, oppure attendere la raccolta di un campione più significativo.

Per ogni classe di prodotto GP5, è stata effettuata la *rimozione degli outliers univariati*, individuati per ogni variabile di interesse descritta nel paragrafo 1: sono state rimosse le osservazioni aventi una variabile con valore assoluto dello *Z-Score* maggiore di 3.5.

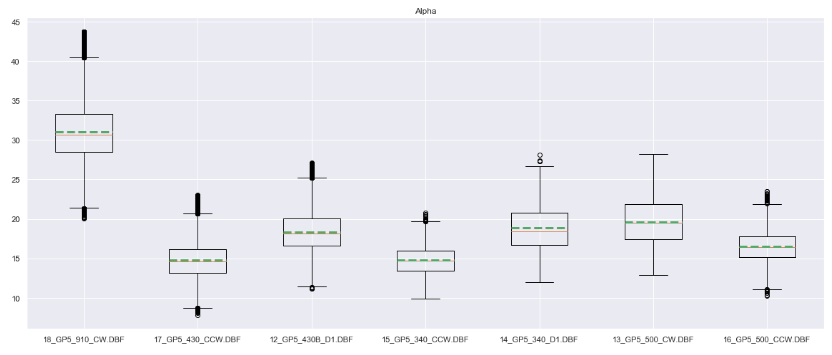
Immagine che contiene ricevuta, testo

Descrizione generata automaticamente

Di seguito vengono riportati i grafici delle distribuzioni delle variabili relativi al gruppo *18\_GP5\_910\_CW.DBF*, avente 227552 osservazioni, ovvero il 77% delle osservazioni totali.

*Immagine che contiene barca, acqua, galleggiante, nave

Descrizione generata automaticamente*

Di seguito, vengono riportate le distribuzioni del coefficiente di perdita **α**, calcolato sapendo il valore della *portata teorica*, per ogni classe di prodotti GP5 attraverso la rappresentazione mediante *boxplot*:  


Di seguito, invece, vengono riportati i valori di *correlazione* tra le grandezze di pressione e portata rilevate durante la fase di controllo @140rpm per ogni classe di prodotti GP5:

*Immagine che contiene ricevuta, testo

Descrizione generata automaticamente*

Si nota come vi è alta correlazione tra pressione e portata sia per il *test a regime*, sia per il *test di controllo*, confermando l’esistenza della relazione descritta nel paragrafo 1.

È giustificato, quindi, indagare su una relazione in grado di esprimere la portata in funzione della pressione.

**4. Novelty detection**

Con lo scopo di rilevare in real time una coppia di valori portata – pressione “anomala” per il test di controllo (@140 rpm), ovvero non coerente con la distribuzione dei dati a disposizione, è stato utilizzato un algoritmo di “novelty detection” chiamato *Isolation Forest*.

Si tratta di un algoritmo di Machine Learning con apprendimento non supervisionato, in grado di isolare le osservazioni “anomale” seguendo il seguente criterio: essendo un *ensemble method* basato su alberi decisionali, la probabilità che un’osservazione sia “anomala” aumenta al diminuire della lunghezza media del path tramite il quale si arriva dal root al nodo foglia.

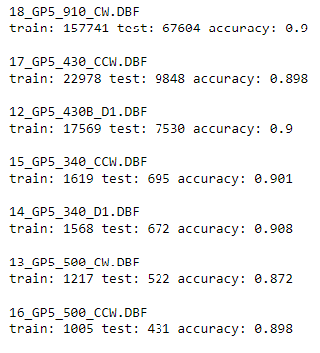
In pratica, un’osservazione “anomala” ha la tendenza ad essere classificata con pochi split.

L’*Isolation Forest*, dopo essere stato opportunamente addestrato sul training set testandone le performance sul test set,è in grado di:

* determinare se un’osservazione è anomala (1) o no (0);
* specificare uno score compreso tra 0 e 1 indicante il grado di “anomalia”.

Per il dataframe oggetto di studio è stato addestrato un algoritmo *Isolation Forest* con *hypeparameters* predefiniti, dividendo le osservazioni in dati di training (70%) e dati di test (30%), usando le variabili relative a portata e pressione per la fase di controllo (@140 rpm).

Presupponendo che le osservazioni nel test set non siano “anomale”, l’accuratezza dell’algoritmo può essere misurata semplicemente contando quante osservazioni sono state definite “anomale” dall’algoritmo e dividendo per il numero di osservazioni totale del test set.

Di seguito, i valori di accuratezza per l’algoritmo applicato ad ogni classe di prodotti GP5:  


Nella prossima figura viene rappresentato un esempio di applicazione dell’algoritmo *Isolation Forest* ad una nuova osservazione, quindi alla coppia pressione – portata rilevata durante la fase di controllo @140rpm: in questo caso l’osservazione viene definita “*inlier*”, quindi coerente con la distribuzione storica dei valori delle grandezze.

Immagine che contiene testo, mappa

Descrizione generata automaticamente

È opportuno specificare che un’osservazione “anomala” non corrisponde necessariamente ad un prodotto con esito del test di controllo (@140 rpm) fallito, ma potrebbe voler significare semplicemente che i valori delle grandezze fisiche si discostano molto dalla distribuzione storica.

**5. Regressione lineare della portata in funzione della pressione**

**6. Conclusioni: risposte ai task**