



UNIVERSITA' POLITECNICA DELLE MARCHE  
FACOLTA' DI INGEGNERIA

---

Corso di Laurea triennale in Ingegneria Gestionale

**ANALISI DEI LOG DI ACCESSO AD UN CORSO DELLA  
PIATTAFORMA  
LEARN.UNVPM.IT: IMPATTO DELLA DIDATTICA A DISTANZA**

**ANALYSIS OF ACCESS LOGS TO A COURSE ON THE  
LEARN.UNIVPM.IT PLATFORM: IMPACT OF DISTANCE  
LEARNING**

Relatore:

Prof. Domenico Potena

Correlatore:

Dott. Alex Mircoli

Tesi di Laurea di:

Roberto Piccioni

A.A. 2019/2020

## **INDICE**

### **CAPITOLO 1: INTRODUZIONE**

1.1 Inquadramento generale.....	4
---------------------------------	---

### **CAPITOLO 2: PROGETTAZIONE DATABASE**

2.1 Progettazione e implementazione Database.....	5
---	---

2.2 Definizione dettagliata del database.....	6
---	---

2.3 Interrogazioni e analisi dati.....	8
--	---

### **CAPITOLO 3: PROCESS MINING**

3.1 Tecniche di Process Mining.....	27
-------------------------------------	----

3.2 Process Discovery e tool utilizzati.....	28
--	----

3.3 Risultati ottenuti.....	29
-----------------------------	----

### **CAPITOLO 4: CONCLUSIONI E SVILUPPI FUTURI**

4.1 Conclusioni.....	34
----------------------	----

4.2 Sviluppi futuri.....	35
--------------------------	----



# **CAPITOLO 1: INTRODUZIONE**

## **1.1 Inquadramento generale**

L'Università Politecnica delle Marche mette a disposizione di docenti e studenti una piattaforma di e-learning basata su Moodle. Si tratta di un software open source che permette ai docenti di caricare, condividere i materiali utili per il loro corso e agli studenti di scaricare e visualizzare tali documenti. Causa COVID-19, e la conseguente impossibilità di svolgere le attività didattiche in presenza, questa risorsa si è rivelata preziosissima, garantendo l'opportunità d'insegnamento.

Da questo scenario sono emerse molteplici situazioni d'interesse. In particolare quelle che hanno stimolato e guidato l'indagine di questo elaborato, sono quelle che concernono l'analisi di un log di accesso relativo ad un corso. A partire da un log degli accessi, ossia una raccolta di dati memorizzati su un file, riguardanti gli ingressi degli utenti al corso e le singole attività svolte dagli stessi, si è sviluppato un database atto alla memorizzazione più significative per l'analisi. Quest'ultima è stata condotta in base a varie prospettive: i tipi di attività più frequenti, le sezioni del corso maggiormente visitate, le differenti attività/operazioni svolte dai singoli studenti, i periodi di maggiore afflusso sulla piattaforma. Grazie a queste analisi è stato possibile evidenziare quelle che sono state le conseguenze della didattica a distanza, obiettivo finale di questo lavoro di tesi.

Nello specifico il lavoro è stato articolato nei seguenti punti, che saranno approfonditi nei rispettivi capitoli e di cui si propone una sintetica descrizione:

- 1) Progettazione e implementazione di un Database: attraverso l'utilizzo di MySql, un software open source di gestione di database relazionali, è

stato possibile organizzare i dati estrapolati dal log in un database appositamente creato (CAPITOLO 2);

- 2) Analisi dei dati: individuazione delle query (i.e., interrogazioni) per l'ottenimento delle informazioni d'interesse per l'analisi e il raggiungimento dello scopo prefissato (CAPITOLO 2);
- 3) Process Mining: tecnica avanzata di analisi dei processi che, partendo dai dati di un log di eventi, permette di generare modelli di processo (Process Model) derivati dall'esecuzione di algoritmi di Process Discovery (CAPITOLO 3);
- 4) Conclusioni e sviluppi futuri: al termine della trattazione verranno descritti i risultati del progetto, valutati rispetto al conseguimento degli obiettivi e inoltre verranno proposte riflessioni in merito a prospettive di analisi future. (CAPITOLO 4)

## **CAPITOLO 2: Progettazione Database**

### **2.1 Progettazione e implementazione Database**

A partire dal log degli accessi al corso preso in esame, è stato necessario creare un database per poter organizzare l'insieme di dati contenuti nel file. Grazie al sistema di gestione di base di dati relazionali MySQL, questo è stato possibile mediante il linguaggio di programmazione SQL, acronimo di Structured Query Language. Tale linguaggio comprende un Data Definition Language (DDL) (insieme di comandi per la definizione dello schema della base di dati relazionale) e un Data Manipulation Language (DML) (insieme di comandi per la modifica e l'interrogazione dell'istanza di una base di dati).

Il log in questione comprende 20268 righe corrispondenti agli accessi degli utenti, nell'arco temporale che va da Gennaio 2020 a luglio 2020. Ogni riga oltre all'istante temporale di accesso (timestamp), contiene anche informazioni come l'attività svolta, il codice identificativo dell'utente (un numero da 1 a 6 cifre), l'indirizzo web associato ed altre informazioni meno rilevanti.

Un database relazionale permette di rappresentare le entità che vogliamo memorizzare mediante tabelle composta da righe (tuple) e colonne (attributi), che definiscono le caratteristiche di ciascuna entità.

## 2.2 Definizione dettagliata del database:

La definizione del database e dello schema è stata eseguita tramite il linguaggio di programmazione SQL. Il database e la tabella contenente i log sono stati creati mediante le seguenti istruzioni:

```
-create database corso;
```

```
-use corso;
```

```
-create table logcorso(
```

```
id int not null auto_increment,
```

```
data1 varchar(20) ,
```

```
desk varchar(150) ,
```

```
fonte varchar(150) ,
```

```
evento varchar(200) ,
```

```
utente varchar(500) ,
```

```
address varchar(40) ,
```

```
indirizzo varchar(30),
```

-primary key (id));

L'utilizzo di tale sintassi ha generato un database di nome "corso" e una tabella "logcorso" contenente i seguenti attributi:

- id: con 2 domini che definiscono che tale valore sia un intero con sviluppo incrementale per ogni tupla;

-data1: stringa di massimo 20 caratteri, che rappresenta l'istante temporale dell'accesso registrato;

-desk e fonte: stringhe di massimo 150 caratteri che non contengono però informazioni rilevanti ai fini delle successive interrogazioni;

-evento: stringa di massimo 200 caratteri che specifica l'attività svolta;

-utente: stringa di massimo 500 caratteri che specifica quale utente, con il corrispettivo codice identificativo, ha compiuto tale azione/attività;

-address e indirizzo: stringhe rispettivamente di al massimo 40 e 30 caratteri che definiscono l'indirizzo ip di provenienza di tale accesso;

ed infine abbiamo la definizione dell'attributo id come chiave primaria, ovvero un vincolo intrarelazionale, cioè definito nella relazione stessa, che fa sì che attraverso tale attributo ciascuna tupla sia identificabile.

Ai fini delle interrogazioni successive è stato poi aggiunto un ulteriore attributo per identificare univocamente ciascun utente secondo i codici id associati ad essi nell'attributo utente (e più in generale nel log). Su ogni accesso alla piattaforma registrato sul log, viene riportata una stringa descrivente quale utente e con quale id ha svolto tale azione in tale accesso. In questo caso vi erano stringhe in italiano ed altre in inglese, così da richiedere l'utilizzo di 2 select qui descritte:

```
select distinct (substring(utente,16, 6))
```

```
from logcorso
```

```
where utente like 'l utente con id %';  
  
select distinct (substring(utente, 18, 6))  
  
from logcorso  
  
where utente like 'the user with id %';
```

-identificativo: contenente il valore del codice id associato all'utente estrapolato dall'attributo "utente" della tabella logcorso con due select con funzioni distinct e substring e operatore like nella clausola where.

Sostanzialmente la sintassi di queste due interrogazioni fa sì che vengano restituiti come risultati, tutti i valori distinti dei caratteri delle stringhe dell'attributo utente, corrispondenti rispettivamente ai successivi 6 caratteri dei primi 16, quando la clausola da rispettare espressa nella where e dall'operatore di confronto like è "l'utente con id %", e i successivi 6 dei primi 18 quando è "the user with id %".

È stato poi necessario compiere un'operazione di modifica dello schema del database, al fine di aggiungere un'ulteriore colonna:

```
use corso;  
  
alter table logcorso add identificativo int;
```

Per popolare la base di dati appena creata, cioè per "tradurre" e organizzare le informazioni contenute in ciascuna riga del log, è stato utilizzato il comando di inserimento di righe fornito dal linguaggio (insert into) che permette di specificare il valore degli attributi delle singola occorrenze.

## 2.3 Interrogazioni e Analisi dei dati

Questa terza parte del capitolo riguarda quelle che sono state le interrogazioni eseguite sul database per raccogliere le informazioni di interesse per l'analisi.

Select 1:

Individuazione mese per mese dell'afflusso degli utenti sulla piattaforma:

```
select count(data1)
from logcorso
where data1
like '___/0N/____'; (con N: 1<=N<=7)
```

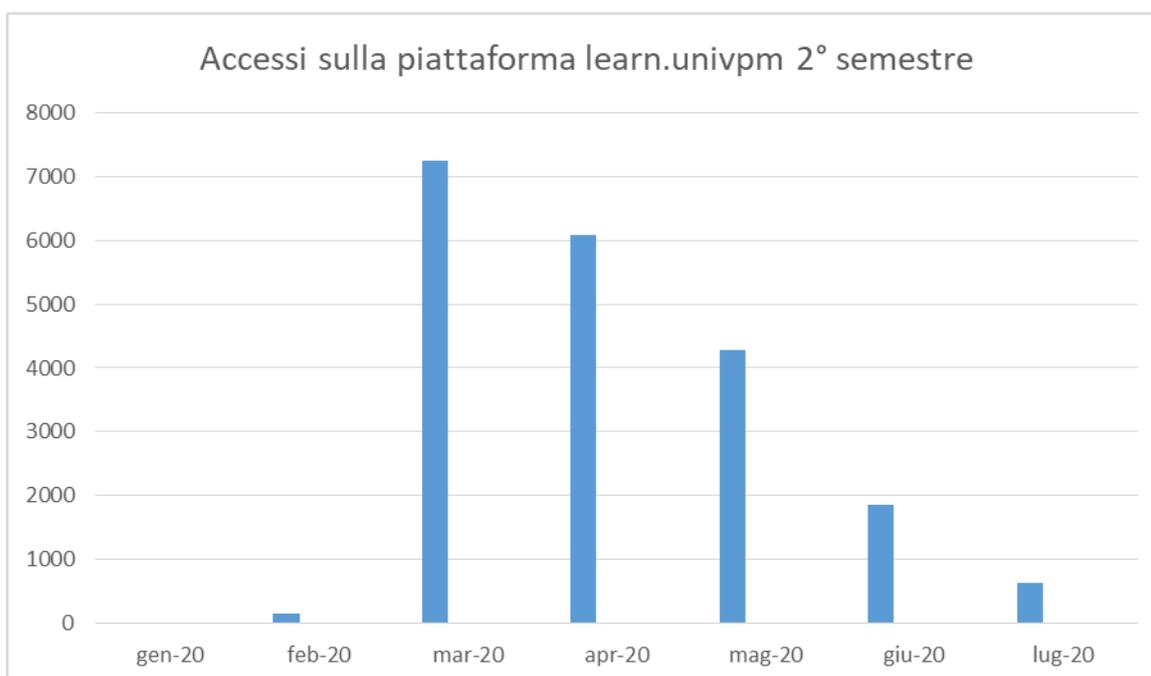
La select attraverso l'operatore count sull'attributo data1, restituisce il numero il numero di righe che soddisfano la condizione espressa nella where dall'operatore di confronto like. Ovvero tramite questa selezione sono stati individuati per i mesi che vanno da gennaio 2020 a luglio 2020 il numero di accessi alla piattaforma.

I risultati ottenuti sono riportati nella *Tabella 2.1* e *Figura 2.1*:

<b>Mese dell'anno</b>	<b>N° accessi totali</b>	
<b>Mese di Gennaio 2020</b>	0	
<b>Mese di Febbraio 2020</b>	148	
<b>Mese di Marzo 2020</b>	7253	
<b>Mese di Aprile 2020</b>	6077	
<b>Mese di Maggio 2020</b>	4287	
<b>Mesi di Giugno 2020</b>	1862	

<b>Mese di Luglio 2020</b>	641
--------------------------------	-----

*Tabella 2.1: Descrizione tabellare che associa mese per mese il numero di accessi alla piattaforma.*



*Figura 2.1: diagramma a barre che permette una maggiore leggibilità dei risultati ottenuti*

Per il mese di Gennaio abbiamo 0 accessi, dovuto al fatto che il corso d'insegnamento ha avuto inizio a Marzo. I mesi di maggiore affluenza sono stati quelli di marzo, aprile e maggio che rappresentano circa l'87% degli accessi. Non a caso questi sono stati i mesi che hanno visto l'inizio delle lezioni online e quindi come era pronosticabile quelli con maggiori accessi. Si può notare poi un forte decremento per i mesi di Giugno e Luglio, probabilmente da rimandare al fatto che sono mesi che coincidono con la sessione d'esami estiva.

Select 2:

Individuazione degli utenti e degli accessi eseguiti:

```
select count(*)
```

```
from logcorso
```

```
where identificativo like '_____';
```

Dopo aver estratto (come precedentemente spiegato) il codice identificativo di ciascun utente dalla stringa dell'attributo utente, è stato possibile conteggiare gli accessi di ciascuno di essi. Vengono riportate una rappresentazione tabellare (tab. 2.2) e una rappresentazione grafica (fig. 2.2) delle informazioni ottenute.

<b>UTENTE</b>	<b>TOTALE ACCESSI</b>
<b>312</b>	11
<b>392</b>	1564
<b>446</b>	12
<b>527</b>	13
<b>550</b>	8
<b>714</b>	48
<b>814</b>	6
<b>1145</b>	6
<b>2627</b>	28
<b>8144</b>	5
<b>8581</b>	34
<b>10494</b>	156
<b>10602</b>	4

<b>13367</b>	345
<b>14245</b>	22
<b>16036</b>	35
<b>16041</b>	480
<b>16628</b>	248
<b>16633</b>	11
<b>16728</b>	12
<b>17358</b>	408
<b>17601</b>	103
<b>17679</b>	355
<b>17752</b>	1079
<b>17934</b>	196
<b>18102</b>	491
<b>18269</b>	352
<b>18515</b>	469
<b>18578</b>	564
<b>18734</b>	381
<b>19218</b>	3
<b>20473</b>	396
<b>20488</b>	654
<b>20502</b>	244
<b>20573</b>	513
<b>20615</b>	8
<b>20622</b>	446
<b>20632</b>	207

<b>20645</b>	67
<b>20928</b>	52
<b>20977</b>	282
<b>21020</b>	238
<b>21023</b>	385
<b>21113</b>	343
<b>21151</b>	420
<b>21178</b>	238
<b>21206</b>	170
<b>21221</b>	26
<b>21285</b>	165
<b>21464</b>	388
<b>21685</b>	127
<b>21850</b>	82
<b>21927</b>	2
<b>21932</b>	512
<b>21987</b>	474
<b>22019</b>	17
<b>22044</b>	2
<b>22050</b>	3
<b>22073</b>	2
<b>22169</b>	376
<b>22294</b>	585
<b>22295</b>	446
<b>22383</b>	216

<b>22389</b>	101
<b>22902</b>	777
<b>22932</b>	652
<b>22996</b>	689
<b>23031</b>	419
<b>23067</b>	506
<b>23099</b>	134
<b>23144</b>	3
<b>23240</b>	172
<b>23248</b>	369
<b>23497</b>	232
<b>23548</b>	15
<b>23579</b>	16
<b>23604</b>	20
<b>23707</b>	406

*Tabella 2.2: Totale delle attività svolte associate ad ogni singolo utente, identificato dal suo codice e per semplicità di lettura ordinati in ordine crescente.*



Figura 2.2: Diagramma a barre dei primi 40 utenti

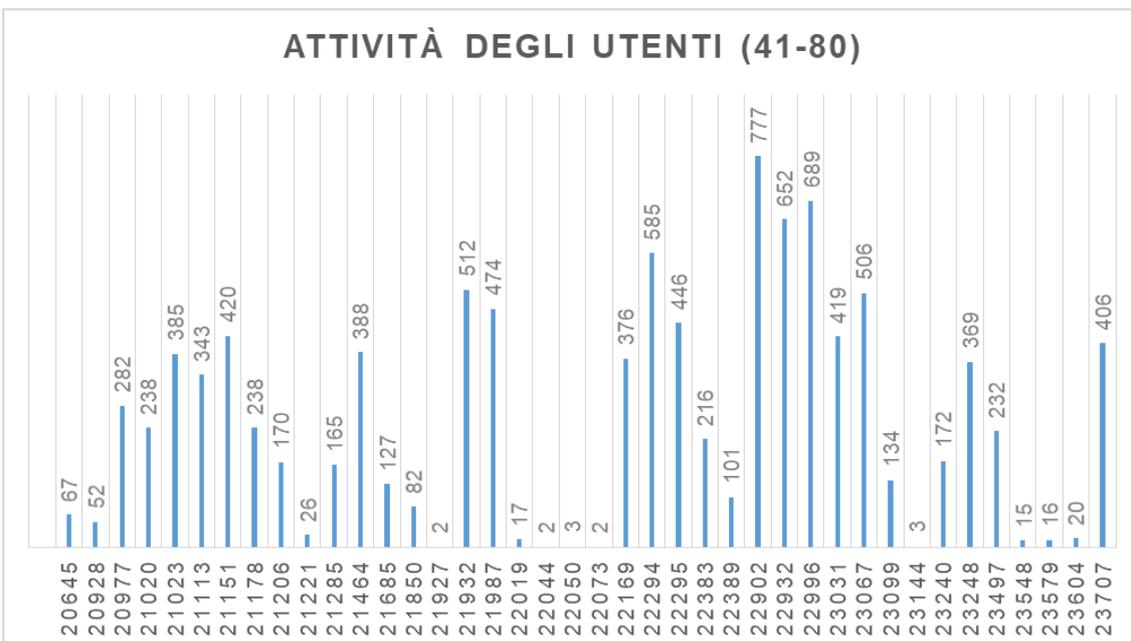


Figura 2.3: Diagramma a barre dei successivi 40 utenti

Il grafico ci mostra come la totalità degli utenti sia pari a 80 ed il totale degli accessi per ciascuno di essi. Il codice 392 è il codice associato al gestore del corso di cui qui di seguito andremo ad analizzare nel dettaglio le attività svolte.

Si può notare come esso sia il maggior utilizzatore della piattaforma con un totale di 1564 accessi eseguiti.

Select 3:

Analisi delle attività svolte dall'utente 392, ossia il gestore del corso:

```
select evento,count(evento)
```

```
from logcorso
```

```
where identificativo like'392'
```

```
group by evento;
```

Nella tabella 2.4 e nel grafico 2.4 vengono riportate tutte le attività dell'utente 392.

<b>ATTIVITA' UTENTE 392</b>		
Aggiornata cartella		7
Aggiornata sezione corso		17
Aggiornato corso		1
Aggiornato evento di calendario		11
Aggiornato gruppo		1
Aggiornato intervento		1
Aggiornato modulo corso		119
Aggiunto membro del gruppo		8
Assegnato ruolo		4
Creata discussione		20

<b>Creata pagina</b>		11
<b>Creato modulo corso</b>		45
<b>Disiscritto utente dal corso</b>		4
<b>Eliminato modulo corso</b>		4
<b>Esportate valutazione in XLS</b>		1
<b>Inserito gruppo nel raggruppamento</b>		29
<b>Inserito intervento nel forum</b>		21
<b>Iscritto utente al corso</b>		4
<b>Revocato ruolo</b>		4
<b>Rimosso gruppo dal raggruppamento</b>		1
<b>Rimosso membro del gruppo</b>		3
<b>Ripristinato corso</b>		1
<b>Scarica tutte le consegne</b>		1
<b>Valutata una consegna</b>		128
<b>Valutato utente</b>		111
<b>Video resource viewed</b>		24
<b>Visualizzata discussione</b>		31
<b>Visualizzata pagina di contenuto</b>		3
<b>Visualizzata scheda di consegna</b>		1
<b>Visualizzata tabella delle valutazioni</b>		13

<b>Visualizzato corso</b>			496
<b>Visualizzato utenti</b>	<b>elenco</b>		80
<b>Visualizzato corso</b>	<b>modulo</b>		201
<b>Visualizzato utente</b>	<b>profilo</b>		5
<b>Visualizzato valutatore</b>	<b>registro</b>		15
<b>Visualizzato report</b>			1
<b>Visualizzato report attività del corso</b>			3
<b>Visualizzato report log</b>			8
<b>Visualizzato report log utente</b>			1
<b>Visualizzato valutazione</b>	<b>schema</b>	<b>di</b>	90
<b>Visualizzato consegna</b>	<b>stato</b>		35

*Tabella 2.4: Descrizione e conteggio delle differenti attività del gestore del corso*



Figura 2.4: Diagramma a barre delle attività del gestore del corso

Il totale è pari a 41. Le maggiori attività svolte da tale utente sono la visualizzazione del corso, la visualizzazione e l'aggiornamento di un modulo del corso. Ma anche attività come la valutazione di una consegna e di un utente, queste sono proprie dell'utente in questione, in quanto, come si potrà notare dalla successiva valutazione di tutte le attività svolte sulla piattaforma, alcune di queste sono state e ovviamente possono essere svolte solo da tale utente.

Select 4:

Individuazione delle differenti attività e conseguente conteggio, la tabella 2.5 e la figura 2.5 ne danno una rappresentazione:

```
select evento, count(evento)
```

```
from logcorso
```

```
group by evento;
```

<b>EVENTO</b>		<b>N° totale</b>
<b>Aggiornata cartella</b>		7
<b>Aggiornata sezione corso</b>		17
<b>Aggiornato corso</b>		1
<b>Aggiornato evento di calendario</b>		11
<b>Aggiornato gruppo</b>		1
<b>Aggiornato intervento</b>		1
<b>Aggiornato modulo corso</b>		119
<b>Aggiunto membro del gruppo</b>		114
<b>Assegnato ruolo</b>		28
<b>Caricato file</b>		31
<b>Creata consegna</b>		34
<b>Creata discussione</b>		20
<b>Creata pagina</b>		1
<b>Creato elemento</b>		4
<b>Creato evento di calendario</b>		2
<b>Creato gruppo</b>		10
<b>Creato modulo corso</b>		45
<b>Disiscritto utente dal corso</b>		5
<b>Effettuata consegna testo online</b>		11
<b>Effettuata ricerca nel corso</b>		8
<b>Effettuata scelta</b>		106

<b>Effettuata una consegna</b>	31
<b>Eliminata valutazione</b>	1
<b>Eliminato elemento</b>	4
<b>Eliminato modulo corso</b>	4
<b>Esportate valutazione in XLS</b>	1
<b>Inserito gruppo nel raggruppamento</b>	29
<b>Inserito intervento nel forum</b>	21
<b>Iscritto utente al corso</b>	29
<b>L'archivio zip della cartella è stato scaricato</b>	125
<b>Revocato ruolo</b>	5
<b>Rimosso gruppo dal raggruppamento</b>	1
<b>Rimosso membro del gruppo</b>	3
<b>Ripristinato corso</b>	1
<b>Salvata consegna da parte di un utente</b>	8
<b>Scarica tutte le consegne</b>	1
<b>Valutata una consegna</b>	128
<b>Valutato utente</b>	222
<b>Video resource viewed</b>	2321
<b>Visualizzata attività recente</b>	4
<b>Visualizzata discussione</b>	264
<b>Visualizzata introduzione al corso</b>	72
<b>Visualizzata pagina di</b>	3

<b>contenuto</b>		
<b>Visualizzata scheda di consegna</b>		52
<b>Visualizzata scheda individuale</b>		13
<b>Visualizzata tabella delle valutazioni</b>		18
<b>Visualizzato commento</b>		123
<b>Visualizzato corso</b>		8465
<b>Visualizzato elenco istanze moduli corso</b>		1
<b>Visualizzato elenco utenti</b>		181
<b>Visualizzato modulo corso</b>		6509
<b>Visualizzato profilo utente</b>		170
<b>Visualizzato registro valutatore</b>		15
<b>Visualizzato report</b>		1
<b>Visualizzato report attività del corso</b>		2
<b>Visualizzato report attività del corso</b>		2
<b>Visualizzato report log</b>		8
<b>Visualizzato report log utente</b>		1
<b>Visualizzato report utente del corso</b>		3
<b>Visualizzato schema di valutazione</b>		91
<b>Visualizzato stato consegna</b>		760

*Tabella 2.5: totale delle differenti attività associate agli accessi degli utenti e il conteggio di ciascuna di esse.*

## DIFFERENTI ATTIVITA'/EVENTI



Figura 2.5: Diagramma a barre orizzontale delle attività e il loro conteggio

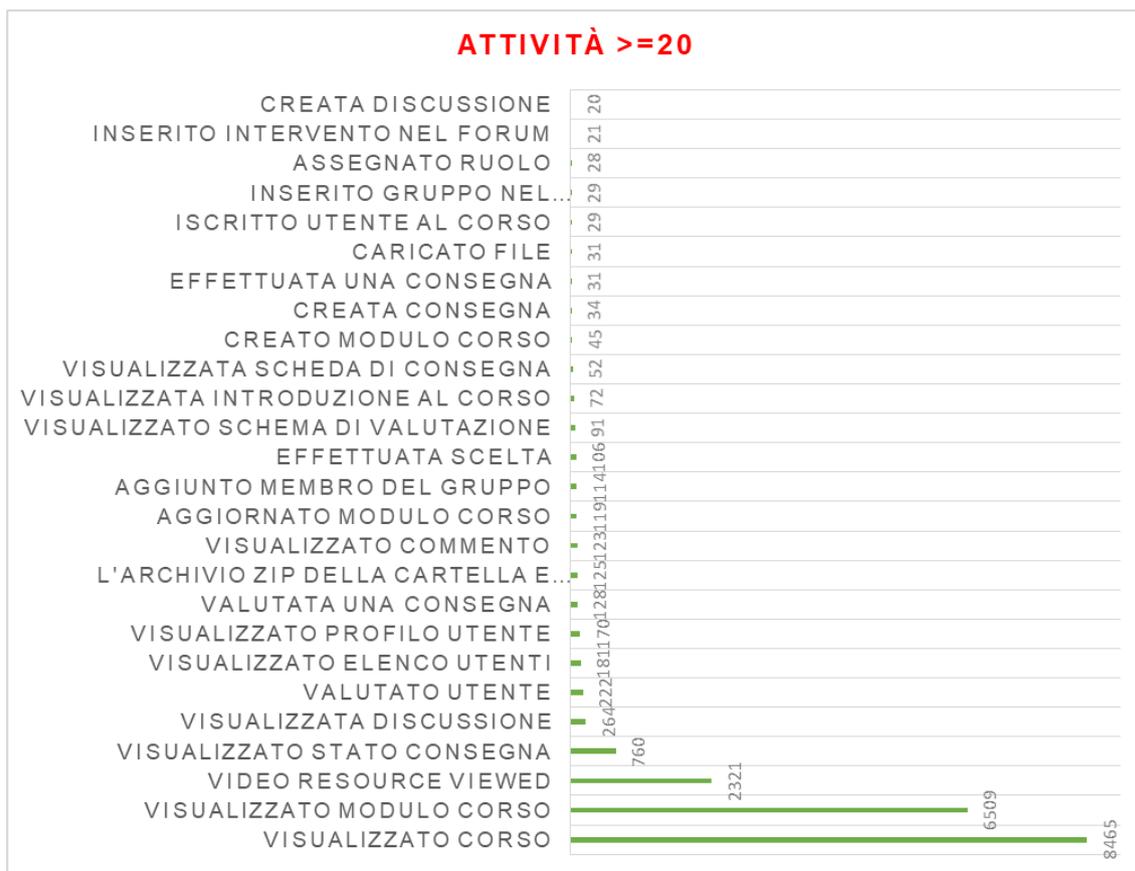
Il totale delle differenti attività è pari a 61. Da questa interrogazione sono emerse attività di scarso interesse e di scarsa frequenza, infatti successivamente è stato fatto un'ulteriore indagine volta ad eliminare tali attività. Quindi qui di seguito si può valutare il risultato di una alle sole attività che sono state eseguite almeno 20 volte. La tabella 2.6 e la figura 2.6 esplicitano i risultati ottenuti:

```
select evento,count(evento) as TotaleEvento
from logcorso
group by evento
having count(evento)>=20;
```

<b>ATTIVITA' &gt;=20</b>		
<b>Visualizzato corso</b>		8465
<b>Visualizzato modulo corso</b>		6509
<b>Video resource viewed</b>		2321
<b>Visualizzato stato consegna</b>		760
<b>Visualizzata discussione</b>		264
<b>Valutato utente</b>		222
<b>Visualizzato elenco utenti</b>		181
<b>Visualizzato profilo utente</b>		170
<b>Valutata una consegna</b>		128
<b>L'archivio zip della cartella e stato scaricato</b>		125
<b>Visualizzato commento</b>		123
<b>Aggiornato modulo corso</b>		119

<b>Aggiunto membro del gruppo</b>	114
<b>Effettuata scelta</b>	106
<b>Visualizzato schema di valutazione</b>	91
<b>Visualizzata introduzione al corso</b>	72
<b>Visualizzata scheda di consegna</b>	52
<b>Creato modulo corso</b>	45
<b>Creata consegna</b>	34
<b>Effettuata una consegna</b>	31
<b>Caricato file</b>	31
<b>Iscritto utente al corso</b>	29
<b>Inserito gruppo nel raggruppamento</b>	29
<b>Assegnato ruolo</b>	28
<b>Inserito intervento nel forum</b>	21
<b>Creata discussione</b>	20

*Tabella 2.6: Descrizione tabellare delle attività con frequenza maggiore o uguale a 20.*



*Figura 2.6: Diagramma a barre orizzontale che elenca le attività con frequenza maggiore o uguale a 20 in ordine crescente.*

Il totale si è ridotto da 61 differenti attività a 26. Dall'indagine è emerso che le attività maggiormente svolte sono:

- La visualizzazione del corso (8465) 42%
- La visualizzazione del modulo di un corso (6509) 32%
- La visualizzazione di una video lezione (2321) 11%

Che con un totale di 17295 rappresentano praticamente la totalità delle azioni svolte sulla piattaforma, pari all'85%.

## **CAPITOLO 3: PROCESS MINING**

Il Process Mining è un'area di ricerca che si colloca tra il data mining, la modellazione e l'analisi dei processi. È una tecnica di Business Process Management usata per analizzare, monitorare e controllare i processi aziendali per come si svolgono nella realtà, con l'obiettivo di eliminare criticità per il processo.

L'idea di base del Process Mining è quella di estrarre conoscenza da un log degli eventi, registrati da un sistema software di un sistema informativo aziendale. Da questo generare un modello di processo da implementare, adattare a quello originariamente in uso e analizzato, oppure per generarne uno ex novo.

### **3.1 Tecniche di Process Mining**

Esistono 3 tecniche di Process Mining:

- 1) Process Discovery: è la tecnica più importante del Process Mining, ha come input un log di eventi e senza utilizzo di ulteriori informazioni produce un modello di processo.
- 2) Conformance Checking: prevede un confronto tra informazioni estratte da un log di eventi relativo al processo e il modello di processo stesso (preesistente) al fine di giudicarne la conformità. È quindi sostanzialmente è una tecnica di diagnosi.
- 3) Enhancement: ha come scopo quello di migliorare un modello di processo già esistente, utilizzando informazioni estratte da un log dello stesso.

Ai fini dello scopo della ricerca, la tecnica che è stata usata è quella del Process Discovery. Nel prossimo paragrafo verrà data una descrizione più approfondita del Process Discovery, dei tool utilizzati per le indagini condotte e del programma ProM (framework open source di riferimento per il Process Mining).

## 3.2 Process Discovery e tool utilizzati

Come precedentemente accennato la Process Discovery costituisce una delle tecniche di riferimento del Process Mining. In linea con l'idea di fondo di questo relativamente nuovo campo di ricerca, l'obiettivo che si prefigge è quello di modellare processi della realtà basandosi sull'analisi e l'interpretazione dei dati contenuti in un log di eventi. Quindi a partire da questi produrre come output una buona caratterizzazione di tutti i possibili percorsi di un processo. Le rappresentazioni di modelli vengono spesso espresse tramite una rete di Petri (un grafo orientato con due tipi differenti di nodi denominati place e transizioni collegati da un arco che può collegare solo nodi diversi) ma anche da altri tipi di formalismi.

Il programma utilizzato per l'esecuzione degli algoritmi scelti per l'analisi è ProM. Si tratta di un framework basato su plugin per il supporto di una grande quantità di tool del Process Mining. ProM è una piattaforma facile da estendere, che garantisce notevoli vantaggi per l'implementazione di nuove tecniche da parte di sviluppatori di algoritmi, e facile da usare per gli utenti. È implementata in linguaggio Java ed utilizza come standard, il formato XES (eXtensible Event Stream). Per la ricerca è stato necessario convertire il file di log da CSV in XES, questo è stato possibile grazie ad un plugin di conversione presente sulla piattaforma ProM. Inoltre, per poter sfruttare i tool offerti per questo tipo di Process Mining, è stato necessario apportare un'aggiunta al log di dati preso in esame. Si è trattato di creare il "concetto" di *sessione*, ovvero un insieme di attività svolte, riconducibili ad un arco temporale pari a 20 minuti. È stato quindi creato un nuovo script contenente l'id di sessione associato ad ogni riga del log.

Tra la vasta gamma di algoritmi a disposizione per la realizzazione di modelli di processo di Discovery Mining sono stati scelti i seguenti 3:

- 1) Alpha Miner: è il primo algoritmo volto a ricostruire la casualità da una serie di sequenze di eventi (ordinati cronologicamente) contenute in un log. Segue 3 regole per la derivazione di un modello da un log di eventi:

*-Dipendenza temporale:* ovvero se un evento A è seguito da un evento B e A non segue mai l'evento B, allora B dipende da A.

*-Indipendenza temporale:* quando esiste un collegamento, una traccia da un evento A verso un evento B e da B verso A, cioè quando entrambi gli eventi possono esser eseguiti in parallelo.

*-Indipendenza:* non esiste una traccia da A verso B o viceversa, e quindi i due eventi sono indipendenti. Non esistono relazioni tra le due attività.

- 2) Heuristic Miner: è un algoritmo migliorativo dell'alpha miner. In questo è possibile tenere conto delle frequenze delle attività registrate nel log. Pertanto possono essere filtrati comportamenti poco frequenti. Permette di rilevare brevi loop e salti di attività. Nonostante questi miglioramenti, questo algoritmo non garantisce sempre modelli di processo validi. Il modello di processo generato è descritto da una Heuristic net, che può essere facilmente convertita in una rete di Petri, tramite un apposito tool presente in ProM.
- 3) Inductive Miner: è un ulteriore miglioramento degli algoritmi alpha miner e heuristic miner, con obiettivo di colmare sempre più il gap tra un log di eventi e la scoperta e modellazione di un modello di processo. Una caratteristica è quella di garantire la solidità, validità del modello generato in output. Per prima cosa l'algoritmo compie una divisione prominente all'interno del registro delle attività (criteri di divisione: sequenziale, parallela, simultanea e loop) quindi identifica un caso base dal quale sviluppare un albero del processo, che possono essere poi convertiti in reti di Petri. Altra caratteristica è quella di garantire buoni valori di fitness, parametro che misura la "bontà" del modello di processo realizzato, che può assumere compresi tra 0 e 1 (valore mai realmente raggiungibile).

Nel prossimo paragrafo verranno descritti i risultati ottenuti dall'utilizzo degli algoritmi di Process Discovery appena trattati.

### **3.3 Risultati ottenuti**

Questa fase della ricerca è stata guidata dalla volontà di generare dei modelli di processo rispetto al log degli accessi preso in esame. Come accennato è stato necessario aggiornarlo con il nuovo concetto di sessione, per poter individuare dei percorsi, ossia delle successioni di attività che frequentemente sono state riscontrate nella realtà. In questo paragrafo saranno discussi singolarmente i risultati scaturiti dall'esecuzione degli algoritmi descritti precedentemente. Per ciascuno di essi verrà riportata la rappresentazione grafica (rete di Petri) ottenuta, rispettivamente la Figura 3.1 per l'alpha miner, la Figura 3.2 per l'heuristic miner, e la Figura 3.3 per l'inductive miner.

- Alpha Miner:

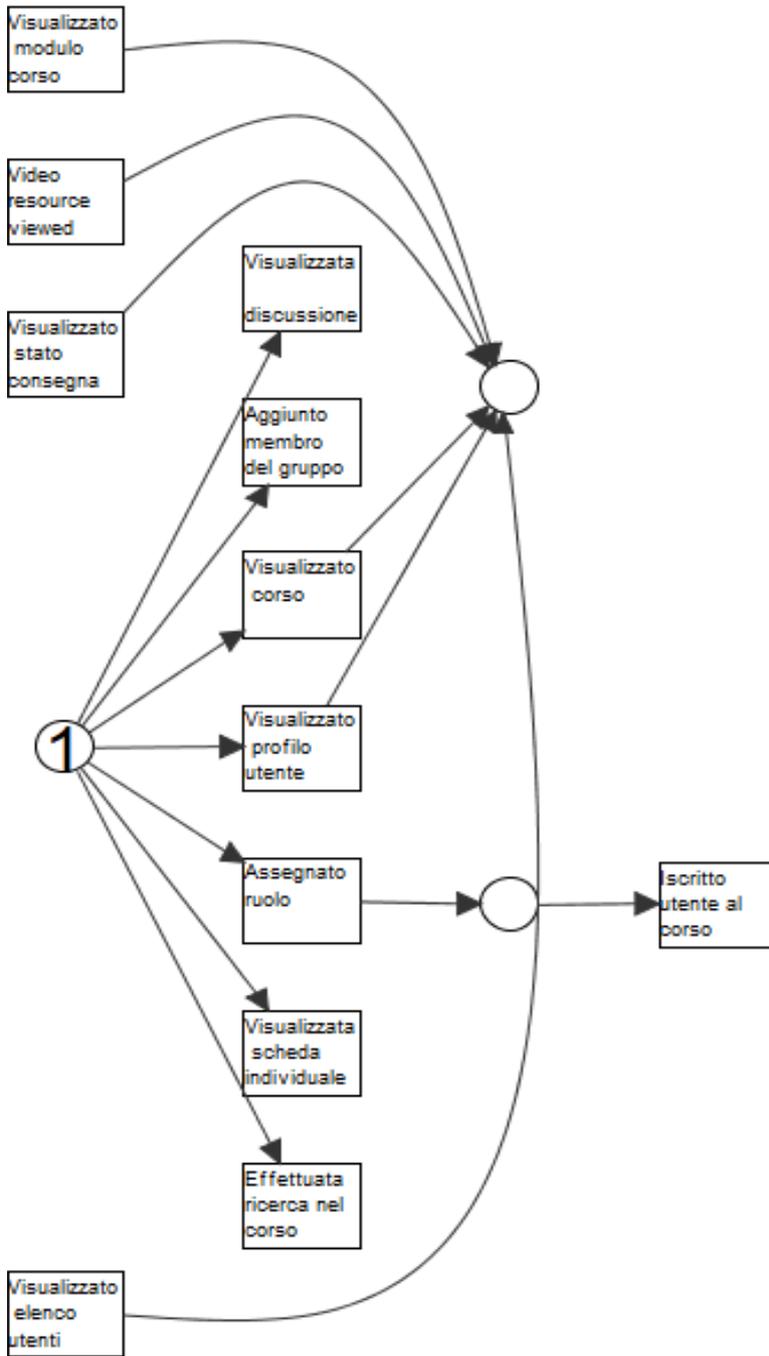
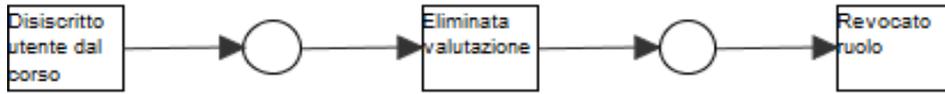


Figura 3.1: rappresenta la rete di Petri ottenuta dall'algorithm Alpha Miner

Le attività riportate sono raffigurate con un rettangolo che rappresenta una transizione, mentre i cerchi vengono chiamati place. Il grafo orientato mostra dei flussi registrati nel log e quindi delle sequenze di attività svolte secondo un preciso ordine. Un posto da cui un arco parte per finire in una transizione è detto *posto di input* della transizione; un posto in cui un arco arriva partendo da una transizione è detto *posto di output* della transizione.

-Heuristic Miner:

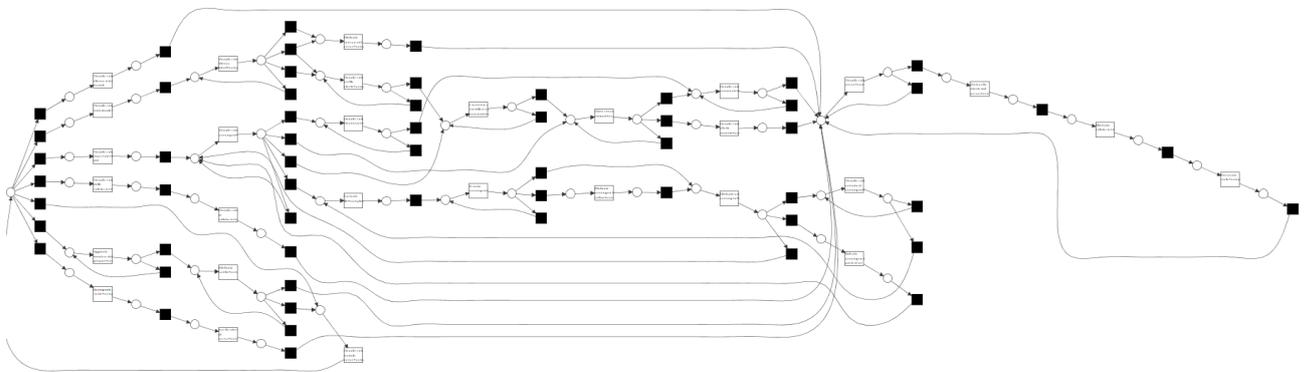


Figura 3.2: riporta la corrispettiva rete di Petri, della rete euristica ottenuta dall'esecuzione dell'algorithm.

A differenza dell'Alpha Miner, in questo algorithm vengono considerate le frequenze dei comportamenti riconosciuti nel log di eventi. Si possono osservare molteplici successioni di attività, cioè dei comportamenti. È interessante notare come per arrivare a svolgere determinate azioni sia necessario eseguire dei percorsi prestabiliti, come alle volte siano state effettuate azioni (transizioni nel grafo) differenti per poi arrivare a compiere allo stesso posto di output. Il valore di fitness associato al grafo è pari a 0.8896, valore che ci garantisce che il modello ottenuto è abbastanza conforme alla realtà descritta dal log. Di seguito sarà riportata anche la rappresentazione della rete euristica, Figura 3.2.1, che è stata convertita nella suddetta rete di Petri.

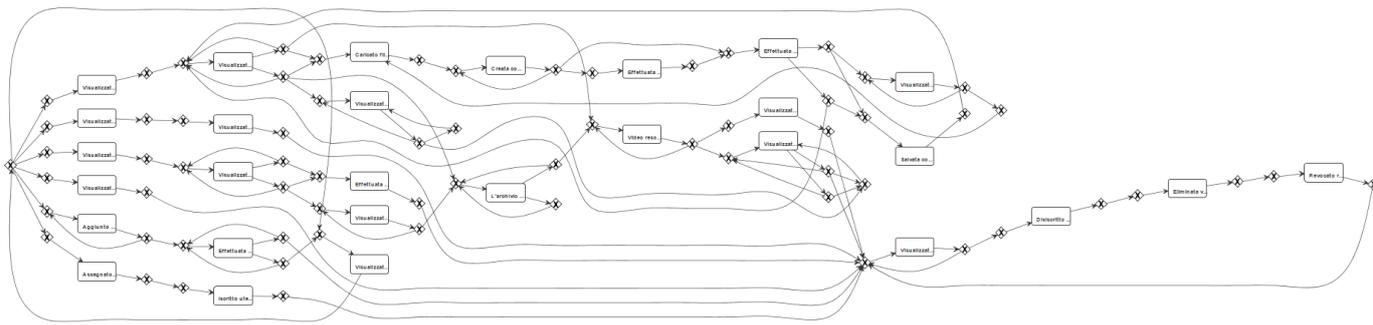


Figura 3.2.1: rete Euristica ottenuta dall'esecuzione dell'algoritmo.

-Inductive Miner:

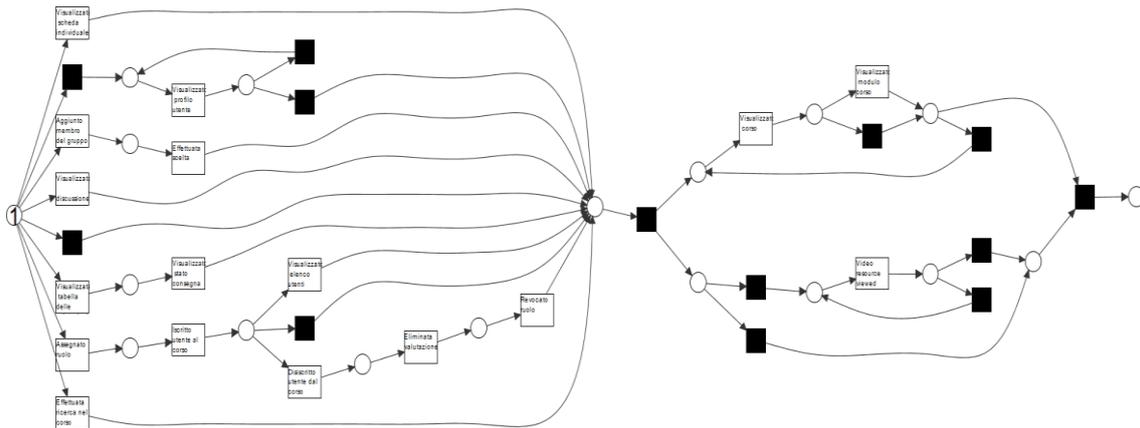


Figura 3.3: rete di Petri ottenuta dall'Inductive Miner

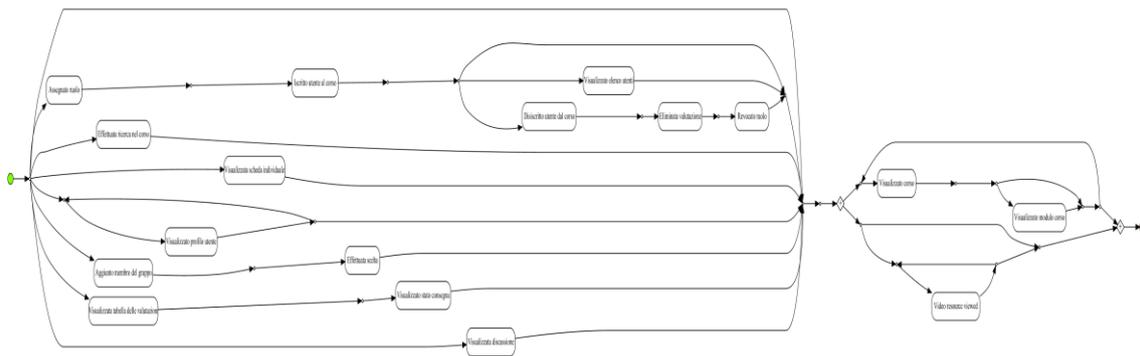


Figura 3.3: rappresentazione derivata dall'utilizzo del tool Inductive Visual Miner

## CAPITOLO 4: CONCLUSIONI E SVILUPPI FUTURI

### 4.1 Conclusioni

In quest'ultimo capitolo vengono tratte le conclusioni riguardanti il raggiungimento degli obiettivi della tesi svolta e le difficoltà incontrate durante le fasi di progetto. Infine, sulla base dei risultati ottenuti, saranno valutate delle ipotesi per futuri approfondimenti delle tematiche trattate.

Lo scopo della ricerca è stato quello di estrapolare informazioni dall'analisi dei log di accesso a un corso ai fini di valutare l'impatto della didattica a distanza. Da questi poi sono stati generati dei modelli di processo attraverso l'utilizzo di tool di Process Discovery. L'obiettivo può considerarsi raggiunto, anche se per quel che riguarda una valutazione più approfondita dell'impatto di questa nuova modalità d'insegnamento, è bene sottolineare che i dati a disposizione concernevano un solo semestre di accessi alla piattaforma. Tuttavia sono state estratte molteplici informazioni, a partire dal riconoscimento dei mesi di maggiore affluenza, alle attività eseguite con maggiore frequenza fino a modelli di processo, che descrivono comportamenti della realtà in merito all'esecuzione di tali attività.

Per quanto riguarda le difficoltà incontrate, queste possono essere inquadrate in due fasi distinte del lavoro svolto: la prima nell'implementazione dei dati sul database appositamente creato e la seconda nella fase di interpretazione dei modelli di processo ottenuti dalla Process Discovery.

Il primo ostacolo è stato proprio quello di riuscire ad estrapolare i codici identificativi degli utenti dal log degli accessi, poiché come spiegato nella trattazione (cap.2) questi si trovavano in stringhe scritte in lingua inglese ed altre in lingua italiana, aspetto che da una prima lettura dei dati non era immediato evincere.

Una seconda criticità è stata riscontrata nell'interpretazione dei modelli ottenuti, in quanto essendo modelli derivanti da tecniche utilizzate soprattutto per l'analisi di processi aziendali, non è stato facile configurare i dati e successivamente compiere una valutazione organica dei risultati.

Si rimanda al paragrafo successivo per valutare le opportunità future che possono prospettarsi.

## **4.2 Sviluppi futuri**

Diverse direzioni di sviluppo potrebbero essere prese in considerazione per approfondire il progetto svolto. Una potrebbe essere quella di raccogliere ulteriori dati di accesso dello stesso corso d'insegnamento, magari riferiti ad anni passati, per poi effettuare un'analisi di confronto, al fine di individuare differenze di comportamento da parte degli utenti della piattaforma e così da compiere ulteriori considerazioni rispetto all'impatto della didattica a distanza.

Un'altra potrebbe essere quella di approfondire l'analisi dei modelli di processo, che potrebbe dar spunti per una riorganizzazione delle varie sezioni del corso.

