UNIVERSITÀ DI PISA

**DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI, DEL TERRITORIO E DELLE COSTRUZIONI**

**RELAZIONE PER IL CONSEGUIMENTO DELLA LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE**

# *Progetto di Business Intelligence a supporto della piattaforma di monitoraggio dei disservizi. Il caso TIM*

## SINTESI

---


RELATORI

Prof. Antonella Martini
*Dipartimento di Ingegneria dell'Energia, dei Sistemi, del Territorio e delle Costruzioni*

Dott. Antonio Forti
*ELIS Consulting & Labs, ROMA*

CANDIDATO

Edoardo Bruni



Sessione di Laurea del 19/06/2019
Anno Accademico 2018/2019

**Progetto di Business Intelligence a supporto della piattaforma di monitoraggio dei disservizi. Il caso TIM**

**Edoardo Bruni**

## Sommario

Questa tesi si sviluppa grazie ad un'esperienza di stage, della durata di cinque mesi, svolta presso ELIS Consulting & Labs (Roma). Lo stage ha riguardato lo sviluppo di un progetto di consulenza nell'ambito IT. In particolare, l'ambito del progetto è inerente la piattaforma di intrattenimento digitale offerta da TIM.

Inizialmente la dashboard di monitoraggio della piattaforma non consentiva di individuare agevolmente le cause che generavano disservizi, andando ad inficiare sulla qualità percepita dal cliente. Verrà trattata la riprogettazione della dashboard, al fine di migliorare il monitoraggio dei dati e sarà evidenziato il processo che permette di descrivere le evidenze sulla base di una logica descrittiva. Questo potrà successivamente permettere di passare ad una logica predittiva, allo scopo di prevenire eventuali disservizi, utilizzando algoritmi di Machine Learning.

In conclusione, gli obiettivi del progetto sono:

- Ridisegnare la dashboard Digital Support Decoder
- Creare un modello per ridurre i disservizi

## Abstract

This thesis is developed thanks to a five-month internship at ELIS Consulting Academy (Rome). The internship involved the development of an IT consulting project. In particular, the scope of the project is inherent to the digital entertainment platform offered by TIM.

Initially, the platform's monitoring dashboard did not allow to easily identify the causes that generated inefficiencies, affecting the quality perceived by the customer. The redesign of the dashboard will be discussed, in order to improve the monitoring of data and the process will be highlighted that allows the evidence to be described on the basis of a descriptive logic. This will then allow you to switch to a predictive logic, in order to prevent any disruption, using Machine Learning algorithms. In conclusion, the objectives are:

- Redesigning the dashboard Digital Support Decoder
- Create a model that aims to reduce inefficiencies

# 1. CONTESTO, OBIETTIVI E METODOLOGIE UTILIZZATE

## 1.1 Contesto del lavoro

TIM è il brand con cui Telecom Italia commercializza i vari servizi di telefonia cellulare in Italia e in Brasile. Attualmente Telecom Italia è il settimo gruppo economico italiano per fatturato e tra i primi 500 mondiali; esso fornisce vari servizi e, in Italia, gestisce una grossa fetta della connettività Internet ed intranet.

Oltre ai servizi di telefonia, fissa e mobile, dal 2009, Telecom si è focalizzata sul settore adibito al servizio on demand, operante nella distribuzione on line di serie televisive, programmi di intrattenimento e film. Tramite l'utilizzo del decoder "TimBox", l'azienda offre una vasta gamma di contenuti articolati in specifiche aree, al fine di soddisfare i vari target di clienti.

A livello tecnico, il decoder (nome tecnico STB, "set-top-box") fornisce dati per poter rilevare eventuali non conformità, come ad esempio nel caso in cui il segnale durante la visualizzazione di un film diventi criptato oppure sia presente un errore di caricamento a display.

In un primo momento, l'azienda non riusciva ad identificare i disservizi in maniera esemplificata, ovvero la dashboard creata per il controllo delle occorrenze risultava ancora complessa e di difficile lettura.

Da qui è stata posta l'attenzione a perfezionare ed arricchire le viste che compongono la dashboard e, in un secondo momento, identificare gli errori in maniera proattiva, ovvero poter prevenire i disservizi prima che accadano identificandone possibili cause. La tesi si concentra sulla descrizione delle varie fasi progettuali ponendo l'enfasi relativamente alle attività adibite all'esplorazione del dataset (Data Exploration).

## 1.2 Fasi del lavoro, obiettivi e metodologie

L'intero progetto è strutturato sul rilascio di deliverable (output intermedi) in un tempo relativamente breve (circa ogni 2-3 settimane), utilizzando un metodo "agile" nell'affrontare il problema. Nel concreto significa realizzare il progetto per fasi, chiamate sprint, al fine di verificare l'allineamento con i desiderata dell'azienda committente, così da permettere il controllo lungo tutte le varie fasi progettuali e verificare la soddisfazione del cliente mostrando ciò che è stato realizzato fino a quel momento. Con tale approccio è possibile decidere, insieme al cliente, se perseverare lungo una direzione o modificare le attività di progetto, tenuto conto dei vincoli che delimitano il project scope. In *Tabella 1* è riportata la metodologia seguita per sviluppare ciascuna delle fasi in cui

il progetto è stato segmentato, con le relative attività, obiettivi, metodologie e ruoli svolti dal candidato ('R': responsabilità diretta; 'C': collaborazione).

| Fase | Macro – Attività | Obiettivi | Metodologie | Cap. | Ruolo |
|---|---|---|---|---|---|
| **Analisi Preliminare** | Analisi AS - IS | • Comprensione del progetto e delle criticità<br>• Mappatura delle viste presenti nella dashboard | Mappatura puntuale della dashboard ed analisi documenti forniti | 3.1 | C |
| **Proposte dashboard TO-BE e mapping del dato** | Proposte di dashboard TO-BE | • Proposta di nuove viste a supporto che favoriscono l'identificazione dei disservizi | Interviste agli utilizzatori della dashboard.<br>Utilizzo di bozze "draft" per proposte di nuove viste e ottenimento di feedback | 3.2 | R |
| | Mappatura del dato | • Mappatura completa del percorso effettuato dal dato | Lettura della documentazione per redigere i documenti di data flow | 3.3 | C |
| **Processo di data exploration finalizzato ai disservizi** | Esplorazione del dataset | • Identificazione puntuale dei disservizi<br>• Identificazione del dataset da analizzare<br>• Comprensione delle dimensioni | Focus sui disservizi.<br>Identificazione del database contenente il maggior quantitativo di informazioni al fine di poter effettuare un'esplorazione completa del dataset. Estrazione di un dataset più piccolo per facilitare il calcolo distribuito | 4.1 | R |
| | Analisi delle evidenze | • Identificazione delle evidenze riscontrate relative ai disservizi<br>• Validazione delle evidenze | Manipolazione e pulitura del dataset, inserimento di funzioni che facilitano la rilevazione delle occorrenze | 4.2 | R |
| **Risultati** | Verifica delle criticità | • Verifica di implementazione di nuovi KPI nella Dashboard<br>• Utilizzo delle evidenze per il modello predittivo | Verifica di implementazione delle evidenze nel modello predittivo | 5 | C |

*Tabella 1: Attività svolte, metodologie utilizzate e obiettivi per fase*

## 2. DESCRIZIONE E ANALISI STATO DELL'ARTE

Per facilitare la lettura è presente un richiamo sul concetto di Business intelligence (BI).

La BI è da sempre definita come un sistema di modelli, metodi, processi, persone e strumenti che rendono possibile la raccolta regolare ed organizzata del patrimonio dati generato da un'azienda.

In particolare, questo sistema coinvolge ogni livello dell'infrastruttura tecnologica aziendale, al fine di distillare le informazioni ed efficientare i processi decisionali (si veda la sua struttura in Figura 1).

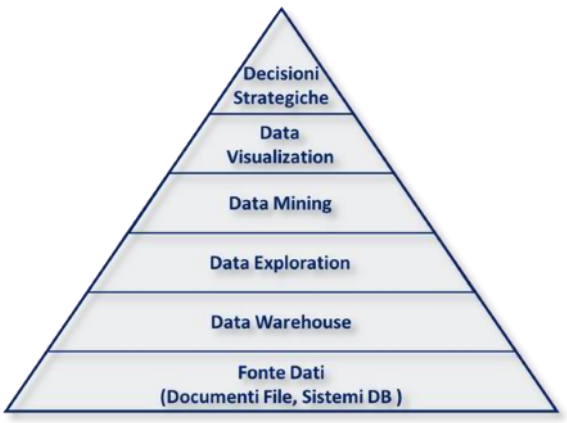


*Figura 1: Piramide Informativa BI*

Partendo dalle diverse fonti dati (come documenti file, sistemi DB), dopo alcuni processi di modifica

di questi ultimi, si arriva alla loro visualizzazione grafica, permettendo di concretizzare le decisioni strategiche. Tra i processi di modifica vi è l'attività di analisi esplorativa (Data Exploration).

Lo scopo principale è quello di mettere in evidenza i dati rilevanti e le caratteristiche di ogni attributo contenuto in un dataset, utilizzando metodi grafici e statistiche di sintesi per identificare l'eventuale relazione tra gli attributi.

Nello sviluppo della seconda fase progettuale vi è una descrizione dettagliata delle metodologie utilizzate. Le principali fonti sono state prelevate da "Google Scholar" e "Scopus".

## 3. SVILUPPO PRIMA FASE PROGETTUALE

### 3.1 Analisi AS-IS

Durante lo sviluppo della prima fase progettuale, per accertare cosa l'organizzazione stesse monitorando, è stata svolta un'attività di mappatura delle viste appartenenti al cruscotto aziendale, descrivendo KPI e grafici presenti. Questo ha permesso una maggior comprensione del contesto in analisi e reso le successive attività più facilitate.

### 3.2 Proposte di dashboard TO-BE

Allo stato iniziale la dashboard è suddivisa in 11 viste, ciascuna con lo scopo di verificare una parte specifica dei decoder. Tramite le interviste agli operatori è stata posta enfasi sulle criticità e sono state proposte nuove viste di supporto al fine di migliorare la visualizzazione ed il controllo dei decoder. Viene riportato di seguito l'esempio di una delle viste adibite al monitoraggio.

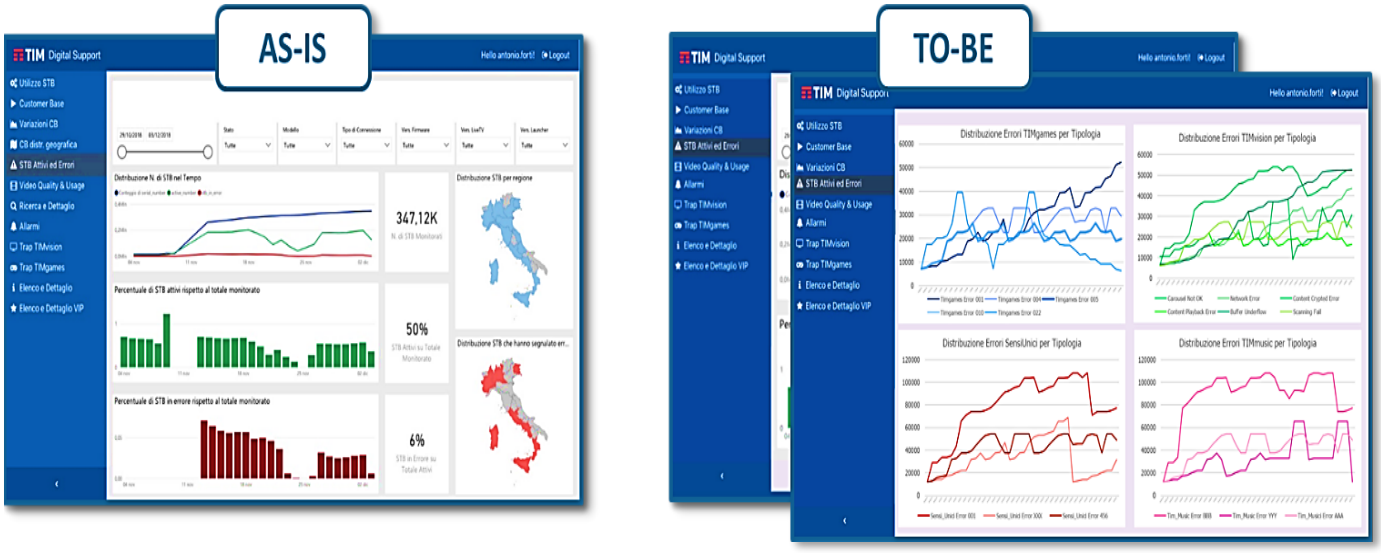


*Figura 2: Vista "STB Attivi ed Errori" (Confronto AS IS – TO BE)*

La vista “STB Attivi ed Errori” verifica lo stato dei decoder attivi sul totale monitorato (attivi e non attivi) evidenziando le percentuali in errore, ovvero i STB che hanno riscontrato un disservizio.
Tale vista si concentra sull’intera numerosità dei disservizi senza porre il focus nelle app specifiche offerte dall’azienda (ad esempio TIM Music, TIM games, TIM Vision). Nelle proposte è stata posta attenzione a identificare gli errori esclusivi per ciascuna app, permettendo all’operatore una migliore comprensione delle inefficienze ed eliminando possibili ambiguità di lettura.

### 3.3 Mappatura del dato

Al fine di verificare se le implementazioni proposte fossero fattibili è stata effettuata un’intera mappatura del percorso del dato fino all’arrivo alla dashboard. Questa attività ha avuto lo scopo di controllare la completezza dei dati che confluivano alla dashboard e capire se fosse possibile estrarre informazioni aggiuntive dai middleware precedenti, osservando, ad esempio quali dati non venivano considerati e si arrestavano nelle varie infrastrutture. Oltre a ciò questa attività è risultata necessaria al fine di capire su quale architettura fosse più efficace effettuare il processo di Data Exploration. Attraverso la figura riportata di seguito si può esplicare il percorso e le varie trasformazioni che effettua il dato fino ad arrivare alla dashboard.

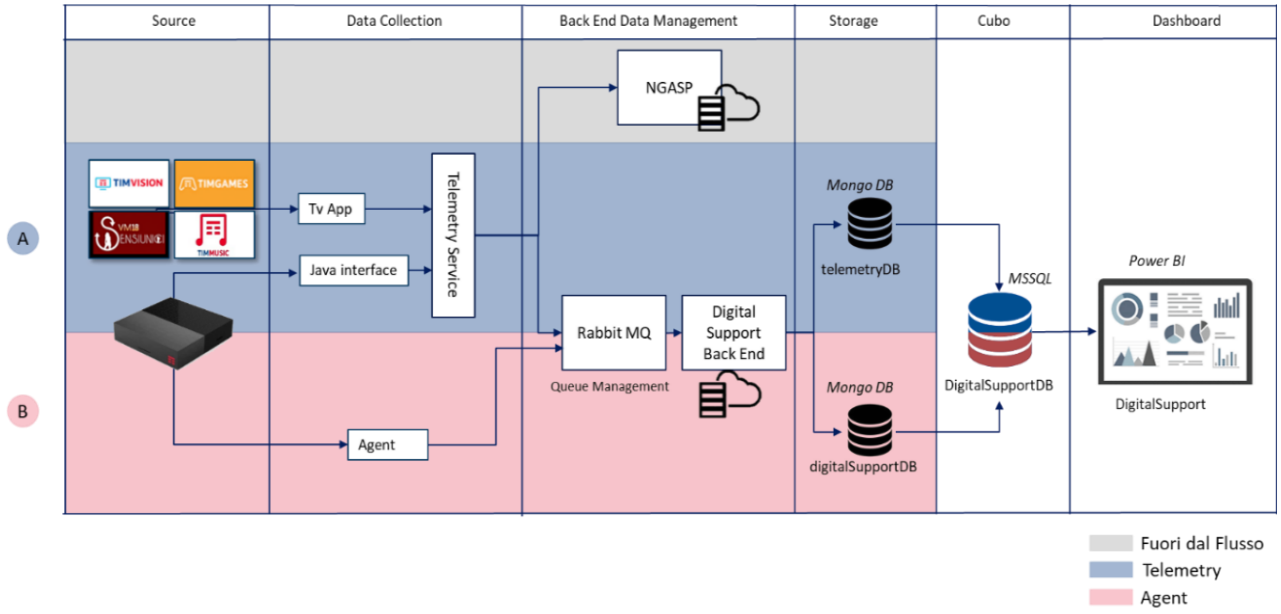


*Figura 3: Architettura del flusso del dato*

Partendo dal decoder fisico, il dato attraversa varie infrastrutture intermedie e convoglia al MongoDB (accumulatore di dati, Data Warehouse); successivamente passa dal MongoDB alla tecnologia Cubo (modello multidimensionale che ha la finalità di rilasciare i dati alla dashboard).

La figura 3 è stata suddivisa in due parti ponendo enfasi sulla classificazione del dato. In particolare, ciascun decoder genera due tipologie di dati:

- Telemetry (colore blu): parametri che monitorano le attività effettuate dagli utenti finali durante l'utilizzo del decoder (es. richiesta di visualizzazione di un video, termine di visualizzazione di un video);
- Agent (colore rosso): parametri tecnici relativi alle prestazioni del decoder in un determinato istante (es. CPU utilizzata, percentuale RAM, spazio della memoria interna utilizzato).

## 4. SVILUPPO SECONDA FASE PROGETTUALE

### 4.1 Esplorazione del dataset

Nello sviluppo della seconda fase progettuale è stato eseguito il processo di esplorazione del dataset, evidenziando alcune problematiche dovute alla manipolazione del grosso quantitativo di dati a disposizione. È stata quindi creata una documentazione apposita relativa ai difetti riscontrati nel dataset studiato.

Sulla base della documentazione fornita e grazie al continuo contatto con il cliente sono stati selezionati i 9 disservizi più impattanti. Per trarre il massimo valore aggiunto dalle informazioni è stata effettuata un'estrazione dei dati dal database non relazionale MongoDB. Questa decisione è dovuta dal fatto che, per poter operare efficacemente sulle informazioni, è necessaria una forma di dati strutturata. Nell'infrastruttura "Cubo" infatti, vengono caricate informazioni di sintesi (aggregate), portando incertezze ed inefficienze nell'attività di prelievo ed analisi del contenuto informativo.

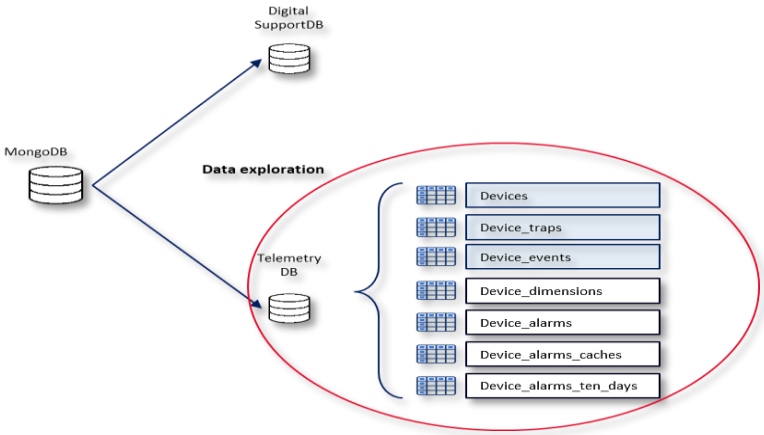


*Figura 4. Esplorazione del Dataset, in azzurro le tabelle sottoposte ad un'esplorazione approfondita*

La figura 4 illustra le principali *collection* analizzate durante la prima iterazione esplorativa.

Dato l'enorme quantitativo di dati (circa 550 GB) riguardante circa l'ultimo mese di monitoraggio, per poter iniziare ad effettuare un'esplorazione più rapida ed efficiente è stato estratto un campione relativo ad una giornata (13 Febbraio). Successivamente sono state applicate le analisi all'intero dataset utilizzando tecnologie performanti per il calcolo distribuito (tecnologia Hadoop-Spark).

Questo processo è stato reiterato più volte al fine di supportare in maniera solida le analisi successive. Il dataset inizialmente analizzato presentava 1.643.155 record (righe), il cui 5% rappresentava eventi di disservizio.

Da qui è stata posta una prima analisi:

| Descrizione | Numerosità | Media di disservizi per utente |
|---|---|---|
| Numero di disservizi registrati in giornata | 79243 | 6.39 |
| Numero di decoder che hanno registrato almeno un disservizio | 12395 | |

*Tabella 2. Prime evidenze riscontrate nell'analisi giornaliera del dataset*

## 4.2 Analisi delle evidenze

L'analisi delle evidenze è stata eseguita con un duplice scopo. In primo luogo, per informare il cliente sui progressi effettuati ed espandere la sua percezione riguardo variabili che, allo stato AS-IS, non monitorava (inizialmente non erano presenti sulla dashboard). In secondo luogo, per incrementare la robustezza del processo costruttivo di un modello predittivo, infatti tale procedura si avvaleva di tecniche statistiche. L'attività è stata effettuata utilizzando Python e le sue librerie associate per poter operare sui dataset, studiando minuziosamente le variabili che componevano le tabelle ed effettuando operazioni di pulitura su valori che presentavano difetti di importazione. In questa prima iterazione, sono state identificate tre evidenze principali:

- *Disservizi*: su 9 disservizi segnalati dalle interviste e dalla mappatura della documentazione, solamente 3 costituiscono più del 95% della numerosità totale, incidendo in maniera significativa sul campione analizzato. In particolare, i 3 disservizi influenti sono legati ad eventi di tipo "buffering" ("TtFP", "Buffer_underflow" ed "End_buffering").
- *SNR:* il rapporto segnale-rumore, il quale identifica la qualità del segnale durante la fruizione di un contenuto, incide sugli eventi allo stesso modo. È stata effettuata un'analisi della variabile su tutti gli eventi (eventi di disservizio e non). Per quanto riguarda i disservizi, a seconda di questi, il valore SNR presenta range leggermente differenti, tuttavia non risulta possibile classificarli utilizzando tale variabile (se pur influente quando si evidenziano errori di tipo "buffering").

  Sopra la soglia di valore 90 non viene segnalato alcun disservizio.

  Il box-plot seguente mostra la distribuzione dell'SNR in relazione ai disservizi sottoposti ad analisi.

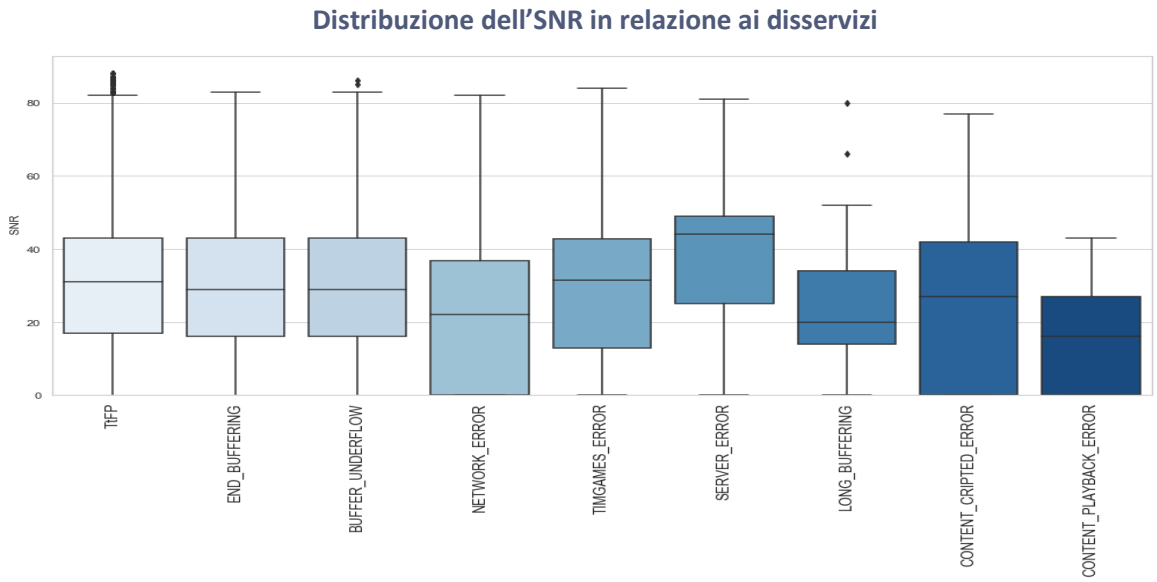


*Figura 5 Distribuzione giornaliera del valore SNR (asse y) in relazione ai disservizi (asse x)*

- *Relazione tra disservizi e range orari:* la maggior parte dei disservizi si presenta nella fascia oraria tra le 19 e le 22. È stata eseguita un'analisi dei disservizi suddividendola in range orari (intervallo di 1 ora), osservando che la numerosità degli errori incrementa in maniera significativa nella fascia appartenente alla seconda serata (Figura 6).

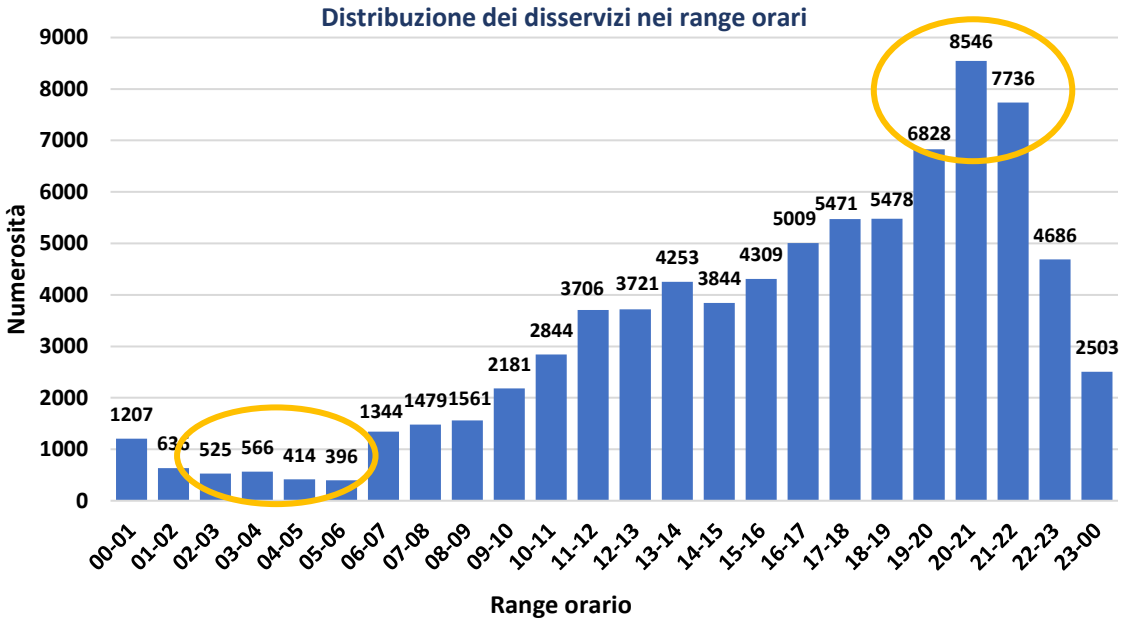


*Figura 6 Distribuzione giornaliera disservizi nei range orari*

Oltre a ciò è stato effettuato uno studio in cui viene osservato che, a prescindere dal range orario, i disservizi sono distribuiti nello stesso modo.

Tale studio consiste nell'analizzare la ripartizione dei singoli disservizi nelle aree temporali che, in termini di numerosità totale, costituiscono il "picco più elevato" (h: 19-22) e la "valle più bassa" (h: 02-06) di figura 6.

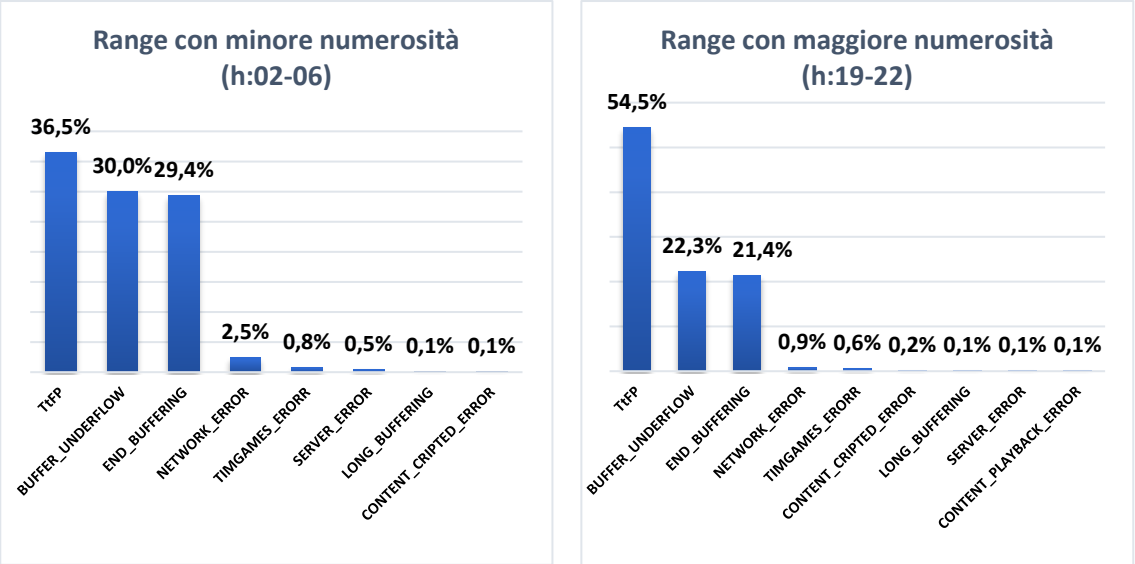


*Figura 7 Confronto distribuzione disservizi nei range orari selezionati*

La figura 7 evidenzia un pattern simile nella distribuzione dei disservizi a meno di un incremento significativo dell'evento "TtFP" (acronimo di "time to first picture": tempo che intercorre tra la richiesta effettuata dall'utente per visualizzare il video on demand e la comparsa del primo frame).

## 5. RISULTATI RAGGIUNTI E SUCCESSIVI PASSI

### 5.1 Evidenze finali

Per rendere significative le evidenze osservate è stato necessario effettuare un'analisi su più giornate. Il dataset finale utilizzato nello studio presentava 25778309 occorrenze registrate dal 04 al 17 Febbraio 2019.

Inizialmente viene riportata la tabella che restituisce la media dei disservizi per utente:

| Descrizione | Numerosità | Media di disservizi per utente |
|---|---|---|
| Numero di disservizi registrati nel periodo | 1289933 | 7,5 |
| Numero di decoder che hanno registrato almeno un disservizio | 171723 | |

*Tabella 2. Evidenze riscontrate durante l'analisi settimanale*

Sono stati analizzati i 9 disservizi presenti in documentazione arricchendo lo studio precedentemente fatto con ulteriori osservazioni.

- *Disservizi*: I tre disservizi principali ("TtFP", "Buffer_underflow" ed "End_buffering") rappresentano anche qui più del 97% della numerosità del campione.

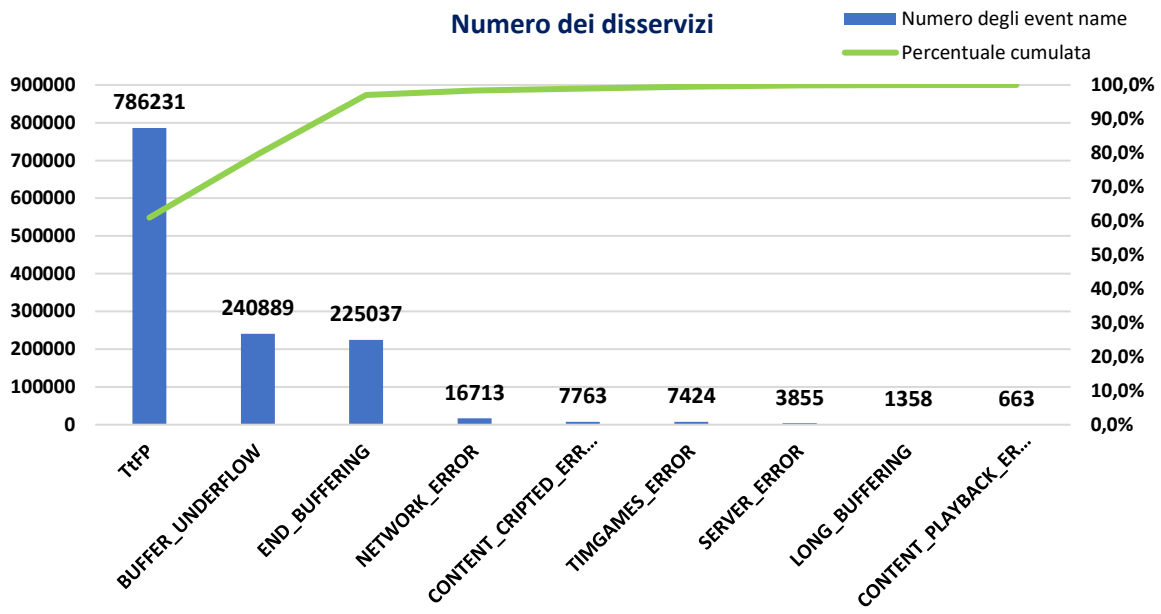


*Figura 8. Distribuzione disservizi in termini di numerosità*

- *SNR*: come era stato evidenziato durante l'estrazione giornaliera è stato visto che l'SNR incide sugli eventi allo stesso modo, pur presentando range differenti a seconda del disservizio. Sopra il valore di 90 non viene segnalato alcun disservizio ed il 75% dei valori SNR si trova nel range compreso tra 0 e 50. È da osservare tuttavia che il valore SNR per quanto riguarda i disservizi impattanti ("TtFP", "Buffer_underflow" e "End_buffering") ha un range quasi identico.

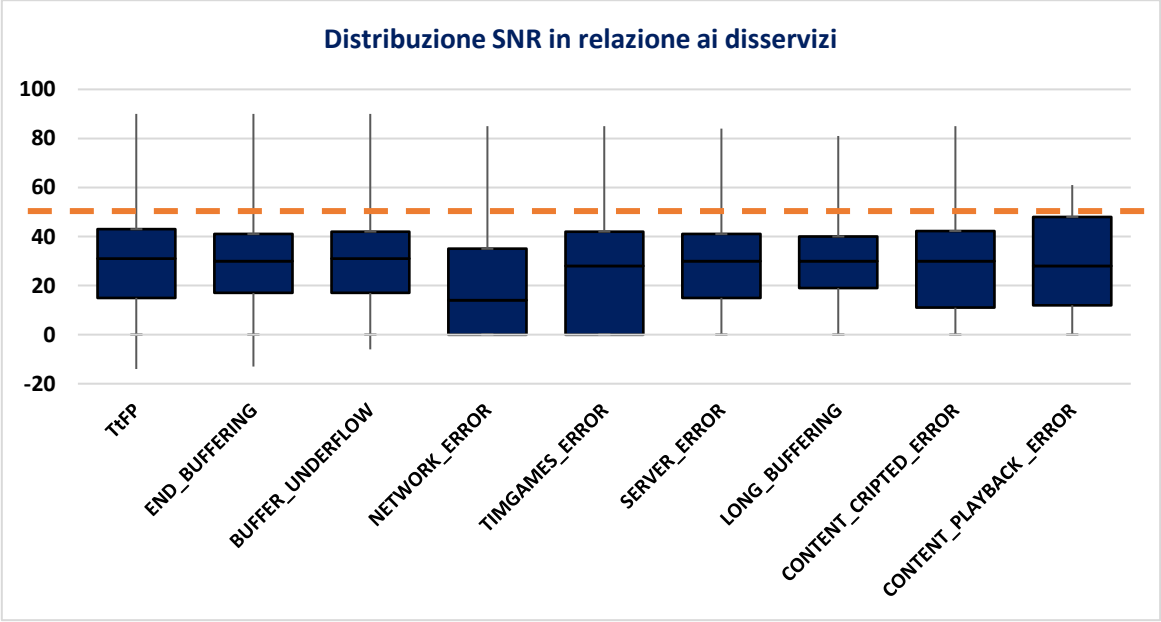


*Figura 9. Distribuzione SNR settimanale per i disservizi*

- *Relazione tra disservizi e range orari*: come dalle analisi precedenti è stato osservato che, la maggior parte dei disservizi, si presenta nella fascia oraria tra le 19 e le 22. La figura sottostante riporta la numerosità media giornaliera dei disservizi.

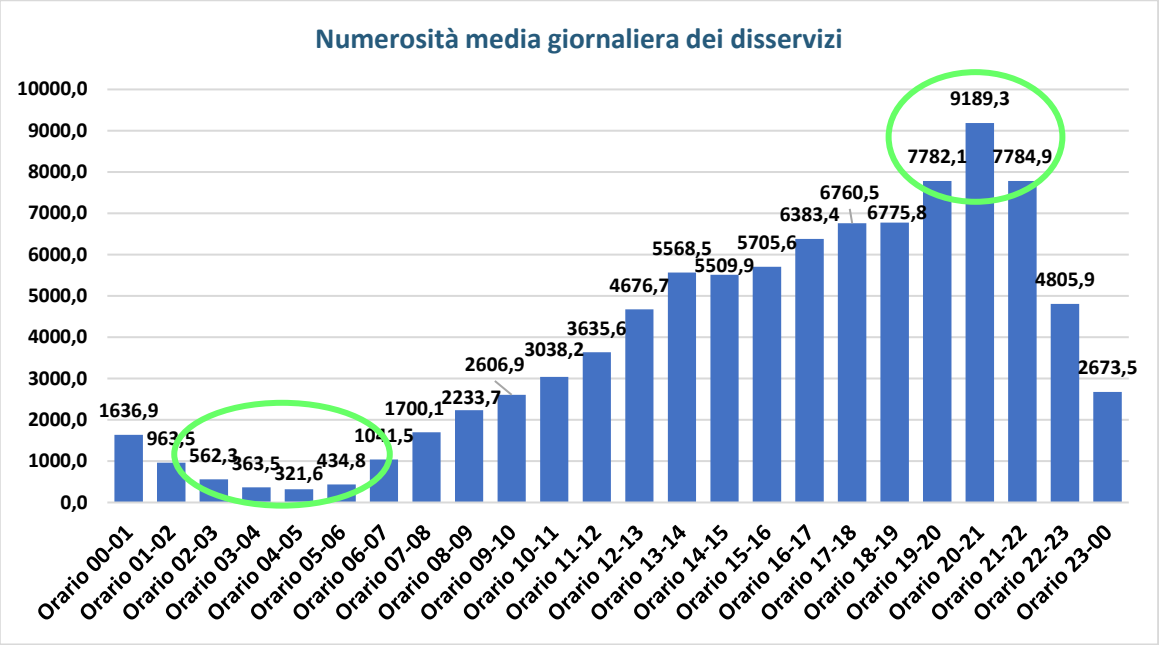


*Figura 10 Distribuzione settimanale media dei disservizi, suddivisa per range orari*

È stato eseguito il solito procedimento effettuato durante lo studio giornaliero, analizzando gli intervalli temporali che presentano i valori massimi e minimi in termini di numerosità dei disservizi (nella figura sopra evidenziati dal cerchio verde).

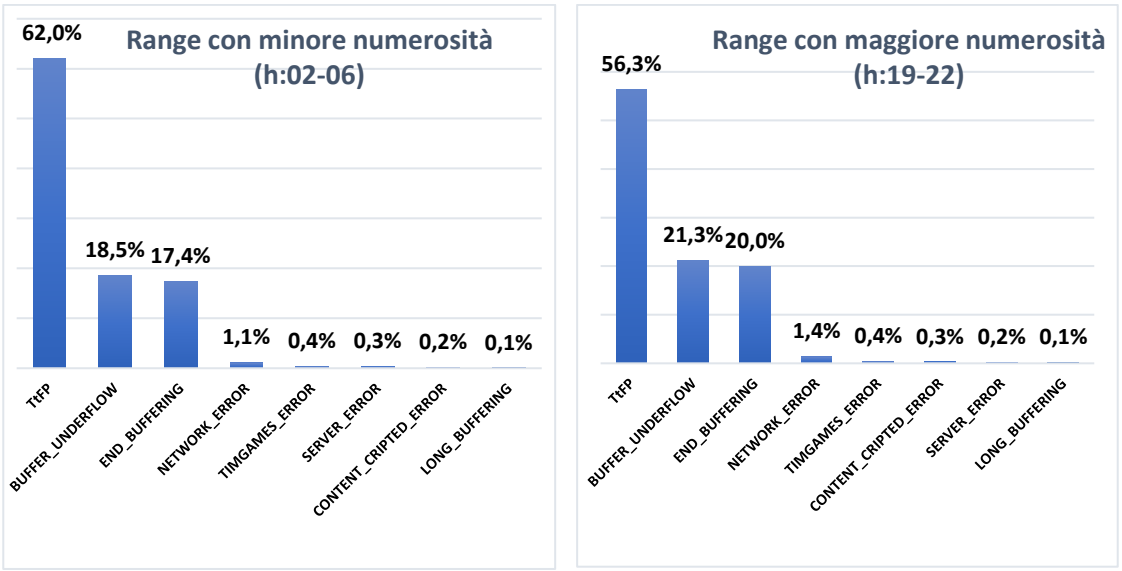


*Figura 11 Confronto distribuzione disservizi settimanale, nei range orari selezionati*

Da qui è possibile osservare che i disservizi impattanti mantengono un pattern regolare.

## 5.2 Conclusioni

La parte conclusiva ripercorre i risultati raggiunti in relazione alle richieste iniziali effettuate dal cliente.

Per quanto riguarda l'attività di reingegnerizzazione della dashboard, molte delle proposte di modifica delle viste sono in fase di implementazione e sono state accolte dall'azienda. Grazie alle evidenze riscontrate potrebbe essere possibile implementare successivamente nuovi KPI che pongano un livello di dettaglio granulare aggiuntivo per la visualizzazione delle occorrenze.

Per quanto riguarda l'evidenza relativa ai range orari, è possibile affermare che i disservizi impattanti mantengono un pattern regolare, indipendentemente dalle fasce orarie soggette ad una maggiore utenza. Questo può far escludere un eventuale problema legato al server di rete che, nel caso di sovraffollamento degli utilizzatori, avrebbe potuto impattare negativamente sulle prestazioni.

Al termine del periodo è stato creato un modello di classificazione binaria, che identifica tramite correlazioni, quali attributi sono associati ai disservizi. Questo ultimo tuttavia non è stato consegnato al cliente in quanto, per iniziare ad ottenere percentuali di affidabilità elevata, sono necessarie diverse operazioni di pulitura del dataset.

Il progetto svolto ha portato alla luce alcune criticità quando si opera nell'ambito big data.QUALI?

A Giugno partirà un nuovo progetto collaborativo, in ottica di miglioramento continuo, al fine di arricchire con ulteriori studi le evidenze riscontrate e creare un modello predittivo efficace e consistente.

## APPENDICE

Lo scopo dell'appendice è quello di descrivere l'esperienza offerta dal programma formativo di Junior Consulting. Verrà descritta la metodologia con la quale Il programma colma il gap tra il mondo universitario e lavoro.

Durante il periodo magistrale presso l'Università di Pisa, in un percorso ricco di alti e bassi, di rallentamenti, di ostacoli, di gioie e di soddisfazioni, la professoressa Antonella Martini mi ha proposto di partecipare al programma JC. Personalmente l'ambito della consulenza era un mondo che mi affascinava così, pur mancandomi qualche esame, ho deciso letteralmente di buttarmi in questo percorso. Il programma JC ha soddisfatto a pieno i miei interessi guidandomi parallelamente sia sull'acquisizione di competenze tecniche sia sullo sviluppo personale.

Per quanto riguarda lo sviluppo personale, JC offre un mese di formazione durante il quale viene svolto un piano educativo al fine di migliorare le soft skill, affrontando le seguenti aree:

Team Building, Personal Leadership, Public Speaking, Stress Management, Self-Marketing, gestione del PM ed utilizzo dei principali software aziendali (Outlook, Excell, Word e Power Point).

Per quanto riguarda le competenze tecniche, nei successivi mesi viene affrontato un vero e proprio progetto di consulenza che permette letteralmente di mettersi in gioco, interfacciandosi con clienti e figure professionali appartenenti al top management. Inoltre, è comunque presente una figura senior (team - leader) che ha lo scopo di fare da tutor ed insegnare, a noi laureandi, una metodologia corretta per poter affrontare questa tipologia di lavoro.

Ho partecipato all'edizione JC-35, un'edizione soprannominata "pisana" (il nome è dovuto dal fatto che 6 dei 12 talenti Junior provenivano dall'università di Pisa), questo ha permesso di creare un forte legame con i colleghi e con lo staff facendo modo di distinguerci anche dalle edizioni precedenti.

In un contesto turbolento dove il mercato del lavoro è mutevole e in continuo cambiamento, il programma JC ha permesso di orientarmi in maniera proattiva su questa strada facendomi capire che esiste un gap elevato tra il mondo universitario e quello lavorativo.

Mi sento vivamente di dire che anche se non esiste il momento perfetto a volte devi avere l'ardire di buttarti.

Consiglio a tutte le persone che si vogliono mettere in gioco di farlo il prima possibile.