

LOAN APPLICATION

Ikram Bourras

Business Information Systems

Prof. Paolo Ceravolo

a.a. 2024 - 2025

Contents

1. Introduzione
2. Descrizione e organizzazione del dataset
 - 2.1 lunghezza log
 - 2.2 Volume dei Dati
 - 2.3 Attributi dei Dati
 - 2.3.1 A livello di Caso
 - 2.3.2 A livello di Evento
 - 2.3.3 Attributi Chiave del Log
- 3.Organisational Goals
 - 3.1 Comprendere i processi aziendali sottostanti
 - 3.2 Identificare opportunità per migliorare l'efficienza e l'efficacia
4. Definizione degli obiettivi strategici dell'organizzazione (KPI)
5. Definizione degli obiettivi operativi e tattici
6. Knowledge Uplift Trail
 - 6.1 Data Cleaning
 - 6.1.1 Normalizzazione dei attributi
 - 6.1.2 Riconciliazione di duplicati e incoerenze.
 - 6.2 Data Filtering
 - 6.2.1 Filtering Noise
 - 6.2.2 Filering Irrelevant Data W_Valideren aanvraag
 - 6.3 Process Discovery
 - 6.4 Descriptive analysis
 - 6.6 Conformance checking
 - 6.2.2 Filering Irrelevant Data df_AO_dc
 - 6.3 Process Discovery
 - 6.4 Descriptive analysis
 - 6.6 Conformance checking
 - 6.7 Intervention strategies

6.7.1 Sviluppi futuri

1. Introduzione

Questo progetto si propone di condurre un'analisi approfondita del processo di gestione delle richieste di prestito all'interno di un istituto finanziario olandese. Utilizzando un **log di eventi reale** proveniente dal BPI Challenge 2012, che comprende **13.087 casi e 262.200 eventi**, ci immergeremo nelle dinamiche operative per identificare inefficienze e opportunità di miglioramento. L'analisi si focalizzerà su tre sottoprocessi interconnessi: la **gestione delle domande** (dalla presentazione alla decisione finale), la **gestione delle offerte** (dalla preparazione all'invio) e le **attività manuali** legate ai *work item*, come i controlli antifrode e i *follow-up*. Particolare attenzione verrà posta all'attributo **AMOUNT_REQ (importo richiesto)**, un fattore chiave che potrebbe influenzare significativamente i tempi di elaborazione e gli esiti finali. L'obiettivo è rilevare i numerosi punti decisionali e le attività manuali che caratterizzano il processo, individuando i potenziali **colli di bottiglia** per ottimizzare l'efficienza complessiva.

2. Descrizione e organizzazione del dataset

2.1 lunghezza log

Il log descrive gli eventi legati al processo di richiesta di prestiti personali e scoperti bancari, registrati in un periodo di circa sei mesi: dal 1° ottobre 2011 al 14 marzo 2012.



```
#periodo del log
```

```
import pandas as pd
# convert to datetime
df['time:timestamp'] = pd.to_datetime(df['time:timestamp'])
# find min and max date
min_date = df['time:timestamp'].min()
max_date = df['time:timestamp'].max()
# print difference
print(f"Min date: {min_date}")
print(f"Max date: {max_date}")
print(f"Difference: {max_date - min_date}")
```



```
Min date: 2011-10-01 00:38:44.546000
Max date: 2012-03-14 16:04:54.681000
Difference: 165 days 15:26:10.135000
```

2.2 Volume dei Dati

- 13.087 casi
- 262.200 eventi totali
- 24 classi di eventi uniche all'inizio dell'analisi



```
# prompt: numero dei casi, eventi totali e numero classi di eventi uniche

print("Numero di casi:", df['case:concept:name'].nunique())
print("Eventi totali:", len(df))
print("Numero di classi di eventi uniche:", df['concept:name'].nunique())
```



```
Numero di casi: 13087
Eventi totali: 262200
Numero di classi di eventi uniche: 24
```

2.3 Attributi dei Dati

Il dataset contiene sia attributi a livello di caso sia a livello di evento

2.3.1 A LIVELLO DI CASO

- **AMOUNT_REQ**: importo richiesto dal cliente
- **case:reg_date**: data di registrazione dell'applicazione
- **case:concept:name**: Identificativo unico del caso (case ID)

2.3.2 A LIVELLO DI EVENTO

- **time:timestamp**: data e ora esatte di completamento dell'evento
- **concept:name**: nome dell'attività o fase del processo
- **lifecycle:transition**: fase del ciclo di vita dell'attività (Schedule, Start, Complete)
- **org:resource**: risorsa (persona o sistema) responsabile dell'evento

2.3.3 ATTRIBUTI CHIAVE DEL LOG

org:resource

I campo identifica la persona, il sistema o il reparto che svolge un'attività nel processo, sono 68 le risorse uniche.

```
[#] # prompt: org:resource valori unici
print("\nValori unici per 'org:resource':", df['org:resource'].unique())
print("Numero di valori unici per 'org:resource':", df['org:resource'].nunique())

[+] Valori unici per 'org:resource': ['112' nan '10862' '10913' '11049' '10629' '11120' '10809' '10912' '11201'
'11119' '10861' '11203' '11181' '11189' '10609' '11111' '10982' '11019'
'11180' '10899' '10138' '11002' '11122' '10889' '10972' '11121' '10939'
'11029' '11009' '11000' '10863' '11169' '11179' '11001' '10971' '10228'
'11202' '10789' '10881' '10909' '10188' '10910' '10929' '10931' '11259'
'11200' '10779' '10880' '10914' '10859' '11339' '10933' '11079' '10932'
'10935' '11254' '11003' '10125' '11269' '10821' '11289' '10124' '11299'
'11309' '11300' '11302' '11319' '11304']
Numero di valori unici per 'org:resource': 68
```

Questa informazione è **fondamentale** per l'analisi organizzativa perché si può andare a capire Distribuzione del carico di lavoro → capire chi lavora di più.

```
Statistics for 'org:resource':
Max count: 45687
Resource(s) with max count:
['112']

Min count: 2
Resource(s) with min count:
['10821']

Median count: 2413.5
Mode count(s):
[6]
Resource(s) with mode count(s):
['10125', '11269']
```

Il carico di lavoro delle 68 risorse varia molto: da 2 a quasi 46.000 eventi. La risorsa con il massimo carico è un sistema automatizzato ('112'). La mediana (2413.5) indica che metà delle risorse ha un carico di lavoro inferiore a circa 2400, mentre la moda (6) mostra che molte risorse hanno un carico basso, segnalando un forte squilibrio tra risorse con carico alto (come sistemi automatizzati) e risorse con carico basso (umane o poco usate).

lifecycle:transition

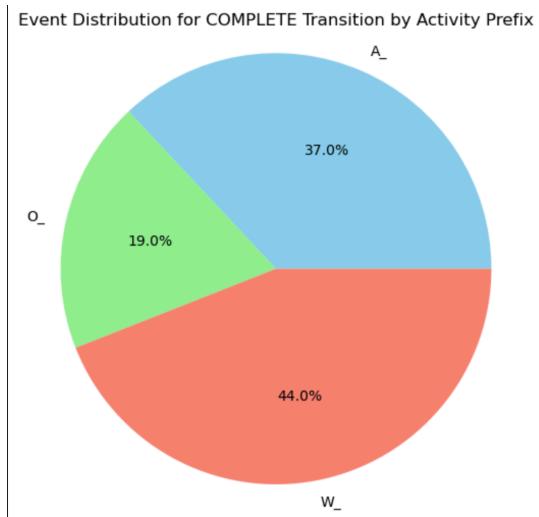
Indica la fase del ciclo di vita dell'attività.

```
[24] #lifecycle:transition valori unici
print("\nValori unici per 'lifecycle:transition':", df['lifecycle:transition'].unique())
print("Numero di valori unici per 'lifecycle:transition':", df['lifecycle:transition'].nunique())

[+] Valori unici per 'lifecycle:transition': ['COMPLETE' 'SCHEDULE' 'START']
Numero di valori unici per 'lifecycle:transition': 3
```

- COMPLETE: Indica la conclusione di un'attività o di un work item. Segna il momento in cui l'attività è stata portata a termine.
- SCHEDULE: Indica che un work item è stato **pianificato o messo in coda** per essere eseguito in futuro

- START: Indica l'**avvio** dell'esecuzione di un'attività o di un work item
Per attività di tipo **A_ (Application)** e **O_ (Offer)**, l'evento di **complete** è l'unico registrato. Per attività di tipo **W_**, si distinguono: schedule (programmato), start (avviato) e complete (completato).



concept:name

Eventi dell'Applicazione (A_)

Tracciano lo stato complessivo dell'applicazione, dal momento in cui il cliente la presenta fino all'esito finale:

- **A_SUBMITTED / A_PARTLYSUBMITTED**: richiesta inviata, completa o parziale.
- **A_PREACCEPTED**: accettazione preliminare, in attesa di informazioni aggiuntive.
- **A_ACCEPTED**: richiesta accettata, in attesa di verifica.
- **A_FINALIZED**: richiesta completa, pronta per la decisione finale.
- **A_APPROVED / A_REGISTERED / A_ACTIVATED**: applicazione approvata e, in alcuni casi, attivata.
- **A_CANCELLED**: richiesta annullata.
- **A_DECLINED**: richiesta rifiutata.

Eventi dell'Offerta (O_)

Si concentrano sulla gestione delle offerte verso il cliente:

- **O_SELECTED / O_PREPARED / O_SENT / O_SENT BACK / O_ACCEPTED**: diverse fasi dell'offerta.
- **O_CANCELLED / O_DECLINED**: annullamento o rifiuto dell'offerta.

Eventi del Flusso di Lavoro (W_)

Riguardano le attività manuali e le verifiche svolte dal personale:

- **W_Afhandelen leads**: gestione delle richieste iniziali.
- **W_Completeren aanvraag**: completamento delle informazioni.
- **W_Nabellen offertes**: follow-up dopo l'invio delle offerte.

- **W_Valideren aanvraag:** validazione finale della richiesta.
- **W_Nabellen incomplete dossiers:** recupero di informazioni mancanti.
- **W_Beoordelen fraude:** controllo frodi.
- **W_Wijzigen contractgegevens:** modifica dati del contratto.

time:timestamp

Registra data e ora dell'evento. È essenziale per ordinare gli eventi, calcolare le performance e analizzare colli di bottiglia e tempi di esecuzione.

case:REG_DATE

Indica la data di registrazione dell'applicazione e consente di analizzare come le performance cambiano nel tempo.

case:concept:name

Identificativo unico del caso (case ID), che collega tutti gli eventi della stessa richiesta e permette di formare le **tracce** nel Process Mining.

case:AMOUNT_REQ

Importo del prestito richiesto. È utile per segmentare le richieste e analizzare l'impatto dell'importo sull'esito o sulla durata del processo.

3.Organisational Goals

Il progetto di analisi del processo di richiesta di prestiti/scoperti bancari ha un duplice scopo principale, orientato sia a una comprensione approfondita dei processi aziendali sottostanti sia al miglioramento continuo delle performance operative. Questa finalità si allinea all'obiettivo generale della Business Intelligence (BI), che è fornire un supporto decisionale basato su dati empirici e metriche misurabili

3.1 Comprendere i processi aziendali sottostanti

L'obiettivo è ottenere una visione dettagliata di **come il processo si svolge realmente**, basandosi sui dati degli eventi raccolti nei log. Gli elementi principali di questa analisi includono:

- **Frequenza e distribuzione degli esiti delle richieste:** Calcolare e visualizzare il numero e la percentuale di casi conclusi con esiti quali **Approvato**, **Rifiutato** (ad esempio, evento finale A_DECLINED), **Annullato** (A_CANCELLED, O_CANCELLED), e casi ancora aperti. L'analisi ha evidenziato che l'esito A_DECLINED è quello più frequente tra i rifiuti, mentre A_APPROVED non è mai evento finale, indicando che chiudere casi su questo evento non riflette la realtà operativa.
- **Tempi medi di ciclo per esito:** Calcolare il throughput time medio per ogni categoria di esito, visualizzando la distribuzione dei tempi tramite boxplot o violinplot per identificare variazioni e outlier.
- **Mappatura e scoperta del modello di processo (AS-IS):** Utilizzare strumenti di Process Mining come pm4py per generare modelli basati su reti di Petri inductive (con soglia di inclusione pari a 0.9), Directly-Follows Graph o Heuristic Miner, al fine di rappresentare fedelmente il flusso effettivo delle richieste e ridurre il rumore.
- **Analisi delle varianti del processo:** Contare le varianti presenti (tramite funzioni come get_variants_df di pm4py), identificare le varianti più comuni e valutarne la frequenza, al fine di capire le differenze operative tra i casi.
- **Analisi delle risorse e della loro attività:** Esaminare la distribuzione delle attività tra le risorse coinvolte, verificando chi svolge multitasking e chi invece ha un ruolo più specializzato, per studiare l'impatto sulla performance.

3.2 Identificare opportunità per migliorare l'efficienza e l'efficacia

- **Individuazione di colli di bottiglia e attività lente:** Calcolare i tempi medi delle singole attività e identificare le fasi critiche con tempi di esecuzione elevati che rallentano l'intero processo.
- **Analisi comparativa tra risorse multitasking e non multitasking:** Confrontare tempi di esecuzione e turnaround per capire se il multitasking influisce positivamente o negativamente sulle performance..
- **Conformance checking:** Allineare i log reali al modello di processo AS-IS per identificare deviazioni significative e valutarne l'impatto operativo e sulla qualità.
- **Miglioramento della qualità e affidabilità dei dati:** Procedere alla pulizia e normalizzazione dei log per un'analisi più completa e precisa.

4. Definizione degli obiettivi strategici dell'organizzazione (KPI)

I KPI sono misure quantitative che collegano direttamente le attività del business agli obiettivi strategici, operativi e tattici. La loro definizione deriva dalla comprensione del processo e dai requisiti specifici, misurando risultati in confronto agli obiettivi prestabiliti.

4.1 KPI Quantitativi

- Frequenza e distribuzione degli esiti (**Approvato, Rifiutato, Annullato, In Sospeso**).
- Tempi medi di ciclo per esito.
- Numero e frequenza delle varianti del processo.
- Tempi medi e deviazioni standard delle singole attività.
- Percentuale di casi che superano la soglia temporale di 30 giorni (che può causare cancellazione automatica).

4.2 KPI Pratici

- Numero di domande processate (sottomesse, accettate, rifiutate).
- Numero di offerte di prestito inviate (**O_SENT**).
- Percentuale di attività svolte per risorsa, con verifica del multitasking.
- Identificazione e localizzazione dei colli di bottiglia (attività con throughput time superiore alla media).
- Differenze di performance tra risorse multitasking e non multitasking.

4.3 KPI Direzionali

- Analisi temporale della distribuzione delle attività (es. dotted chart).
- Evoluzione nel tempo dei tassi di approvazione e rifiuto.
- Percentuale di deviazioni dal modello AS-IS (conformance checking).
- Trend di deviazione e non conformità nel tempo.

4.4 KPI Azionabili

- Mappatura del modello di processo AS-IS (Directly-Follows Graph, Heuristic Miner).
- Individuazione di attività a rischio (fasi con tempi elevati o anomalie/outlier).
- Identificazione di varianti critiche con performance peggiori o migliori.
- Modelli predittivi per stimare probabilità di successo o ritardo.

4.5 KPI Finanziari

- Importo medio richiesto per esito (**AMOUNT_REQ**).
- Distribuzione degli importi per casi approvati e rifiutati.

5. Definizione degli obiettivi operativi e tattici

Il successo organizzativo si basa sull'integrazione di due forme di conoscenza fondamentali:

- Descriptive Knowledge: come sono realmente le cose, basata su dati osservati e KPI che misurano le performance attuali.
- Prescriptive Knowledge: come dovrebbero essere, indicando azioni necessarie per raggiungere obiettivi, guidata da KPI target e best practice.

5.1 Obiettivi Strategici

- Descriptive Knowledge
- Nel report ho applicato strumenti di esplorazione dati (KUT) e analisi descrittiva per comprendere lo stato attuale dell'organizzazione a livello macro. Questo include la pulizia dei dati (data cleaning) e la selezione di dati rilevanti tramite filtri per assicurare la qualità dell'analisi.
- L'obiettivo è stato ottenere una fotografia precisa del contesto, individuare trend e pattern nei dati aggregati, senza definire azioni o target specifici.

5.2 Obiettivi Tattici

- Descriptive Knowledge
- A livello tattico, il lavoro ha riguardato l'analisi dettagliata di subset di dati (filtraggio e segmentazione), per evidenziare comportamenti o anomalie specifiche in determinati periodi o aree.
- L'obiettivo è stato descrivere con chiarezza come si comportano le singole componenti organizzative o processi, basandosi sui dati effettivi, per permettere una base solida a future analisi o decisioni.

5.3 Obiettivi Operativi

- Descriptive Knowledge
- A livello operativo, ho esaminato dati di dettaglio a livello giornaliero o di singole transazioni, utilizzando filtri e tecniche di pulizia dati per rimuovere rumore e outlier.
- L'obiettivo è stato garantire che i dati analizzati siano accurati e affidabili, per descrivere fedelmente le attività quotidiane, senza però intervenire direttamente con azioni correttive o modifiche operative.

6. Knowledge Uplift Trail

6.1 Data Cleaning

Data cleaning prepares the event log by standardizing data, filling missing values or converting categorical variables to numerical ones;

6.1.1 NORMALIZZAZIONE DEI ATTRIBUTI

In questo caso il tipo degli attributi sono giusti, infatti abbiamo date:time per time:timestamp e case:Reg_date

```
▶ #controllare gli attributi di che tipo sono

print("\nData types of attributes:")
for col in df.columns:
    print(f"- {col}: {df[col].dtype}")

→ Data types of attributes:
- org:resource: object
- lifecycle:transition: object
- concept:name: object
- time:timestamp: datetime64[ns]
- case:REG_DATE: datetime64[ns]
- case:concept:name: object
- case:AMOUNT_REQ: object
```

6.1.2 RICONCILIAZIONE DI DUPLICATI E INCOERENZE.

- confrontato START e COMPLETE per tutte le attività W_.
- identificato le attività incoerenti.
- associato i COMPLETE ai rispettivi START, eliminando i COMPLETE "orfani" per ripristinare la coerenza.

```
▶ #Confrontare il numero di eventi START e COMPLETE per ogni attività che inizia con W_

# Filter events starting with 'W_'
df_W = df[df['concept:name'].str.startswith('W_')]

# Count START and COMPLETE transitions for these events
start_W_count = len(df_W[df_W['lifecycle:transition'] == 'START'])
complete_W_count = len(df_W[df_W['lifecycle:transition'] == 'COMPLETE'])

print(f"\nNumero di eventi 'START' per attività che iniziano con 'W_': {start_W_count}")
print(f"Numero di eventi 'COMPLETE' per attività che iniziano con 'W_': {complete_W_count}")
print(f"differenza con 'W_': ", complete_W_count - start_W_count)

→ Numero di eventi 'START' per attività che iniziano con 'W_': 71376
Numero di eventi 'COMPLETE' per attività che iniziano con 'W_': 72413
differenza con 'W_': 1037
```

```
▶ print( len(df),"-->",len(df_cleaned))

→ 262200 --> 261160
```

6.2 Data Filtering

Data filtering sets conditions to filter the data, keeping only cases that meet specific criteria

6.2.1 FILTERING NOISE

Informazioni errate, imprecise o incomplete nel log di eventi.

- Ordinato il log per caso (case:concept:name) e timestamp (time:timestamp) per garantire la corretta sequenza temporale.
- Estratto per ogni caso il primo e il secondo evento (concept:name) e calcolata la frequenza di occorrenza.
- Calcolata la durata di ogni caso come differenza tra ultimo e primo timestamp.
- Unito gli eventi consecutivi A_SUBMITTED e A_PARTLYSUBMITTED in un unico evento A_SUBMITTED/A_PARTLYSUBMITTED, usando il timestamp del secondo evento per ridurre ridondanze.

Vantaggi del mio approccio

- **Semplificazione dell'analisi:** Riduco il numero di stati distinti, rendendo i log più semplici da leggere e interpretare.
- **Maggiore focus sulla rilevanza:** Mi concentro sugli eventi che hanno un impatto effettivo o una durata misurabile.
- **Riduzione del “rumore”:** Elimino eventi che potrebbero essere considerati dettagli secondari o “rumore” nella mia analisi principale.
- **Migliore rappresentazione del processo:** Poiché "A_PARTLYSUBMITTED" è solo un'eco o una conferma temporanea di "A_SUBMITTED", consolidarli mi permette di avere una visione più pulita e lineare del flusso reale.

Successivamente, ho notato anche la presenza di altri due eventi che terminano con "schedule". Tuttavia, dato che questi eventi saranno già filtrati nella fase successiva, non li ho uniti in questo momento.

Infine, avevo preso in considerazione anche l'idea di raggruppare gli eventi identici che si ripetono in sequenza. Tuttavia, ho verificato che la computazione richiesta per questo tipo di raggruppamento è troppo elevata rispetto ai benefici che potrei ottenere. Per questo motivo ho deciso di non procedere in questa fase.

Gestione Casi Completi e Incompleti

- Diviso il dataset in due parti:
 - **Casi andati a buon fine:** quelli con attività finale W_Valideren aanvraag, che indica conclusione regolare del caso (come da documentazione).
 - **Casi rifiutati:** includono eventi finali A_DECLINED, A_CANCELLED e O_CANCELLED A_APPROVED
- Filtro ulteriore:
 - Solo i casi che finiscono con W_Valideren aanvraag con ultima attività marcata come complete nel campo lifecycle:transition vengono considerati validi e mantenuti.

- Eventuali casi senza questa conferma sono considerati incompleti o irrilevanti e quindi eliminati.

```
array(['COMPLETE', 'START'], dtype=object)
```

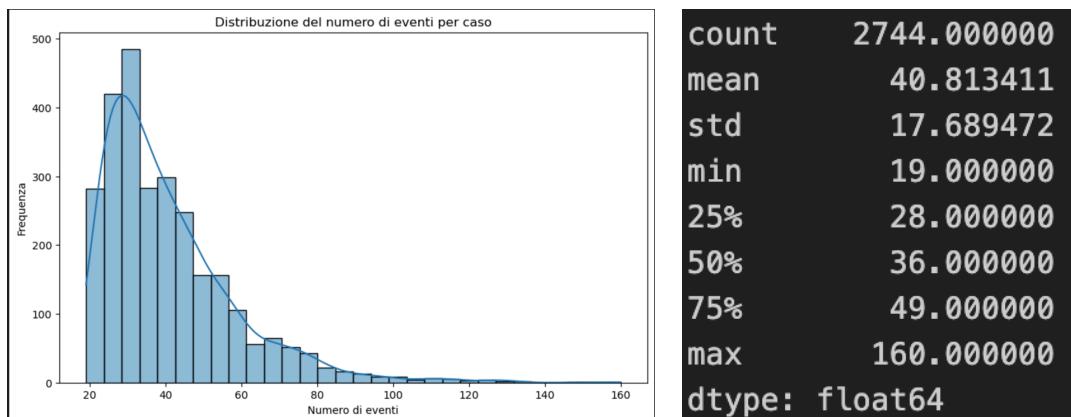
```
array(['COMPLETE'], dtype=object)
```

6.2.2 FILTERING IRRELEVANT DATA W_VALIDEREN AANVRAAG

Focalizzarsi su specifici segmenti del processo.

LAVORO CON DF che finisce con attività **W_Valideren aanvraag**

Idea è quella di rimuovere casi anomali o eccezionali, come quelli troppo lunghi o troppo corti, per concentrarsi sui casi rappresentativi.



Metodo:

- Calcolate le lunghezze delle sequenze (numero di eventi per caso).
- Identificati i quartili Q1 (25° percentile) e Q3 (75° percentile) delle lunghezze.
- Segmentati i casi in tre gruppi:
 - **Short**: lunghezza \leq Q1
 - **Medium**: $Q1 < \text{lunghezza} \leq Q3$
 - **Long**: lunghezza $> Q3$

L'idea è utilizzare principalmente i casi **short**, perché questi rappresentano situazioni in cui il processo è andato senza troppi intoppi: l'attività finale **W_Valideren aanvraag** indica un esito positivo (prestito rilasciato e utente soddisfatto dell'importo).

I casi **short** permettono di osservare il processo “ideale”, senza difficoltà o ritardi.

I casi **medium** e **long** verranno invece mantenuti per analisi mirate, per identificare colli di bottiglia, difficoltà o comportamenti anomali.

6.3 Process Discovery

Per la scoperta del processo ho utilizzato le reti di Petri inductive con una soglia (threshold) di 0.9.

Ho scelto questo approccio perché:

- Garantisce un modello completo e deterministico, privo di blocchi o comportamenti non validi;
- Consente un buon bilanciamento tra accuratezza e generalizzazione, grazie alla soglia di 0.9, che limita l'inclusione di rumore o casi troppo rari;
- Produce modelli più leggibili e interpretabili rispetto ad altre tecniche.

Non ho utilizzato le altre due principali tecniche di discovery, come Alpha Miner e Heuristics Miner, perché:

- **Alpha Miner** tende a produrre modelli incompleti o non connessi, soprattutto in presenza di rumore o dati incompleti, come nel mio dataset;
- **Heuristics Miner** genera modelli più complessi e meno deterministici, con molte transizioni spurie e cicli poco chiari, rendendo difficile l'interpretazione e l'analisi.

Inizialmente, ho applicato il metodo inductive al dataset completo (df_W_complete), ottenendo una rete di Petri molto lunga, caratterizzata da numerosi cicli.

Successivamente, ho analizzato separatamente i casi **short**, che corrispondono ai dataframe più piccoli (percentile 25%), e poi i casi **middle** e **long**.

Ho notato che il processo rappresentato dal caso **short** fornisce un'idea chiara di come dovrebbe essere il processo principale, anche se presenta diversi cicli, che sono stati approfonditi in seguito.

Analizzando i casi **middle** e **long**, ho osservato che, paradossalmente, il caso **long** è molto più simile al caso **short**, mentre il caso **middle** presenta molte più attività ed è complessivamente più lungo.

Tutte e tre le versioni presentano cicli soprattutto nelle stesse aree, che corrispondono alle cosiddette transizioni **W_**.

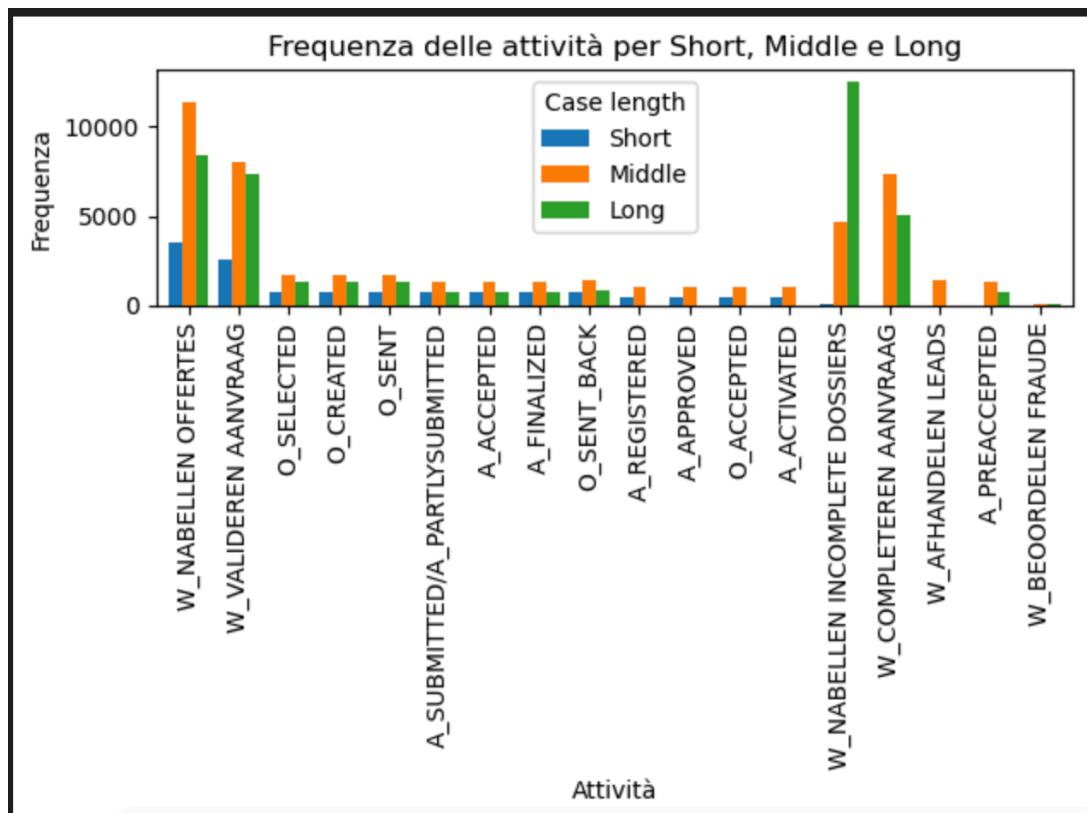
Un'altra osservazione interessante è che nei casi corti (short) non è presente la transizione **W_boolean Fraud**.

6.4 Descriptive analysis

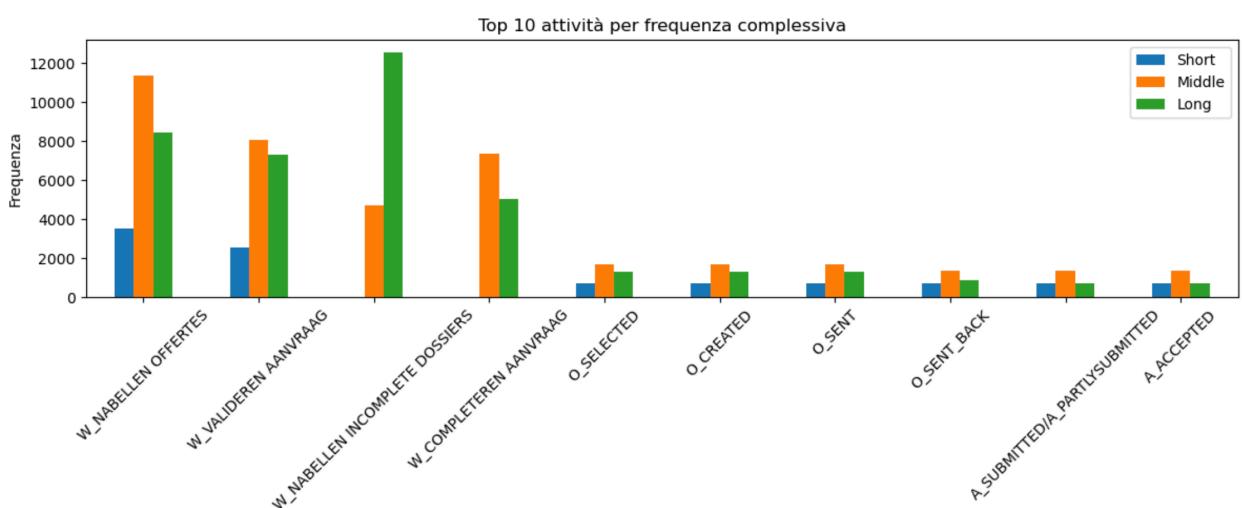
Descriptive analysis uses statistical tools to understand the distribution and the characteristics of the data.

Analizzando la rete di Petri, sono emersi numerosi cicli, e ho voluto capire quali attività fossero all’origine di questi. Per questo motivo, ho esaminato la frequenza delle attività in ciascuno dei tre dataframe a disposizione (short, middle e long).

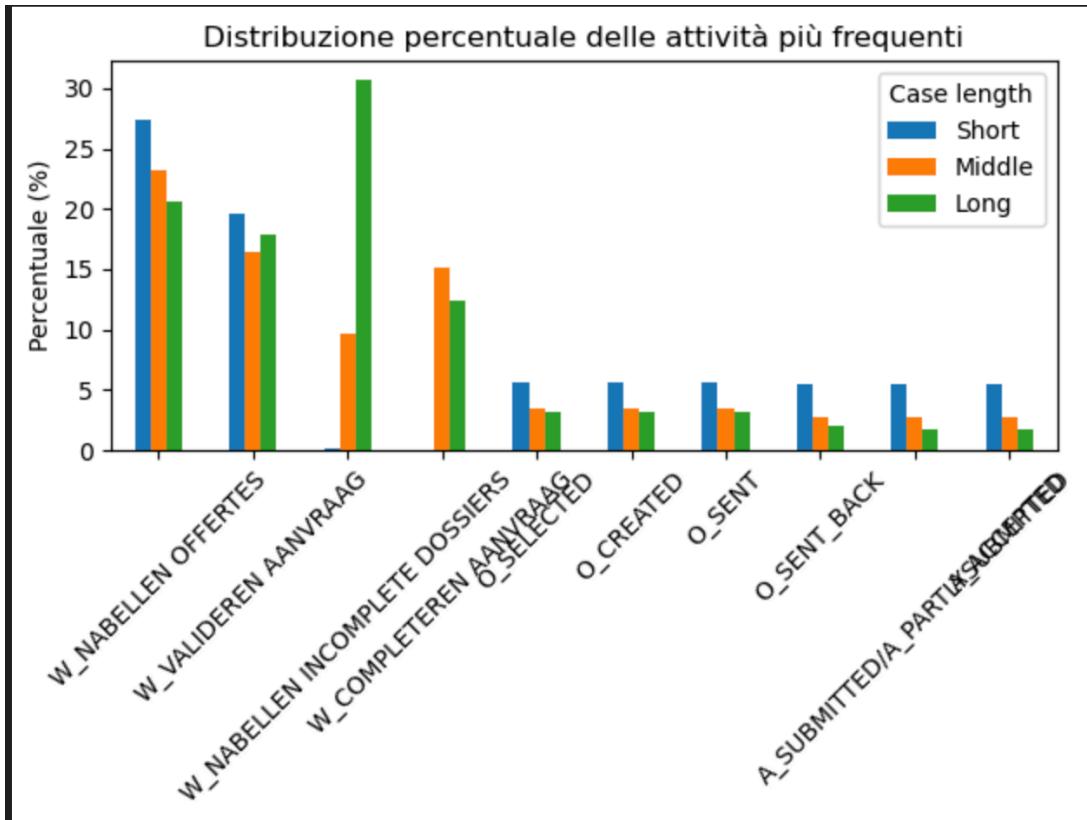
Un risultato interessante è che tutti e tre i dataframe presentano lo stesso problema: un elevato numero di attività ripetute, corrispondenti proprio ai cicli osservati nelle reti di Petri.



Successivamente, ho individuato le 10 attività più frequenti per ciascun dataframe e ne ho calcolato la distribuzione percentuale.



L'analisi della distribuzione percentuale delle attività è fondamentale per ottenere una valutazione normalizzata. Poiché i dataframe contengono numeri diversi di eventi, utilizzare le frequenze assolute non consentirebbe un confronto diretto e corretto tra i diversi casi. La percentuale, invece, mette in relazione la frequenza di ciascuna attività con il totale degli eventi nel dataframe, fornendo una misura relativa. Questo permette di confrontare con facilità l'importanza e l'incidenza delle attività nei processi di dimensioni differenti, oltre a facilitare l'identificazione di pattern comuni o differenze significative.



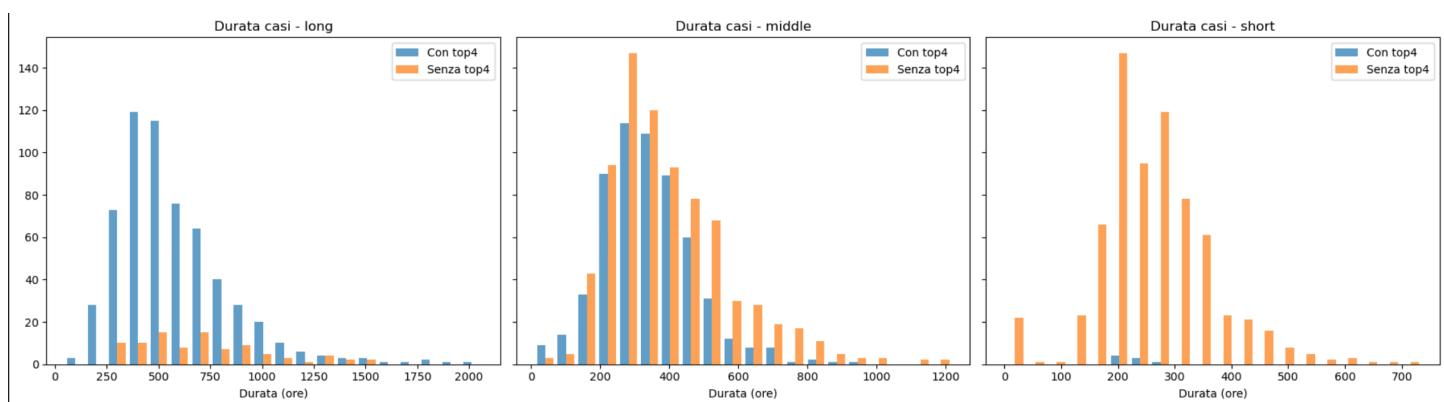
Tra tutte, le attività con frequenza più elevata sono quattro, tutte con prefisso **W_**. Ho quindi voluto verificare come la presenza simultanea di queste quattro attività incidesse sulla durata complessiva dei casi, confrontandoli con casi in cui una o nessuna di queste attività fosse presente.

L'analisi mostra che, nei casi medi e corti, non è la quantità di queste attività a far aumentare la durata complessiva del processo.

Per approfondire, ho calcolato la durata media per ciascuna attività. Sebbene nel codice mostrato non sia definita esplicitamente una funzione `durata_media_attività`, in generale questo tipo di calcolo si effettua così:

- Per ogni caso e per ogni attività si identificano i timestamp di inizio e fine.
- Si calcola la differenza tra questi timestamp, ottenendo la durata di ciascuna esecuzione.
- Si raggruppano tutte le durate per ogni attività.
- Si calcola la media delle durate, ottenendo la durata media di ogni attività.

In ogni caso, il tempo risulta più elevato per le quattro attività più frequenti.



Tuttavia, il calcolo iniziale non era del tutto accurato. Perciò, ho ridefinito i tempi nel modo seguente:

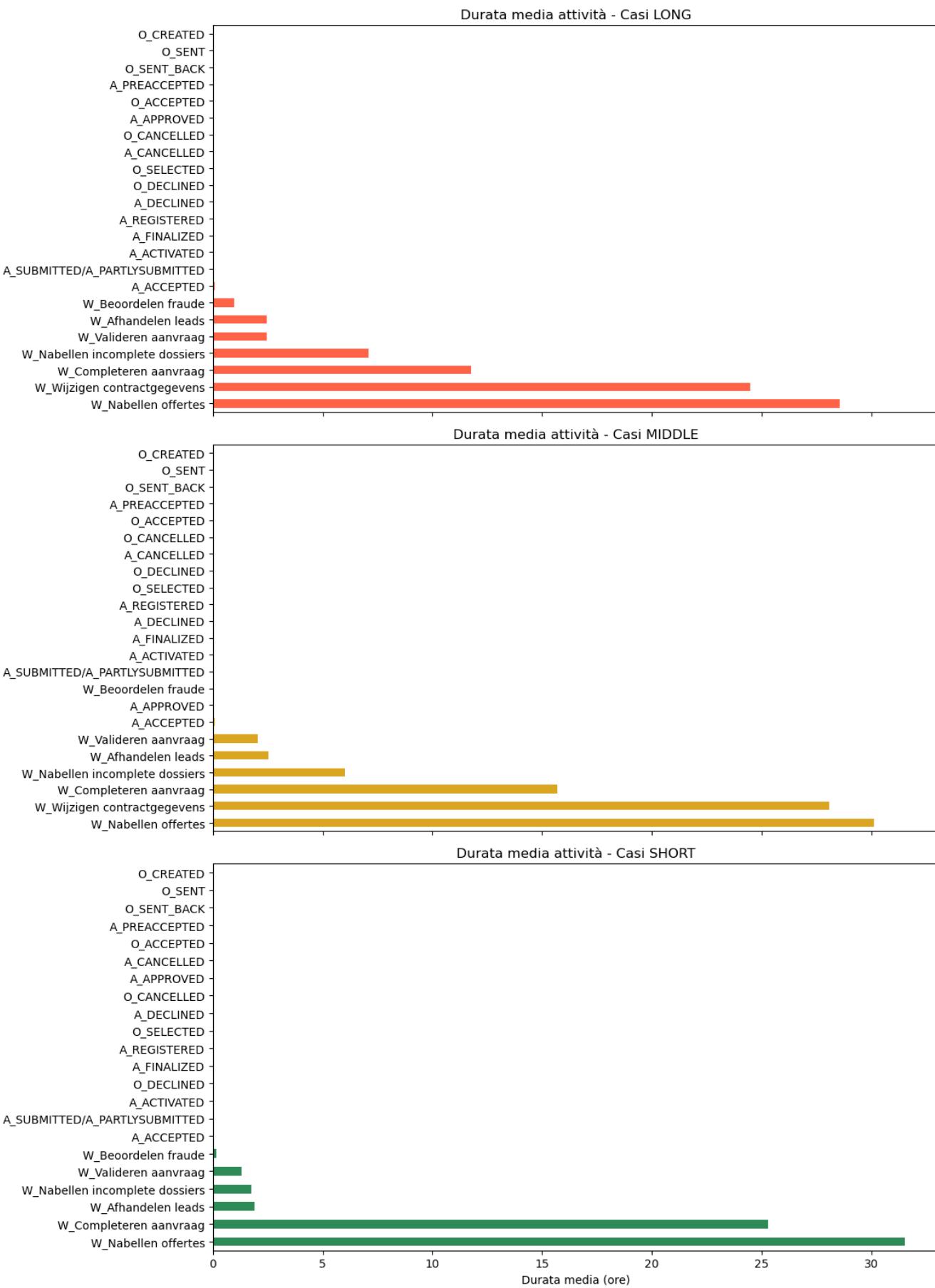
- **Execution time**: differenza tra il timestamp `complete` e `start` della stessa attività nello stesso caso, rappresenta il tempo effettivo impiegato per eseguire l'attività.
- **Waiting time**: tempo trascorso tra la fine dell'attività precedente e l'inizio di quella corrente, calcolato come differenza tra il timestamp `start` dell'attività corrente e il `complete` dell'attività precedente.
- **Total time**: somma di waiting time ed execution time, oppure differenza tra il timestamp `complete` e un evento di riferimento (ad esempio, l'inizio del processo o dell'attività precedente).

Questi calcoli sono stati applicati a ciascun dataframe (`long`, `middle` e `short`), ottenendo risultati più precisi che confermano chiaramente i colli di bottiglia rappresentati dalle seguenti attività:

- **W_Completeren aanvraag**: completamento delle informazioni
- **W_Nabellen offertes**: follow-up dopo l'invio delle offerte
- **W_Valideren aanvraag**: validazione finale della richiesta
- **W_Nabellen incomplete dossiers**: recupero di informazioni mancanti
- **W_Beoordelen fraude**: controllo frodi

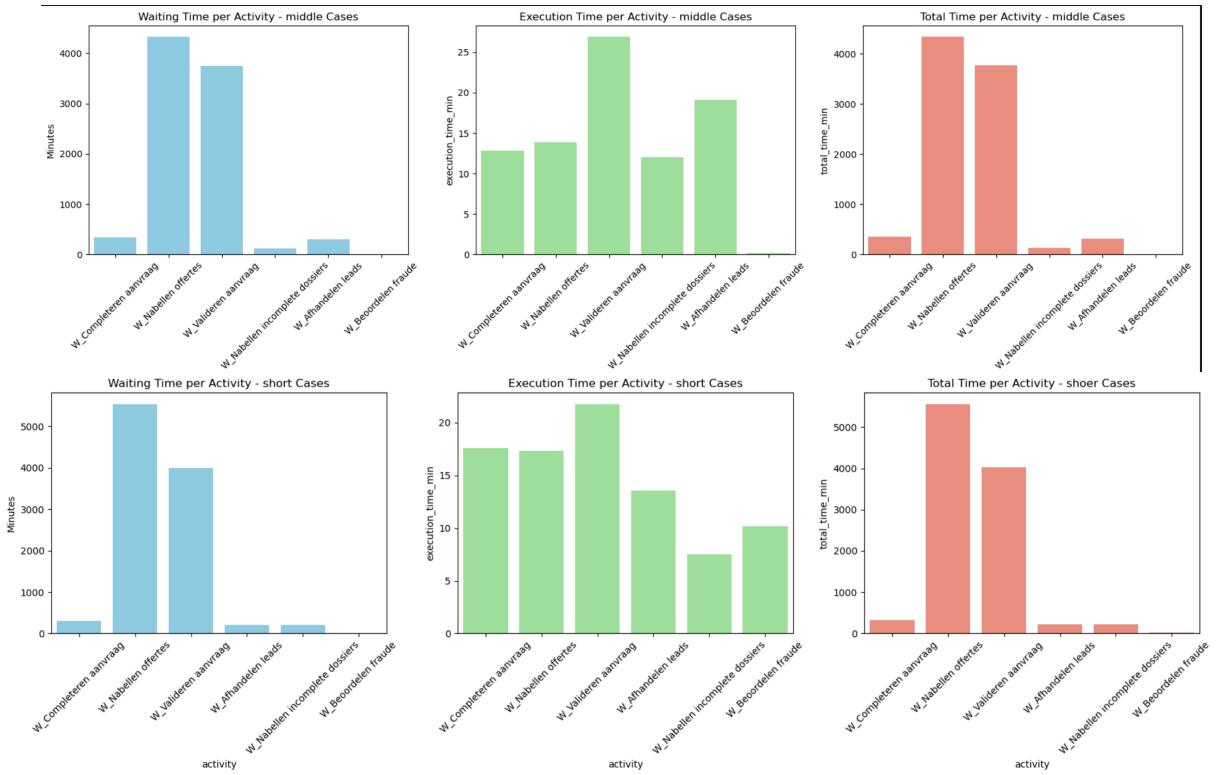
L'attività **W_Beoordelen fraude** risulta praticamente assente o molto rara, indicando che non si sono verificate richieste di modifiche finanziarie o variazioni contrattuali. Questi passaggi sono quindi legati principalmente all'ottenimento del primo contratto.

concept:name

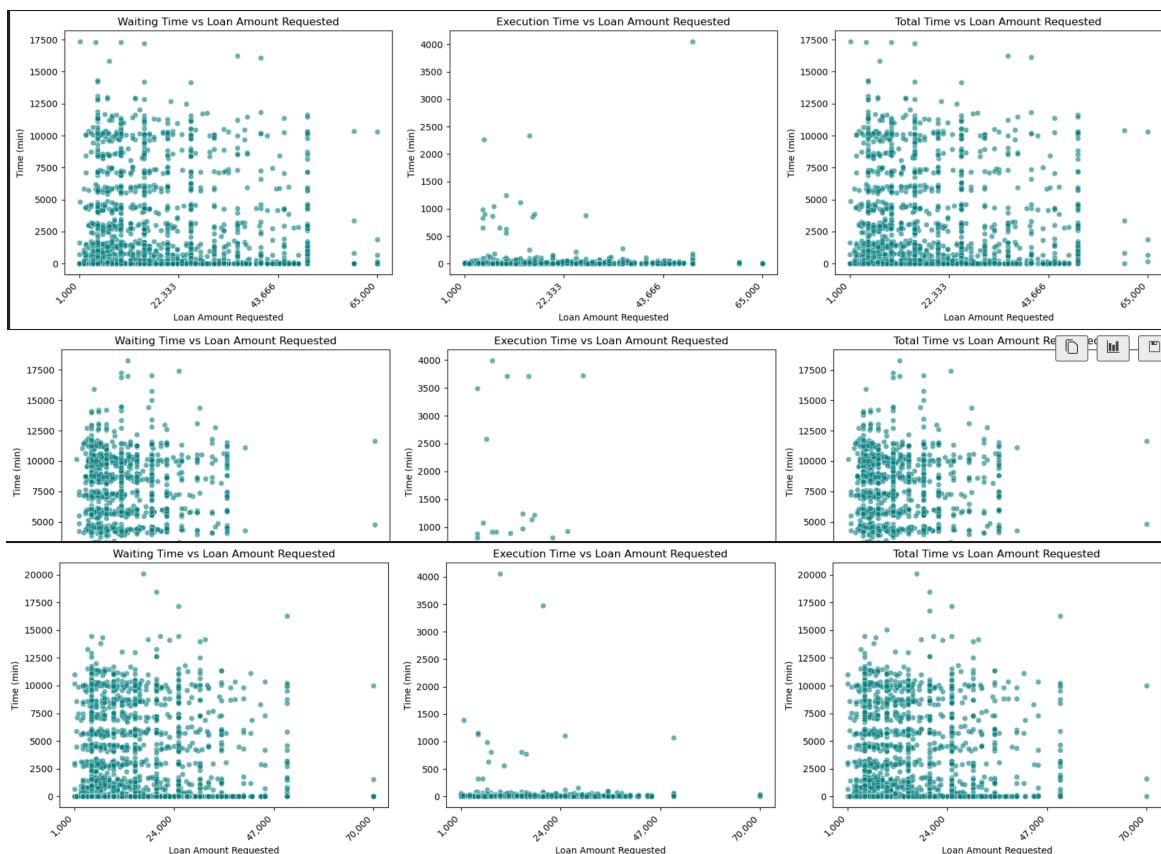


Questo è confermato anche dall'assenza dell'attività:

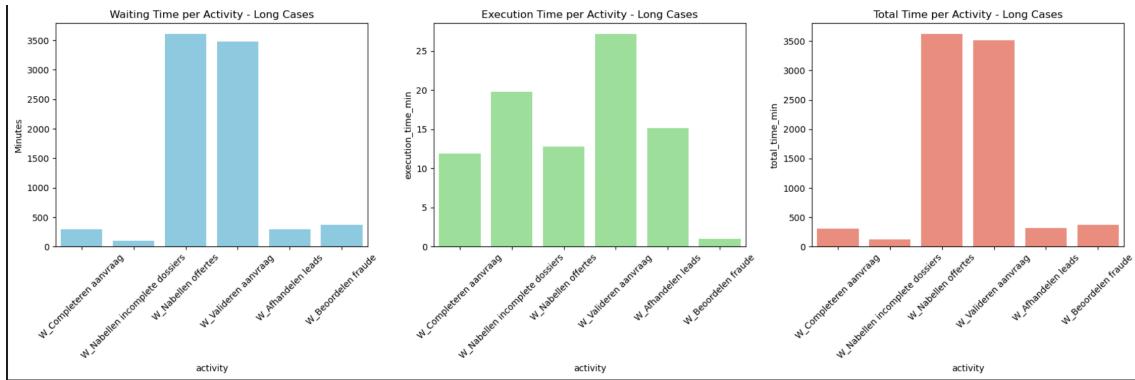
- **W_Wijzigen contractgegevens**: modifica dei dati del contratto, che non compare mai nei dati. Possiamo quindi supporre con buona certezza che non ci siano state modifiche contrattuali successive.



Un altro aspetto che volevo analizzare era se la durata del processo dipendesse dall'importo del prestito richiesto (loan amount). In altre parole, volevo capire se cifre più alte comportassero tempi più lunghi per la gestione della pratica Ho quindi suddiviso l'analisi per ciascun dataframe (short, middle e long) e per ciascuna delle metriche temporali considerate: **waiting time**, **execution time** e **total time**.



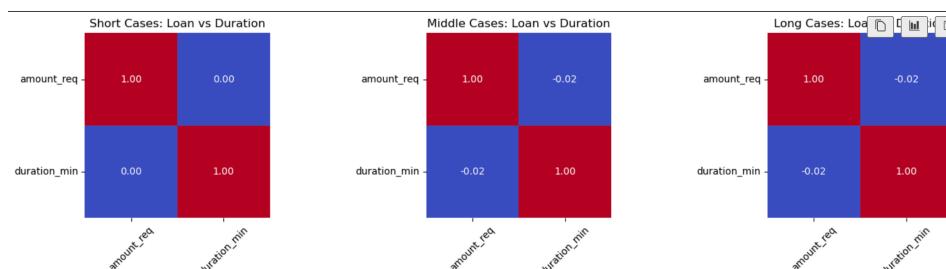
I risultati mostrano chiaramente che non esiste una correlazione significativa tra l'importo richiesto e la durata del processo. In pratica, anche per prestiti di importi elevati si possono osservare tempi brevi, e viceversa, indicando che la dimensione finanziaria del prestito non è un fattore determinante per la durata complessiva.



Nel dettaglio:

- **Rosso** indica una correlazione positiva forte (all'aumentare di un valore, cresce anche l'altro).
- **Blu** indica una correlazione negativa (all'aumentare di un valore, l'altro diminuisce).
- **Bianco** indica assenza di correlazione.

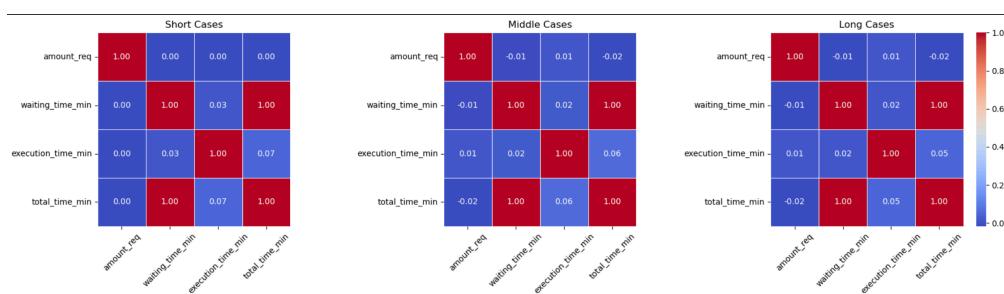
L'analisi conferma quindi un'assenza di correlazione (area bianca) tra **amount** e i tempi misurati.



Tuttavia, dall'analisi emerge una correlazione positiva significativa tra il **waiting time** e le attività con prefisso **W_**(quelle che abbiamo già identificato come possibili colli di bottiglia). Questo conferma che l'attesa tra le attività è un fattore rilevante nel determinare la durata complessiva del processo, specialmente per quelle attività di controllo, validazione e follow-up.

In sintesi, mentre l'importo del prestito non incide sui tempi, il tempo di attesa tra le attività legate ai controlli e alle verifiche (le attività **W_**) gioca un ruolo cruciale nell'allungare la durata totale dei casi.

In



conclusione, risulta fondamentale approfondire quali siano i problemi specifici legati alle attività con prefisso **W_**, dato che queste influenzano significativamente sia i tempi di attesa sia la durata complessiva dei casi. Comprendere le cause che portano a questi allungamenti — come eventuali inefficienze, mancanza di risorse, passaggi ridondanti o necessità di ulteriori verifiche — è essenziale per poter intervenire in modo mirato.

Un’analisi più dettagliata potrebbe aiutare a identificare i veri colli di bottiglia e a sviluppare strategie per snellire il processo, riducendo così sia la lunghezza che la complessità dei casi, migliorando l’efficienza complessiva del flusso di lavoro.

6.6 Conformance checking

identifies any deviations from the typical sequence of events

Calcolo delle metriche di conformità

Abbiamo utilizzato il metodo di *token-based replay* della libreria PM4Py per analizzare tre gruppi di casi rispetto al modello di processo. Le metriche calcolate sono:

- **Fitness media:** misura quanto le tracce si adattano al modello (da 0 a 1).
- **Token mancanti:** elementi previsti dal modello ma assenti nelle tracce, indicano deviazioni.
- **Token residui:** elementi presenti nelle tracce ma non consumati dal modello, anch'essi segnali di deviazione.
- Numero totale di tracce analizzate per ciascun gruppo.

Visualizzazione e analisi dei risultati

Per facilitare il confronto, abbiamo creato un grafico a barre con due assi:

- Asse sinistro (blu): fitness media per gruppo.
- Asse destro (rosso/arancione): token mancanti e token residui per gruppo.

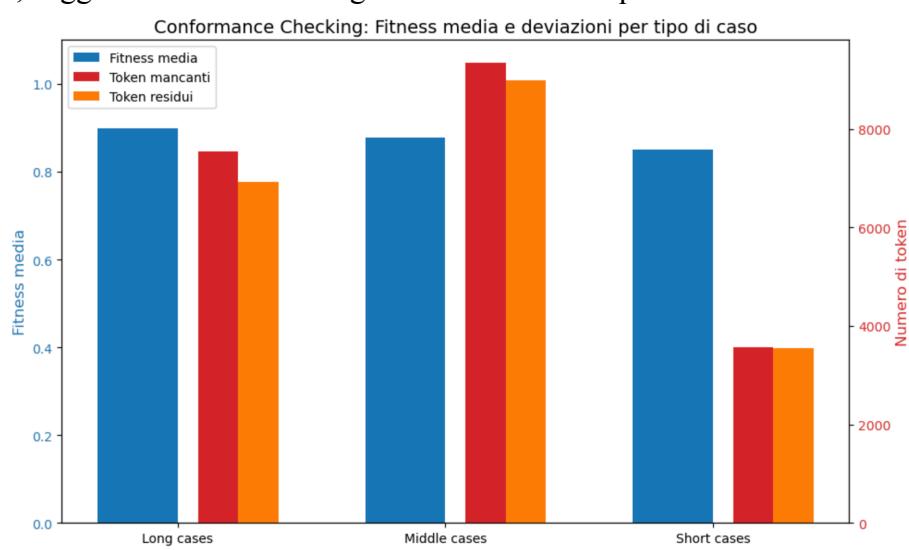
Questo permette di osservare come la conformità vari tra i gruppi, mettendo in relazione l'aderenza al modello con le deviazioni.

Importanza dell'analisi:

La conformance checking individua le tracce che si discostano dal modello, aiutando a capire quali gruppi mostrano maggiori problemi e a identificare potenziali cause di inefficienza.

Risultati chiave:

La fitness media diminuisce da long a short cases, indicando che i casi più brevi si discostano di più dal modello. I token mancanti e residui sono più elevati nei middle cases, suggerendo deviazioni significative anche in quei casi.



6.2.2 FILTERING IRRELEVANT DATA DF_AO_DC

Qui ho suddiviso il dataset in base alle attività finali relative ai casi rifiutati, che includono gli eventi **A_DECLINED**, **A_CANCELLED** e **O_CANCELLED** (quest'ultimo, sebbene non previsto tra gli esiti principali, è rilevante per analisi future).

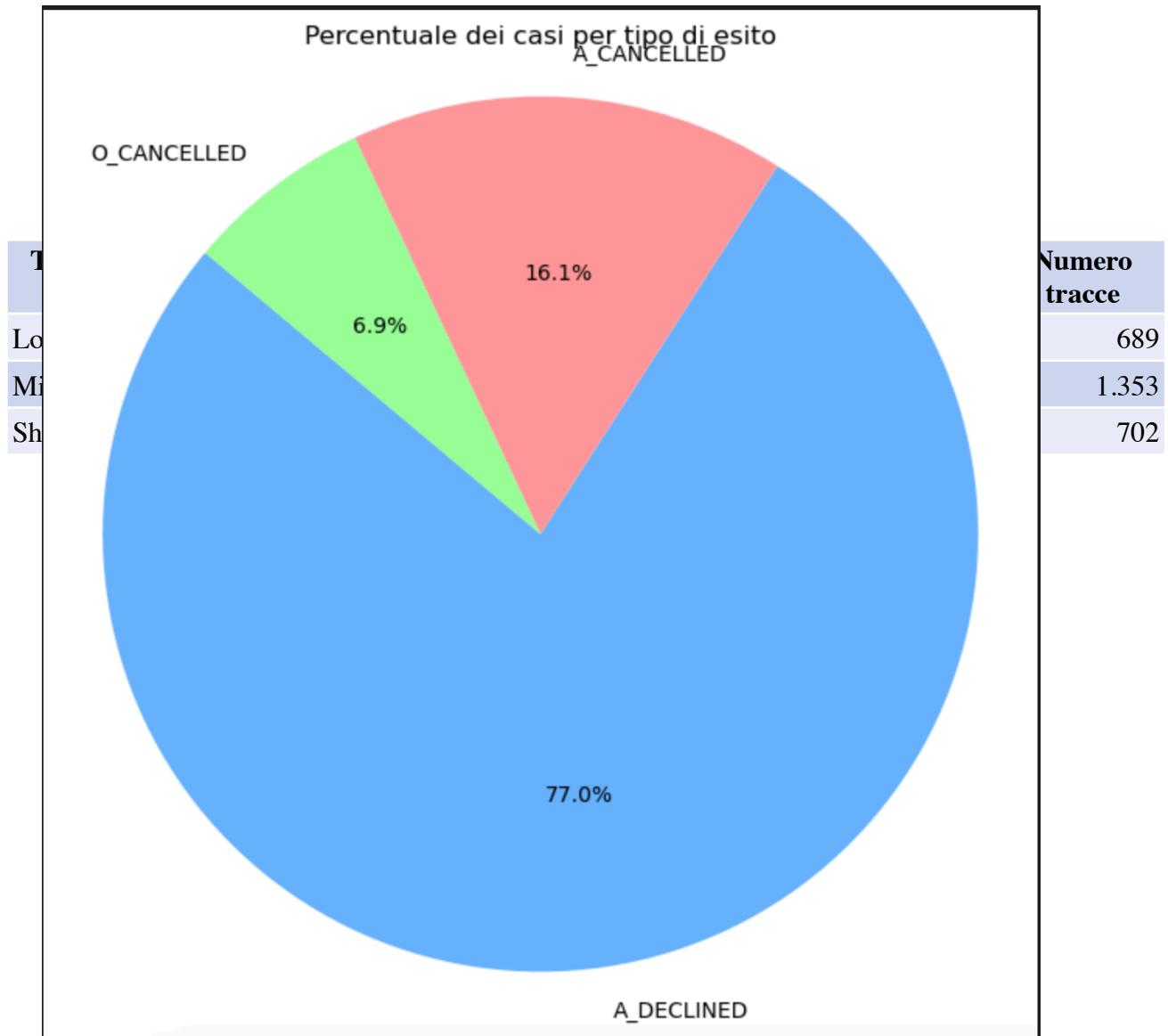
L'evento **A_APPROVED** non compare mai come evento finale. Per questo motivo, non

ha senso forzare una chiusura artificiale del caso basandosi su questo evento, per evitare di considerare come chiuse attività ancora aperte o incomplete.

Come anticipato, ho osservato che l'attività **A_DECLINED** è quella che si presenta più frequentemente come evento finale.

Quindi, ho ulteriormente diviso il dataframe in tre parti utilizzando i filtri sulle attività finali:

df_declined
df_cancelled
df_o_cancelled



6.3 Process Discovery

Per la scoperta del processo ho utilizzato le reti di Petri inductive con una soglia (threshold) di 0.9. Ho scelto questo approccio perché:

- Garantisce un modello completo e deterministico, privo di blocchi o comportamenti non validi;

- Consente un buon bilanciamento tra accuratezza e generalizzazione, grazie alla soglia di 0.9, che limita l'inclusione di rumore o casi troppo rari;
- Produce modelli più leggibili e interpretabili rispetto ad altre tecniche.

Non ho utilizzato le altre due principali tecniche di discovery, Alpha Miner e Heuristics Miner, perché:

- Alpha Miner tende a produrre modelli incompleti o non connessi, soprattutto in presenza di rumore o dati incompleti, come nel mio dataset;
- Heuristics Miner genera modelli più complessi e meno deterministici, con molte transizioni spurie e cicli poco chiari, rendendo difficile l'interpretazione e l'analisi.

Ho applicato la rete di Petri inductive su tre sottoinsiemi del dataset. Per l'attività **A_DECLINED** ho ottenuto un modello molto semplice, con solo due attività: quella iniziale comune a tutti (**A_SUBMITTED**) e quella finale (**A_DECLINED**).

Non ho utilizzato soglie per il filtraggio del rumore in questo caso, e possiamo ipotizzare che questo modello rappresenti casi di errori di applicazione, in cui un'attività viene creata e poi eliminata subito dopo. Questi eventi possono quindi essere considerati rumore inutile e rimossi dal dataset. Infatti, successivamente non ho più utilizzato quel sottoinsieme.

Ho invece analizzato più nel dettaglio i casi con attività finali **O_CANCELLED** (annullamento o rifiuto dell'offerta).

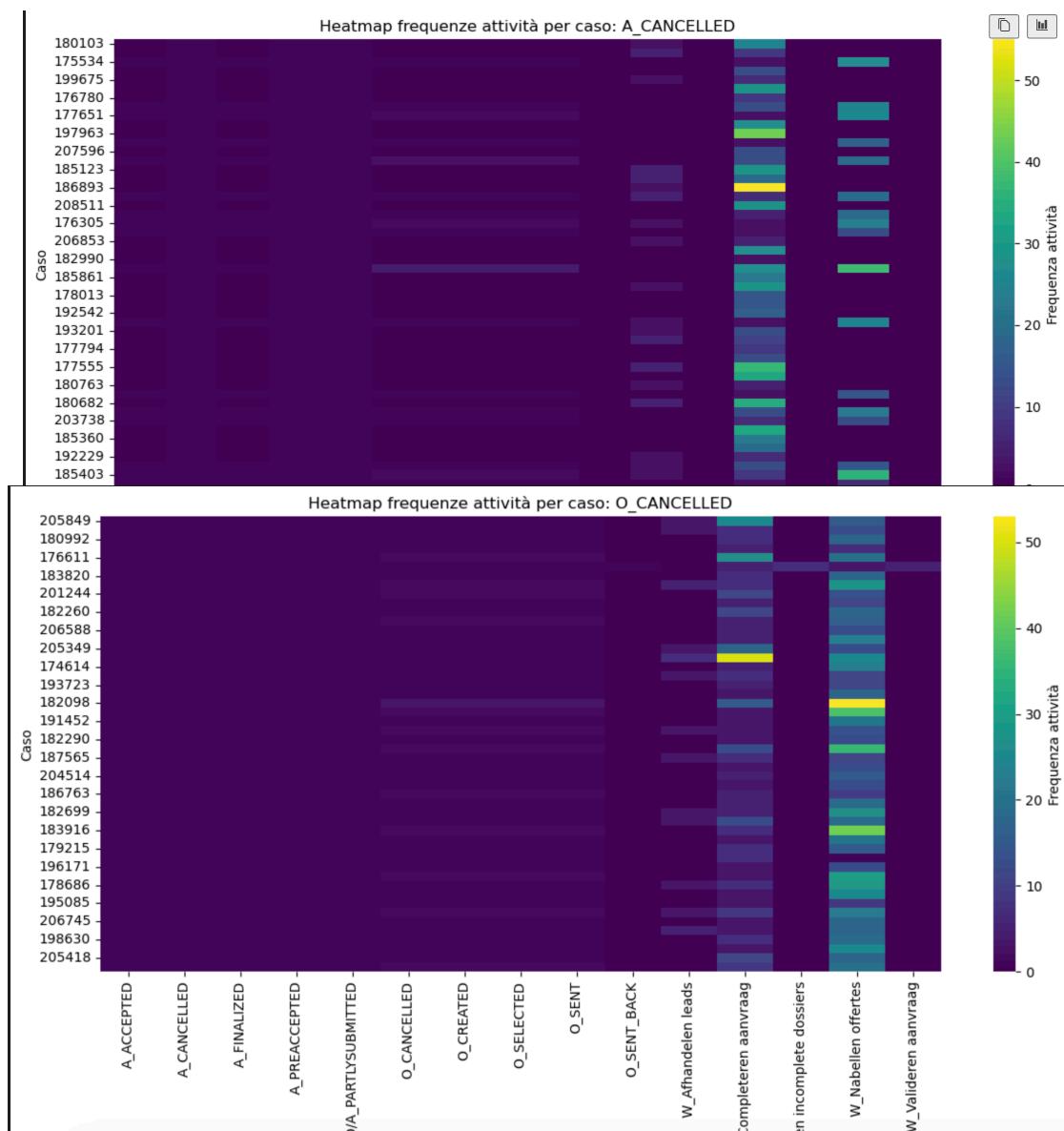
La seconda categoria, cioè i casi che terminano con **A_CANCELLED**, presenta un modello molto più articolato, con diverse attività di offerta e attività di attesa (**W_**) che possono aiutare a capire perché la richiesta sia sempre stata rifiutata dall'applicazione o annullata (come nel caso di **O_DECLINED** e **A_DECLINED**).

6.4 Descriptive analysis

Per cercare di individuare le attività che possono indicare un'elevata probabilità di rifiuto, ho realizzato una heatmap in cui si analizza la frequenza media delle attività nei casi che si concludono con un esito negativo.

Risultati principali:

- Nei casi che terminano con O_CANCELLED, l'attività con valore più alto è W_Nabellen offertes, ovvero il follow-up dopo l'invio delle offerte.
Il termine "follow-up" indica un ricontatto o un sollecito nei confronti del cliente, ad esempio per ricordare o sollecitare una decisione dopo l'invio dell'offerta.
Questo può suggerire che, quando è necessario effettuare follow-up multipli, ci sia una maggiore probabilità che il cliente abbandoni il processo o annulli l'offerta.
- Nei casi che terminano con A_DECLINED, l'attività più ricorrente è W_Completeren aanvraag, ovvero il completamento delle informazioni necessarie per la richiesta.
Questo può indicare che, nei casi in cui sono richiesti più interventi per completare l'applicazione, ci sia maggiore probabilità che la richiesta venga infine respinta.



I due boxplot confrontano il numero di attività eseguite prima della cancellazione nei casi di tipo A_CANCELLED e O_CANCELLED.

A_CANCELLED

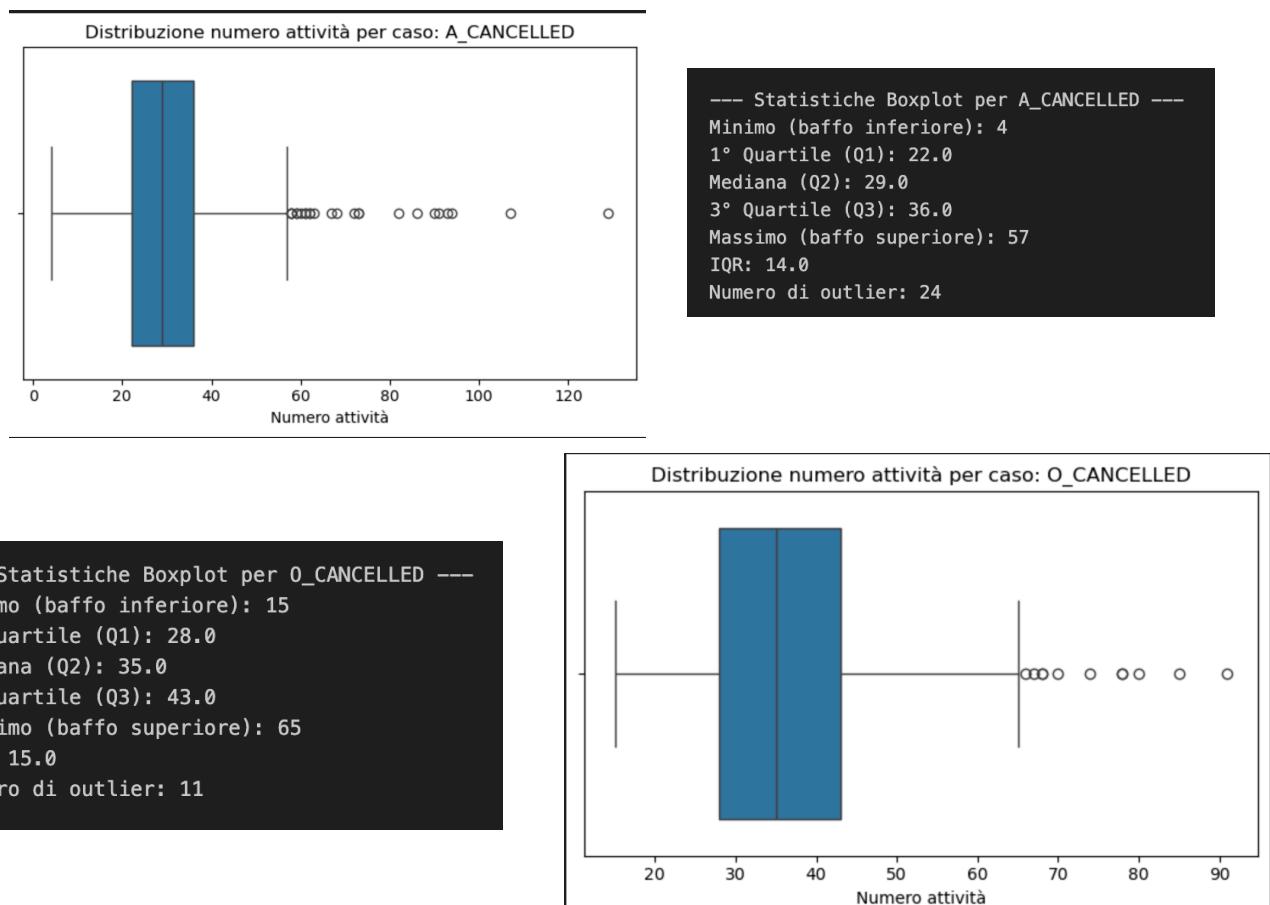
- Mediana: 35 attività → la metà dei casi viene cancellata relativamente presto.
- Intervallo interquartile (IQR): ~25–35 attività → la maggior parte dei casi ha un numero limitato di attività.
- Outlier: Numerosi casi con oltre 60 attività, alcuni superano le 120 → indicano eccezioni molto lunghe.
- Minimo: Può avvenire la cancellazione anche dopo pochissime attività → alcuni casi si interrompono quasi subito.

O_CANCELLED

- Mediana: ~42 attività → cancellazione generalmente più tardiva.
- IQR: ~35–45 attività → processi leggermente più lunghi rispetto ad A_CANCELLED.
- Outlier: Presenti anche qui, ma si fermano attorno a 90 attività.
- Minimo: Attività minime più alte rispetto ad A_CANCELLED → casi meno "precoci".

Confronto

- I casi O_CANCELLED durano mediamente di più rispetto agli A_CANCELLED.
- Entrambi presentano outlier, ma A_CANCELLED mostra una maggiore variabilità con code più estreme.
- Le cancellazioni A_CANCELLED possono avvenire sia molto presto che molto tardi; O_CANCELLED è più stabile ma comunque soggetto a eccezioni.

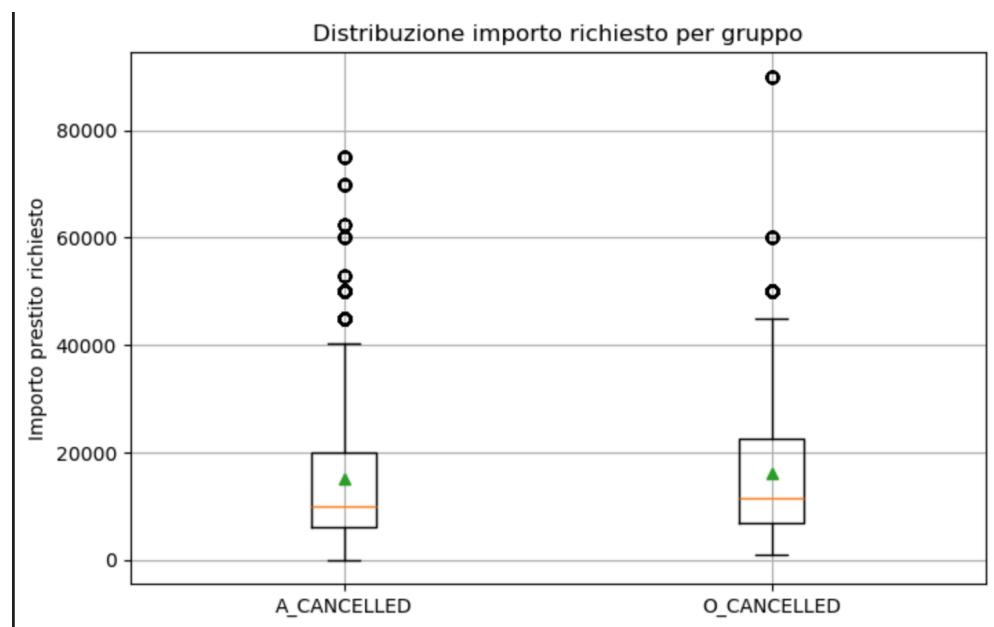


Analisi statistica e visualizzazione della distribuzione degli importi di prestito per i gruppi A_CANCELLED e O_CANCELLED

```
Statistiche per A_CANCELLED:  
min: 0  
max: 75000  
median: 10000.0  
mean: 15209.485712966809  
std: 12445.339828718335  
count_valid: 21663  
Statistiche per O_CANCELLED:  
min: 1000  
max: 90000  
median: 11500.0  
mean: 16233.375254267268  
std: 12920.99438570207  
count_valid: 11307
```

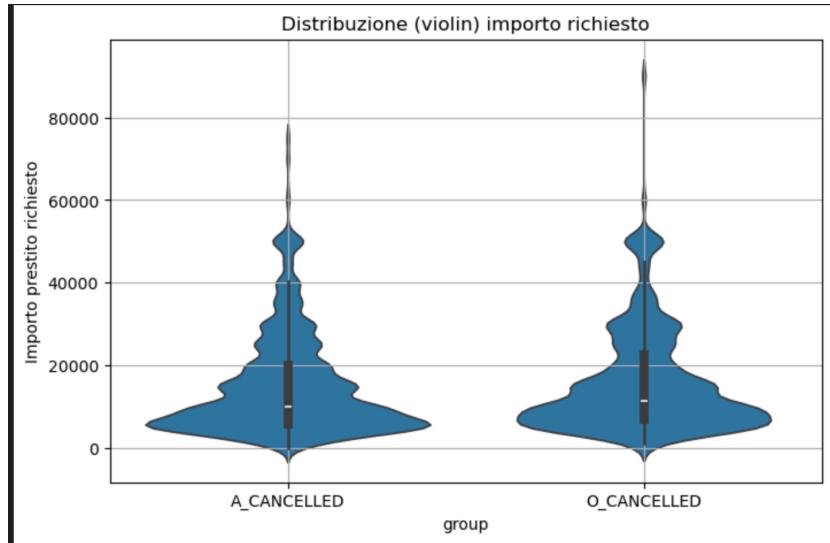
Sono state calcolate le principali statistiche descrittive degli importi di prestito richiesti nei due gruppi:

- Per il gruppo A_CANCELLED, l'importo minimo è 0, il massimo 75.000 euro, con una mediana di 10.000 euro e una media di circa 15.200 euro. La deviazione standard elevata (~12.445) indica una buona variabilità dei dati. Sono stati considerati validi 21.663 casi.
- Per il gruppo O_CANCELLED, l'importo minimo è 1.000 euro, il massimo 90.000 euro, con una mediana di 11.500 euro e una media leggermente superiore, circa 16.200 euro. La deviazione standard è simile a quella del primo gruppo (~12.920), con 11.307 casi validi.



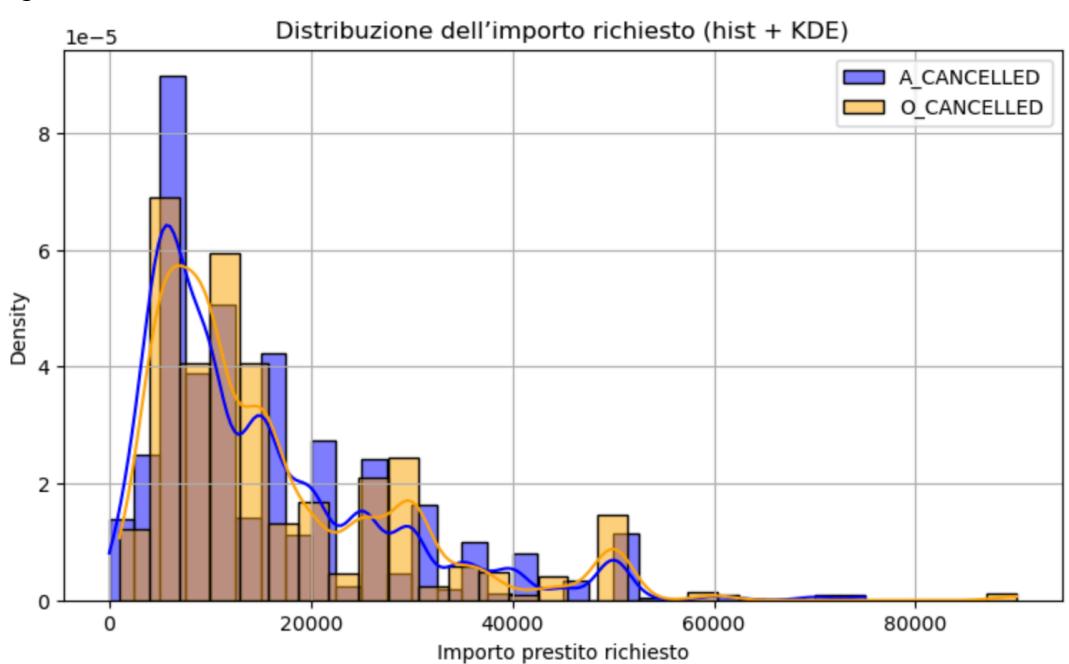
Le distribuzioni sono state esplorate tramite istogrammi accompagnati da stime di densità kernel (KDE), che mostrano la forma e la densità dei dati in modo continuo. Entrambi i gruppi presentano un picco nella fascia di importi compresa tra 5.000 e 15.000 euro, con una coda lunga verso valori più elevati, ma il gruppo O_CANCELLED tende a mostrare una maggiore concentrazione su importi più alti rispetto ad A_CANCELLED.

L'analisi con i violin plot conferma queste osservazioni, evidenziando la distribuzione e la densità dei dati in modo più intuitivo. La larghezza del "violino" in un dato punto indica la densità relativa di casi con quell'importo. Entrambi i gruppi mostrano una distribuzione simile e una variabilità consistente, con O_CANCELLED che presenta una leggera prevalenza di importi più alti e meno casi vicino allo zero. I quartili sono ben evidenziati e



indicano come la maggior parte dei dati si concentri attorno alle mediane.

In sintesi, pur mostrando caratteristiche simili, i prestiti nel gruppo O_CANCELLED tendono ad avere importi mediamente più elevati rispetto ad A_CANCELLED, con una distribuzione più spostata verso valori alti e una presenza minore di importi molto bassi. Entrambe le distribuzioni mostrano comunque una variabilità significativa con alcuni casi di importi estremamente alti.

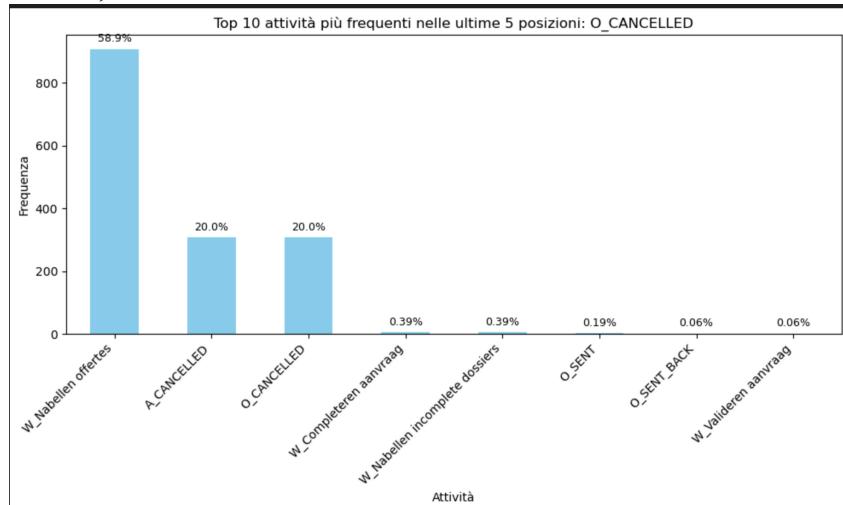


Analisi dei pattern nelle ultime attività prima della cancellazione

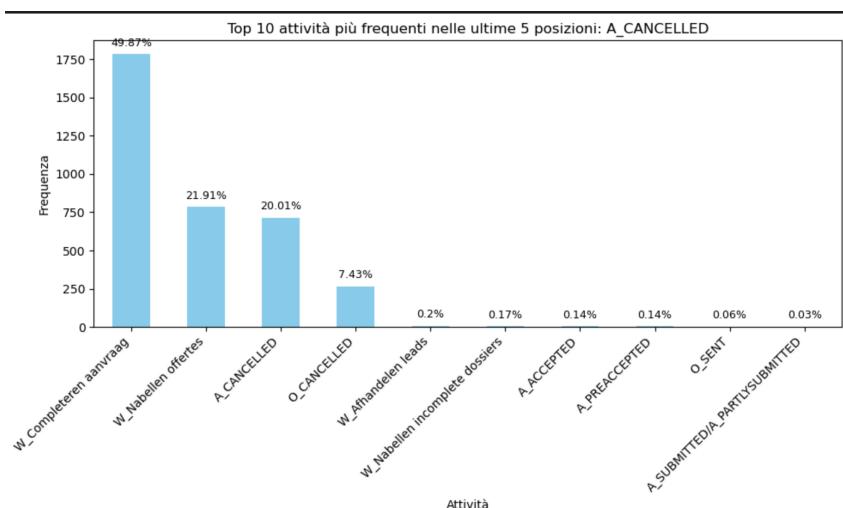
L'obiettivo di questa analisi era verificare se, osservando le ultime 5 attività di ciascun caso, fosse possibile individuare dei pattern ricorrenti che portano con alta probabilità alla conclusione del caso con un'attività di cancellazione, come A_CANCELLED o O_CANCELLED.

In particolare, si è cercato di capire se la presenza ripetuta di un'attività A_CANCELLED o O_CANCELLED nelle ultime posizioni potesse essere predittiva della cancellazione finale. Dai risultati ottenuti emerge un pattern chiaro: se una di queste attività compare tra le ultime, è molto probabile che il caso si concluda proprio con quella stessa attività di cancellazione.

Un risultato particolarmente interessante riguarda O_CANCELLED: quando tra le ultime attività compare W_Nabellen offertes: follow-up dopo l'invio delle offerte, la frequenza è molto alta. Questo suggerisce che si tratti di un pattern forte e ripetitivo. In effetti, per circa il 60% dei casi che si concludono con O_CANCELLED, quest'attività specifica appare tra le ultime, indicando una chiara tendenza.

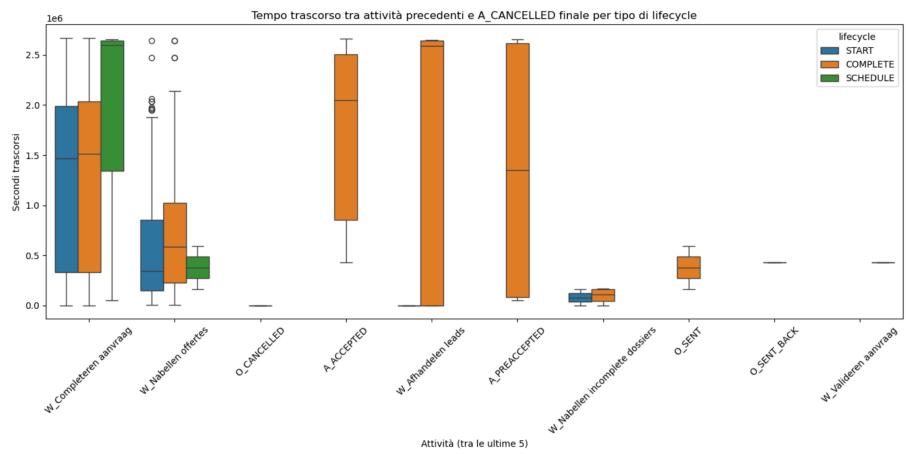


Nel caso di A_CANCELLED, invece, l'attività più frequente tra le ultime è W_Completeren aanvraag: completamento delle informazioni. Anche se meno intuitivo, questo dato è altrettanto rilevante: quasi il 50% dei casi che terminano con A_CANCELLED presenta questa attività tra le ultime, suggerendo che anche in questo scenario c'è una sequenza tipica che conduce alla cancellazione.



Analisi del tempo e del tipo di transizione (lifecycle)

Per approfondire ulteriormente, è stato analizzato anche il tempo trascorso tra le ultime attività e la cancellazione finale, distinguendo i tipi di transizione (schedule, start, complete).

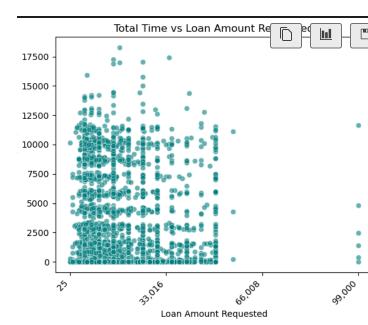
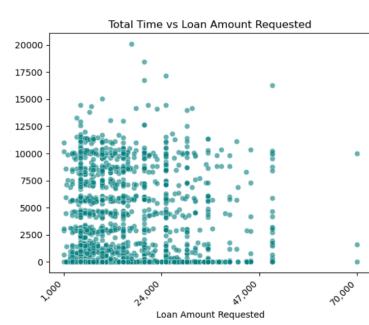
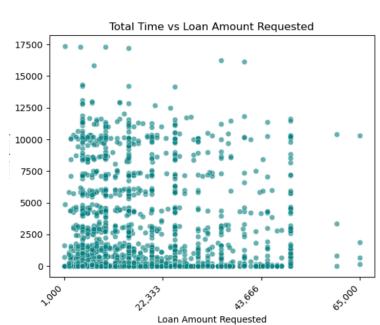
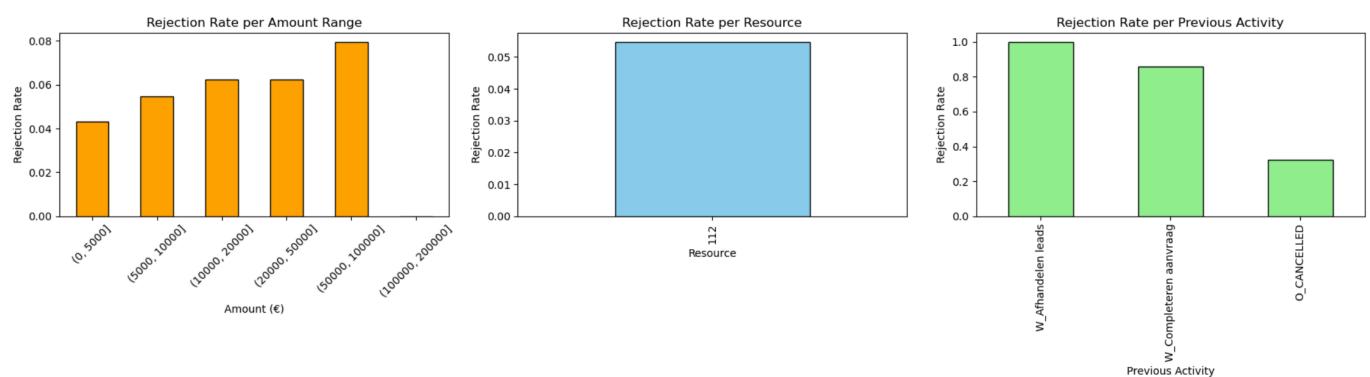


Infine, è stata condotta un'analisi dei tassi di rifiuto (rejection rate), dalla quale emerge che i casi appartenenti alla categoria **W_** presentano un alto tasso di rifiuto. Questo conferma quanto già osservato in precedenza: i casi **W_** non rappresentano un “vaso” (ossia non sono casi neutri o passivi), ma sono spesso oggetto di rifiuto.

In particolare, l'elemento con il più alto tasso di rifiuto è il **112**, che corrisponde a un reparto o a un'attività automatizzata.

Si nota inoltre che i prestiti di importo elevato vengono frequentemente rifiutati.

Al contrario, negli altri dataset la distribuzione dei rifiuti è più concentrata su valori centrali, compresi tra i **1000** e i **50.000 euro**, ovvero prestiti non troppo elevati



6.6 Conformance checking

In questo caso **non ho effettuato un'analisi di conformance checking**, perché ritengo che **non ci sia un reale beneficio ("win-win") nell'applicarla**.

I dati considerati rappresentano casi in cui l'**applicazione è stata rifiutata** (dal sistema o dall'utente), quindi non descrivono un flusso di processo "ideale" da confrontare con un modello di riferimento.

Il conformance checking, in generale, serve per **identificare deviazioni dalla sequenza tipica degli eventi**. Tuttavia, in questo contesto, l'intero processo può essere considerato **una deviazione in sé**, in quanto termina in modo indesiderato per entrambe le parti:

- **Per la banca:** si verifica una **perdita di un potenziale cliente** e, quindi, una mancata opportunità di profitto e fidelizzazione. Inoltre, si perde l'effetto positivo del **passaparola**, che rappresenta ancora oggi uno dei canali di marketing più forti, grazie alla fiducia tra persone che si conoscono.
- **Per il cliente:** la richiesta viene respinta e, nel caso desideri ancora il prestito, dovrà **ripetere l'intero iter da capo**, con ulteriore dispendio di tempo e impegno.

In sintesi, **questi casi non si prestano bene all'analisi di conformità**, perché non esiste un "processo atteso" in cui tali esiti vengano considerati successi o obiettivi da replicare.

6.7 Intervention strategies

Le strategie di intervento vengono sviluppate sulla base delle informazioni e delle conoscenze acquisite nei passaggi precedenti dell’analisi.

Queste strategie non sono statiche, ma vengono continuamente migliorate grazie a un processo di monitoraggio costante e alla misurazione dei risultati ottenuti.

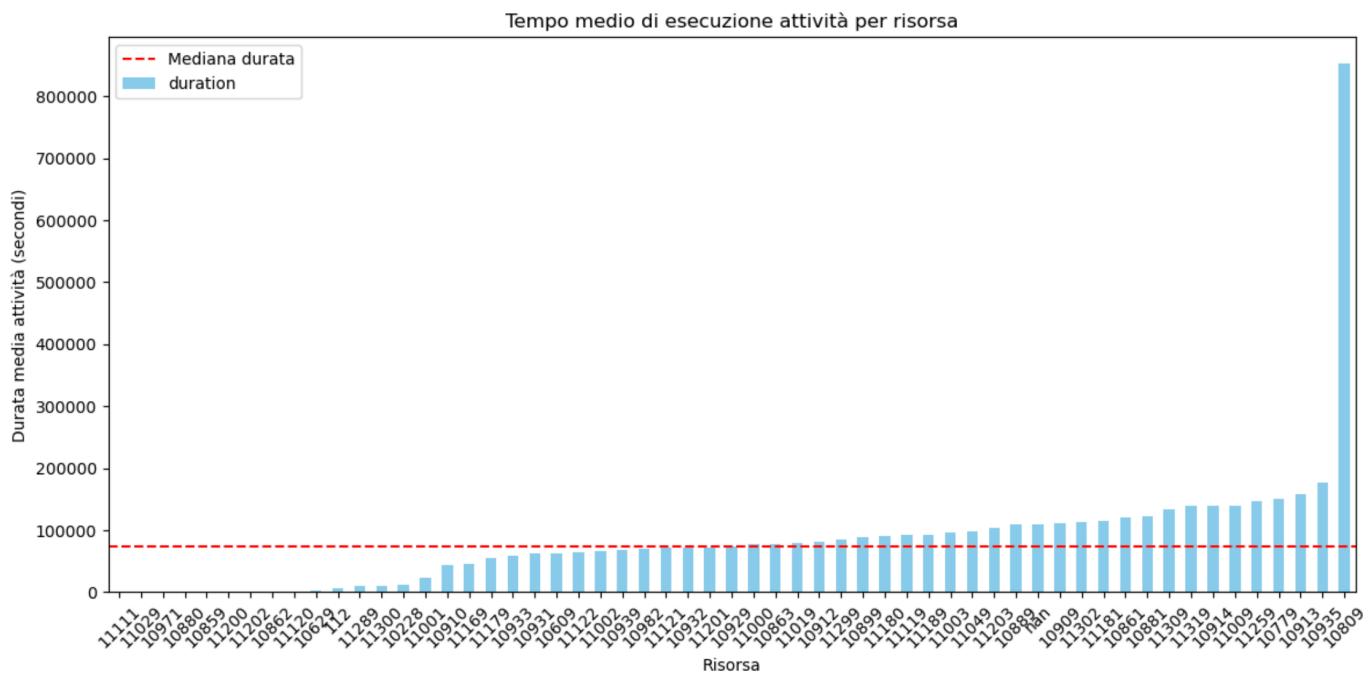
In pratica, si tratta di un ciclo iterativo in cui:

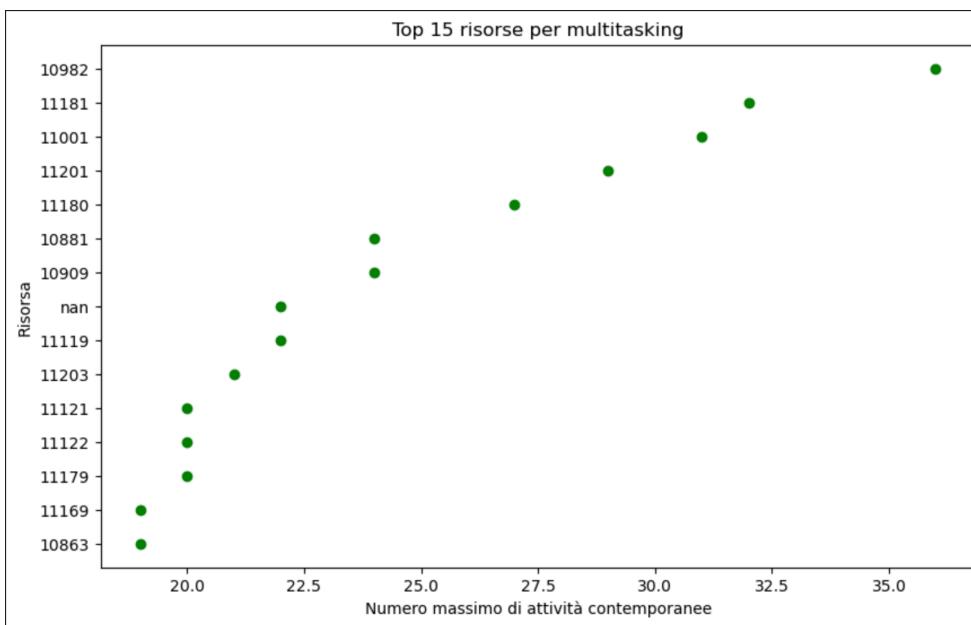
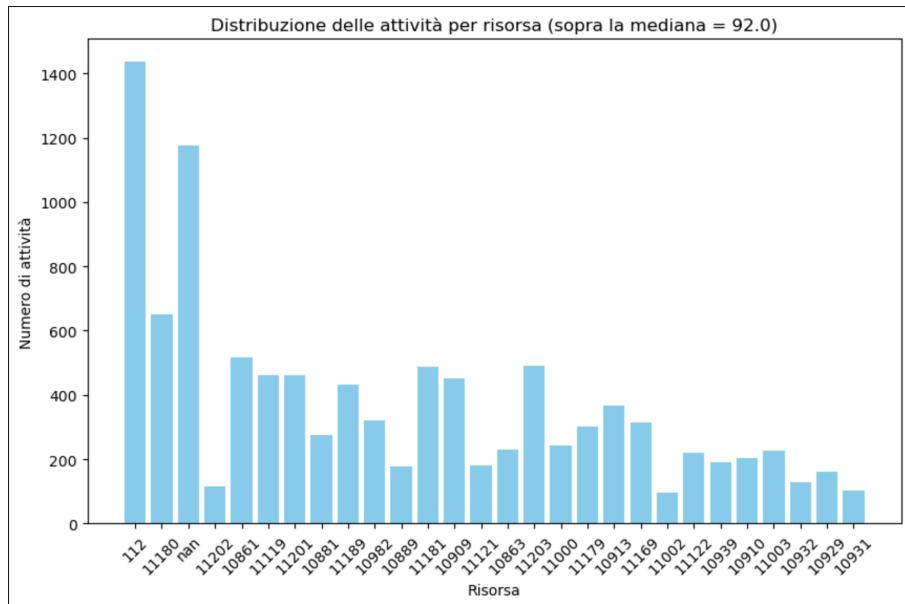
- I dati raccolti e le analisi effettuate guidano la definizione degli interventi da mettere in atto (ad esempio, modifiche strutturali al processo, riorganizzazione delle attività, formazione del personale).
 - Una volta implementate, le strategie vengono monitorate per valutarne l'efficacia nel tempo.
 - Sulla base dei risultati ottenuti e dei feedback ricevuti, vengono apportati ulteriori aggiustamenti per ottimizzare i processi e affrontare eventuali criticità residue.

Questo approccio garantisce che le azioni adottate siano sempre allineate agli obiettivi di miglioramento continuo, permettendo anche una risposta tempestiva a nuove sfide o cambiamenti nel contesto operativo.

Infine, per supportare l'elaborazione delle strategie ho calcolato:

- Il **tempo medio di esecuzione per risorsa**, per identificare possibili colli di bottiglia o sovraccarichi operativi;
 - La **distribuzione delle attività per risorsa**, focalizzandomi su quelle che superano la mediana, per evidenziare eventuali concentrazioni anomale di carico;
 - Le **top 15 risorse multitasking**, per valutare possibili effetti sul tempo di esecuzione e sulla qualità del lavoro svolto.





6.7.1 SVILUPPI FUTURI

Un possibile sviluppo futuro potrebbe riguardare **l'analisi predittiva**, con l'obiettivo di identificare i casi a rischio di rifiuto già nelle fasi iniziali del processo. Questo permetterebbe di attuare interventi mirati (come reminder personalizzati o assistenza dedicata) per aumentare le probabilità di successo dell'applicazione.

Inoltre, si potrebbe approfondire **l'analisi delle sequenze di attività** tramite tecniche di pattern recognition, per individuare configurazioni di comportamento che portano sistematicamente a esiti negativi, migliorando così il supporto decisionale.

<https://github.com/IkramBourras/Business-information-System>