
# UNIVERSITÀ POLITECNICA DELLE MARCHE
# FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

Corso di Laurea Magistrale o Specialistica in Data Science per l'Economia e le Imprese

Analisi della carriera universitaria tramite tecniche di Process Mining

Analysis of university career through Process Mining techniques

Relatore: Prof. Domenico Potena

Correlatrice: Prof.ssa Laura Genga

Tesi di Laurea di: Chiara Mercati

Anno Accademico 2022 – 2023

*"Senza dati sei solo un'altra persona con un'opinione"*

*W. Edwards Deming*

# INDICE

**1. INTRODUZIONE ..... 5**
1.1 CONTESTO E MOTIVAZIONI ..... 6
1.2 CASO DI STUDIO ..... 9
1.3 STRUTTURA DELLA TESI ..... 14
**2. FONDAMENTI TEORICI ..... 16**
2.1 PROCESS MINING ..... 18
2.1.2 Process Model ..... 29
2.1.2 Process Discovery ..... 34
2.1.3 Conformance Checking ..... 46
2.2 REGRSSIONE LINEARE ..... 48
2.3 REGRESSIONE LOGISTICA ..... 51
2.4 SUPPORT VECTOR MACHINES ..... 53
**3. METODOLOGIA DELLA RICERCA ..... 56**
3.1 OBIETTIVI DELLA RICERCA ..... 59
3.2 DATASET E PRE-PROCESSING ..... 60
3.3 APPROCCIO ..... 64
3.3.1 Estrazione dei Modelli di Processo ..... 65
3.3.1 Conformance Checking con i processi estratti ..... 71
3.3.2 Correlazioni tra le variabili ..... 73
3.3.3 Varianti con il minor tempo di laurea ..... 75
3.3.4 Predizione del tempo di laurea ..... 77
**4. RISULTATI ..... 83**
4.1 PROCESSI ANNI ACCADEMICI 2015-2017 ..... 83
4.2 PROCESSI ANNI ACCADEMICI 2017-2020 ..... 93
4.3 CORRELAZIONI TRA LE VARIABILI ..... 103
4.4 VARIANTI CON IL MINOR TEMPO DI LAUREA ..... 107

4.1 PREDIZIONE DEL TEMPO DI LAUREA .....114
4.2 LIMITAZIONI E SVILUPPI FUTURI .....137
**5. CONCLUSIONI.....140**
**6. BIBLIOGRAFIA E SITOGRAFIA .....143**
**7. RINGRAZIAMENTI.....146**

# 1. INTRODUZIONE

Gli studenti universitari devono affrontare sfide di rilievo nel corso della loro esperienza accademica e queste sfide possono avere un impatto sul tasso di completamento degli studi entro i termini prestabiliti. Comprendere e analizzare il percorso didattico svolto dagli studenti si rivela quindi fondamentale.

L'importanza di questo tipo di analisi sottolinea la necessità per le istituzioni universitarie di disporre di strumenti e metriche in grado di individuare le cause che portano all'insuccesso degli studenti.

Il Ministero dell'Istruzione Universitaria ha avanzato diversi indicatori con lo scopo di valutare il percorso accademico degli studenti. Tuttavia, questi indicatori offrono solamente una panoramica generale del comportamento degli studenti ed al fine di individuare gli ostacoli potenziali, risulta necessario condurre un'analisi più approfondita dei progressi degli studenti durante il loro percorso di studio, considerando fattori come la loro capacità di rispettare il piano di studio nel flusso di esami previsti ed entro i tempi previsti.

In questo studio, si presenta l'implementazione delle tecniche di estrazione dei processi educativi (Educational Process Mining - EPM) per affrontare queste sfide attraverso un caso di studio reale. L'obiettivo principale di EPM è individuare modelli e tendenze all'interno dei dati relativi all'ambito educativo, allo scopo di comprendere l'esecuzione dei processi formativi e identificare opportunità di

miglioramento. Lo studio rientra nella sotto disciplina dell'EPM nota come “curriculum mining”, che si focalizza sull'analisi dei dati relativi al percorso di studio degli studenti per ottenere importanti insight sulle scelte curriculari effettuate dagli studenti. In particolare, si applicano tecniche di EPM al curriculum, che rappresenta la sequenza ideale degli esami previsti dall'università per gli studenti, concentrandoci su due obiettivi principali: primo, identificare le differenze tra gli studenti che seguono il curriculum previsto e quelli che presentano ritardi, e secondo, determinare se e in che modo il progresso degli studenti durante il primo anno possa avere un impatto sulla tempistica della laurea.

Utilizzando i dati derivanti dai processi di apprendimento, è possibile acquisire informazioni preziose riguardo alle abitudini di studio degli studenti.

Lo studio mette in luce le complesse dinamiche dei percorsi accademici degli studenti, sottolineando l'importanza del completamento tempestivo degli esami e il rispetto delle linee guida del curriculum per migliorare i risultati educativi. Questi risultati forniscono approfondimenti per le istituzioni educative che cercano di ottimizzare le loro strutture di curriculum e meccanismi di supporto per migliorare i tassi di successo degli studenti.

## 1.1 CONTESTO E MOTIVAZIONI

I dati presi in considerazione si riferiscono a studenti che hanno frequentato il corso di laurea in Ingegneria Informatica e dell’Automazione presso una Università

Italiana. L'Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) ha definito alcuni criteri e metodologie per la valutazione dei programmi di studio. Sulla base di questi criteri il dataset è stato suddiviso in tre sezioni, ognuna delle quali rappresenta rispettivamente un sottogruppo di studenti in base al tempo di laurea, quindi, prendendo in considerazione quanto tempo è trascorso dal giorno dell'iscrizione (inizio anno accademico) al giorno del conseguimento del titolo finale:

1. *In tempo*

Gli studenti che hanno conseguito la laurea entro 3 anni e 6 mesi sono considerati laureati in tempo (in base all'indicatore iC02 che tiene conto della percentuale di laureati entro la normale durata del corso).

2. *Un anno in ritardo*

Questa sezione del dataset include studenti che hanno conseguito la laurea un anno dopo il periodo sopra menzionato (indicatore iC17).

3. *In ritardo*

Gli studenti classificati come "In ritardo" hanno impiegato più di 4 anni e 6 mesi per laurearsi (tutti gli studenti che non rientrano nei due indicatori già menzionati). Si utilizzano questi tre indicatori per estrarre i processi legati a questi sottogruppi di studenti e analizzare le differenze tra di loro, al fine di ottenere considerazioni relative al miglior percorso da seguire e raccomandarlo agli studenti che stanno iniziando i loro studi.

Tuttavia, questi indicatori non rappresentano l'intera popolazione studentesca, in quanto ci potrebbero essere anche studenti che abbandonano gli studi, ma si prende in considerazione la parte più rilevante per comprendere meglio le cause dei fallimenti e dei ritardi degli studenti. Questo studio si concentrerà sulle prestazioni degli studenti che si laureano in tempo rispetto a quelli che si laureano in ritardo, conducendo una valutazione interna del programma di studio del corso presso l'Università Italiana presa in considerazione.

Nell'analisi si considerano sia i corsi obbligatori che quelli a scelta del programma di studio, focalizzandosi esclusivamente sui primi due anni, poiché sono considerati i più critici per il completamento del programma di laurea. Ciò è dovuto alla struttura intrinseca del terzo anno, in quanto solo due corsi sono obbligatori e il resto del piano di studio prevede un totale di crediti a scelta dello studente; quindi, le prestazioni nel tempo dipendono dalle scelte di ciascuno studente e non dalla struttura del programma di studio. Concentrandosi solo sui primi due anni, l'analisi consente un confronto equo tra gli studenti che seguono lo stesso percorso di studio.

In sintesi, l'obiettivo principale di questo tipo di analisi è aiutare le università a individuare aree di miglioramento all'interno dei loro curricula, qualora ce ne fosse bisogno, e suggerire agli studenti il miglior percorso da seguire in modo tale che si possa incrementare il tasso di completamento degli studi entro i tempi previsti e la qualità complessiva dell'istruzione.

## 1.2 CASO DI STUDIO

Il dataset in questione è composto da 410 studenti iscritti agli anni accademici dal 2015-2016 al 2019-2020 del corso di laurea triennale in Ingegneria Informatica e dell'Automazione presso l'Università Politecnica delle Marche.

Il corso in questione ha un piano di studio che comprende diversi corsi obbligatori e opzionali, diffusi su tre anni accademici, ciascuno diviso in due semestri. Alcuni corsi hanno una sequenza logica, in questo caso gli esami avranno lo stesso nome ed un numero progressivo che identifica la continuazione del corso.

I corsi obbligatori sono presenti nei primi due anni del programma di studio, solo nel secondo anno lo studente ha la possibilità di scegliere 6 crediti, che corrispondono quindi ad un esame da sostenere il primo o il secondo semestre a seconda del corso. Mentre il terzo anno è composto maggiormente da corsi opzionali, insieme ad un progetto di ricerca ed uno stage formativo. Considerando il nostro obiettivo di analisi si continua ad analizzare i primi due anni.

Tuttavia, è importante notare che il manifesto degli studi è stato oggetto di cambiamenti a partire dall'anno accademico successivo al 2016-2017. Considerando queste modifiche, gli studenti sono stati suddivisi in due sottogruppi: uno rappresentante il periodo precedente al cambiamento del manifesto e l'altro rappresentante il periodo successivo. Da questo punto in avanti, è importante notare che tutti i risultati e le analisi saranno presentati considerando sempre i due sottogruppi definiti in base al periodo del manifesto degli studi: questa divisione è

fondamentale quando si tratta di situazioni in cui l'ordine degli esami e la struttura dei corsi possono influenzare significativamente i risultati. Mantenendo questa distinzione, si garantisce che le analisi siano coerenti con il reale flusso degli esami in ciascun sottogruppo e che tengano conto delle variazioni introdotte dai cambiamenti nel piano di studio.

In seguito, le figure mostrano rispettivamente il manifesto che gli studenti dovrebbero seguire nel primo e nel secondo anno in base agli anni accademici sopra descritti.

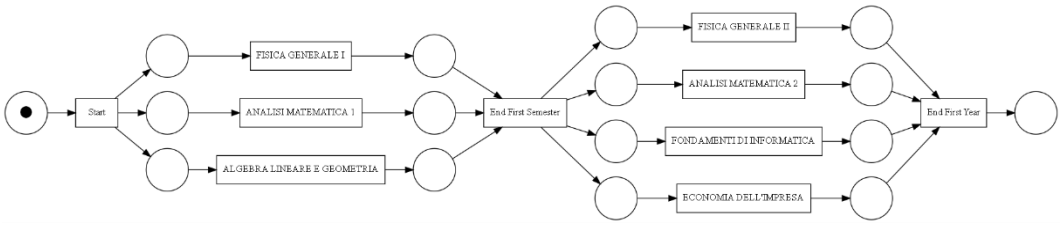


*Fig. I.1 Manifesto degli studi primo anno*

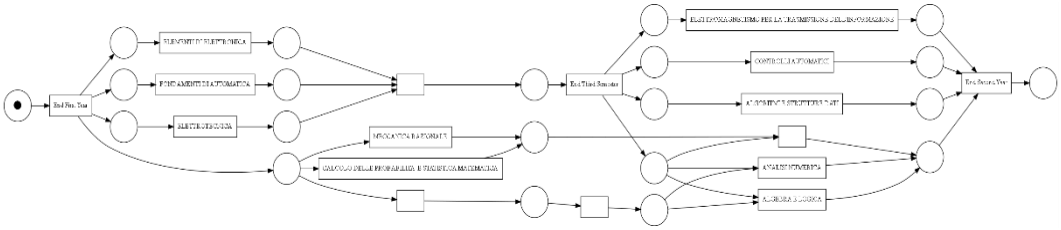

*Fig. I.2 Manifesto degli studi secondo anni accademici dal 2015 al 2016*

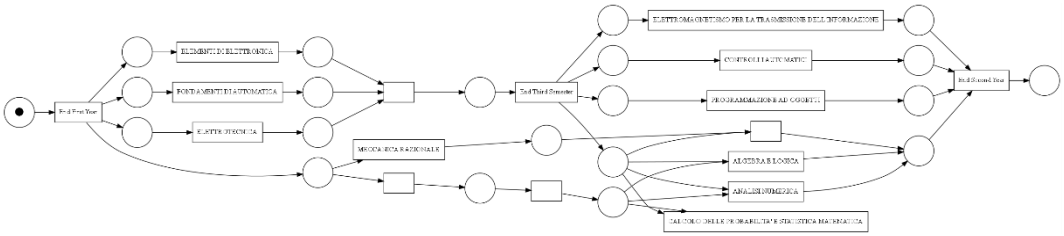


*Fig. I.3 Manifesto degli studi secondo anni accademici dal 2017 al 2019*

È rilevante sottolineare che i cambiamenti al manifesto degli studi sono stati introdotti a partire dal secondo anno accademico. Pertanto, è importante notare che, per il primo anno accademico considerato, il manifesto degli studi rimarrà invariato tra i due sottogruppi di studenti. A partire dal secondo anno accademico invece, i due sottogruppi rifletteranno le differenze nelle strutture dei corsi e negli ordinamenti degli esami: l'esame "Algoritmi e strutture dati" presente nella Fig. I.2 viene sostituito dall'esame "Programmazione ad Oggetti" nella Fig. I.3, inoltre l'esame "Calcolo delle Probabilità e Statistica Matematica" che rappresenta uno degli esami a scelta, passa dal primo al secondo semestre.

Queste riassegnazioni hanno un impatto significativo sull'organizzazione del curriculum accademico, è quindi cruciale tener conto di questo cambiamento nell'analisi. La metodologia utilizzata garantirà che le differenze nell'ordine e nella sequenza degli esami siano correttamente considerate, contribuendo così a una valutazione accurata dei risultati accademici degli studenti nei diversi periodi di studio.

| *Anno accademico* | *Studenti* | *In tempo* | *Un anno in ritardo* | *In ritardo* |
|---|---|---|---|---|
| *2015-2016* | 98 | **38%** | **39%** | **23%** |
| *2016-2017* | 93 | **46%** | **30%** | **24%** |

*Tab. I.1 Indicatori studenti iscritti agli anni accademici 2015 e 2016*

La tabella (Tab. I.1) valuta la performance degli studenti per un determinato anno accademico calcolando la percentuale di studenti che si sono laureati in tempo a paragone con gli studenti che hanno impiegato un anno in più e studenti in ritardo. La seguente tabella presenta dati che rivelano un aspetto cruciale della realtà accademica: circa il 62% degli studenti iscritti all'università nell'anno accademico 2015/2016 si scontra con la sfida di non riuscire a completare i propri studi entro il termine previsto. In contrasto, soltanto il 38% è in grado di conseguire la laurea entro il periodo atteso. Questa rivelazione getta luce su un fenomeno che, sebbene mostri qualche miglioramento nell'anno successivo, si qualifica ancora come una sfida: la percentuale di studenti che concludono il percorso formativo entro il tempo previsto si attesta ancora al di sotto del 50%.

L'analisi dei dati svolta in questo studio si concentra principalmente sugli anni accademici 2015-2016 e 2016-2017, nonostante il dataset includa dati fino all'anno accademico 2019-2020, ma si mostrerà comunque i risultati relativi a tutti gli anni di iscrizione. Questa scelta è motivata dalla necessità di ottenere risultati più

affidabili e rappresentativi, soprattutto quando si considera la categoria degli studenti laureati in ritardo.

La ragione è legata alla necessità di avere un intervallo di tempo sufficientemente ampio per ciascuna categoria. Ciò ci consente di ottenere risultati più realistici e significativi, soprattutto per la categoria oltre un anno di ritardo, in modo da catturare una gamma più ampia di studenti che hanno conseguito la laurea in un arco di tempo più esteso. Questo è particolarmente rilevante per gli studenti che hanno completato gli studi in ritardo poiché includere gli anni accademici fino al 2019-2020 potrebbe sottostimare le percentuali di laureati in tempo, dato che i dati relativi agli anni successivi sono ancora in fase di raccolta.

In sostanza, la limitazione agli anni accademici 2015-2016 e 2016-2017 ci consente di ottenere una panoramica più accurata dei risultati per le diverse categorie di laureati, garantendo che i dati siano rappresentativi e riflettano in modo adeguato la realtà del percorso accademico degli studenti, specialmente quelli che hanno completato gli studi con un ritardo maggiore.

L'analisi dei tempi di completamento degli studi universitari riveste un'importanza di fondamentale rilevanza sia per gli studenti che per le istituzioni accademiche.

L'analisi della tabella (Tab.1) evidenzia inequivocabilmente quanto la percentuale di studenti che riesce a rispettare la tempistica pianificata sia significativamente bassa. Proprio in questo contesto si inserisce il presente studio, con l'intento di scavare nelle radici del tasso di completamento degli studenti e di sondare le

conseguenze di tali risultati per il miglioramento del rendimento accademico e il futuro successo professionale degli studenti stessi.

In tal modo, si aspira a mettere in luce i motivi che sottostanno a questa realtà, aprendo la strada all'identificazione di soluzioni volte a migliorare le prospettive di completamento degli studi e allineare l'esperienza universitaria con le aspettative degli studenti e delle istituzioni accademiche.

## 1.3 STRUTTURA DELLA TESI

Di seguito vengono fornite le principali sezioni che compongono questa tesi:

- *Fondamenti Teorici:*

  In questa sezione, verranno fornite le basi teoriche che costituiscono il quadro concettuale del lavoro. Si approfondiranno concetti chiave come il Process mining e le diverse tecniche di regressione utilizzate. Questa parte svolge un ruolo fondamentale nell'elaborazione delle metodologie applicate.

- *Metodologia della Ricerca:*

  Nella presente sezione verrà descritta in dettaglio la metodologia adottata per condurre l'indagine. Vengono spiegate le domande di ricerca, l'approccio complessivo e le varie fasi coinvolte, inclusa l'estrazione dei modelli di processo, la verifica di conformità, le analisi di correlazione e le previsioni.

- *Risultati:*

  Questa sezione presenta i risultati ottenuti attraverso l'applicazione delle metodologie descritte. Verranno esposti i dati relativi ai processi accademici negli anni considerati, nonché le analisi statistiche e le predizioni effettuate. Inoltre, verranno affrontate le limitazioni dell'indagine e le eventuali aree in cui potrebbero essere approfondite ulteriori ricerche.

Infine, nella sezione *Conclusioni* verranno sintetizzate le conclusioni principali tratte dall'analisi dei risultati. Saranno discusse le implicazioni dei risultati e potenziali direzioni future di ricerca.

La struttura della tesi è stata progettata per fornire un quadro completo e logico dell'intero percorso di ricerca, dall'introduzione delle tematiche all'analisi dei risultati. In modo da guidare il lettore attraverso i vari passaggi dello studio in modo chiaro e organizzato.

# 2. FONDAMENTI TEORICI

In questo capitolo l'obiettivo è quello di fornire una panoramica generale delle principali teorie, concetti e modelli che verranno esplorati nel dettaglio successivamente. Nel corso delle prossime sezioni, verranno affrontate una serie di teorie chiave, esplorando i loro principi fondamentali e la loro rilevanza nell'ambito di questo studio.

Si spiegherà il concetto cruciale di Process Mining (PM), che costituirà il fulcro delle analisi, comprendendo a pieno cosa si intende per PM e come esso fornisce un quadro fondamentale per l'analisi dei processi. Dopo aver gettato le basi concettuali del PM, si vedrà le metodologie quantitative e i modelli statistici che hanno contribuito allo sviluppo di questa analisi. Questi strumenti analitici ci consentono di esplorare relazioni, identificare pattern e trarre conclusioni affidabili dai dati raccolti.

I fondamenti teorici ci offrono una comprensione essenziale delle metodologie e delle tecniche che costituiscono il nucleo di questo studio.

Prima di approfondire nel dettaglio ciascuna teoria, si fornisce una panoramica di alto livello delle principali categorie concettuali che si esamineranno:

- *Regressione Lineare:*

  Un metodo statistico per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti. Si vedrà come questo modello può essere utilizzato per identificare trend e previsioni.

- *Regressione Logistica:*

  Una tecnica utilizzata per modellare la relazione tra variabili indipendenti e una variabile dipendente binaria. Si vedrà come questa tecnica trova applicazioni in previsioni di eventi dicotomici.

- *Support Vector Machine (SVM):*

  Una tecnica di classificazione che utilizza i vettori di supporto per creare un modello in grado di separare efficacemente le categorie dei dati.

Una volta stabiliti i concetti di base dell'analisi quantitativa e dei modelli statistici, ci si concentra sull'introduzione al campo del PM. Questo settore si basa sulla scoperta, la comprensione e il miglioramento dei processi attraverso l'analisi dei dati. Successivamente si spiegherà il concetto di *Process Discovery* e l'importanza del *Conformance Checking*.

Nelle sezioni seguenti, si affronterà più in dettaglio le diverse aree del PM, evidenziando come queste teorie si traducono in pratica nell'analisi e nell'ottimizzazione dei processi.

## 2.1 PROCESS MINING

Il PM è una delle tante branche della Data Science, focalizzata sui processi. Non è una branca del data mining, ma è una disciplina a sé stante che si concentra nell'analizzare, con tecniche nuove, i processi relativi a qualunque tipo di organizzazione. Ci sono molti casi in cui l'aspetto statico del dato, che è quello su cui si basa il data mining, non è sufficiente.

Il focus del PM è legato all'aspetto della generazione di eventi e i dati ad essi correlati: qualsiasi tipo di evento che interagisce con un dispositivo produce informazione. L'obiettivo del PM è estrarre valore dall'evolversi degli eventi e dai dati che gli eventi generano, si concentra sulla traccia che l'esecuzione di un processo lascia nei sistemi informativi.

Il PM aggiunge la prospettiva dei processi al machine learning e data mining, confronta i *dati sugli eventi (event data)*, ossia, il comportamento osservato e i *modelli di processo (process model)* che possono essere creati manualmente o scoperti automaticamente: i dati sugli eventi sono correlati a modelli di processo espliciti, ad esempio reti di Petri o modelli BPMN e i modelli di processo vengono scoperti a partire dai dati sugli eventi o i dati sugli eventi vengono riprodotti sui modelli per analizzare la conformità e le prestazioni.

La prospettiva dei processi è assente in molte iniziative legate ai Big Data e nei programmi di studi della data science. Il PM può essere considerato come il collegamento mancante tra data science e process science. Gli approcci della data

dcience tendono a ignorare il contesto dei processi, mentre gli approcci della process science tendono a essere guidati dai modelli senza considerare le "evidenze" nascoste nei dati.

Si utilizza il termine "process science" per identificare la disciplina più ampia che combina conoscenze provenienti dalla tecnologia dell'informazione e dalle scienze gestionali al fine di migliorare ed eseguire processi operativi.

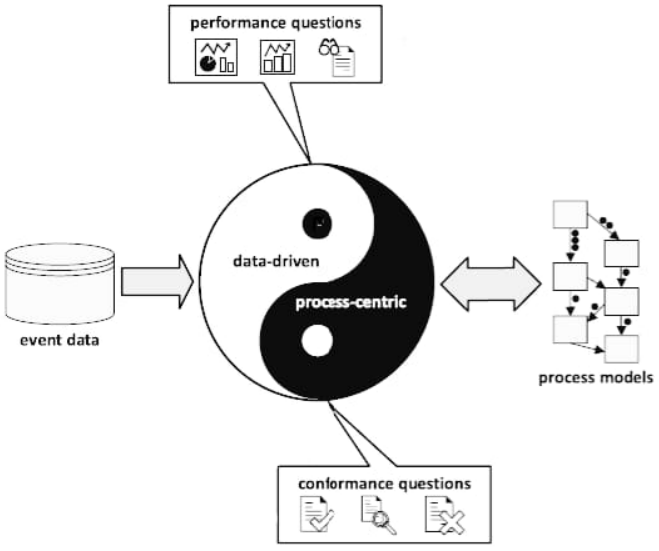


*Fig. II.1 Process mining: l'anello mancante*[1]

Un modo per caratterizzare il PM è mostrato nella Fig. II.1, la quale mostra che il PM parte dai dati sugli eventi e utilizza modelli di processo in vari modi, ad esempio, i modelli di processo vengono scoperti dai dati sugli eventi che fungono da modelli di riferimento o vengono utilizzati per individuare i punti critici.

[1] Fonte: Wil van der Aalst, Process Mining Data Science in Action, Second Edition, Springer, 2016

Il PM è sia data-driven che process-centric: utilizzando una combinazione di dati sugli eventi e modelli di processo si può rispondere a una vasta gamma di domande in riferimento a *performance* e *conformance*, qui di seguito alcuni esempi.

Domande inerenti alla *perfromance*:

- Perché determinati casi sono in ritardo?
- Quali risorse sono sovraccaricate?

Domande inerenti alla *conformance*:

- Quali attività vengono spesso saltate?
- Quali risorse causano deviazioni?

Il PM si concentra sull'analisi dinamica del processo, ossia guarda ad una serie di attività nella loro successione temporale per ricostruire l'insieme delle attività relative ad un particolare processo. L'obiettivo è quello di estrarre conoscenza utile dall'analisi dei processi. Non sarà quindi una ricerca puntale su un determinato evento, perché questa può essere espletata con una query su un database. L'analisi del PM è basata, invece, su aspetti dinamici e questo tipo di analisi si può fare solo considerando n diverse esecuzioni reali di un processo.

Quando si analizza un processo bisogna porsi quattro domande fondamentali:

1. *Che cosa è avvenuto?*
2. *Perché è successo?*
3. *Cosa accadrà?*
4. *Qual è il meglio che può accadere?*

La prima domanda permette di ricostruire la sequenza degli eventi per capire come è andato il processo, se tutti gli eventi accadono nello stesso ordine o se qualche fase devia rispetto alla media dei processi avvenuti. Ottimizzare il processo imparando dagli errori osservati dai dati del passato.

Il PM si pone quindi a metà strada tra **l'analisi dei modelli di processo tradizionale**, che è l'analisi di schema che descrive come vanno fatte le attività, e un **approccio basato puramente sui dati**, ossia data-oriented analysis (tipico del data mining e della business intelligence). Fino all'avvento del PM si potevano fare solo le analisi statiche, si poteva analizzare solo lo schema del processo, per fare pianificazione (a chi assegnare ogni attività) o simulazione (che conseguenze ha sul processo fare un'attività o l'altra), ma ad oggi si estraggono le informazioni legate agli eventi sui sistemi informatici dell'organizzazione.

Inoltre, si pone a metà strada tra altri due tipi di obiettivi: delle **performance** (analisi dei costi e dei tempi) e l'altro è quello di capire se ci sono state delle esecuzioni errate **(compliance).**

L'analisi dei dati prodotti dal sistema, perciò, permette di individuare azioni al di fuori dagli schemi predefiniti, e permette di individuare modi alternativi di eseguire il processo, oppure di individuare anomalie.

Quindi, si può affermare che il PM rappresenta un campo che combina tecniche di data mining, machine learning e *business process*[2] allo scopo di analizzare i dati legati agli eventi e ricavare informazioni di valore riguardo alle dinamiche sottostanti dei processi reali. In questa pratica, l'obiettivo principale è quello di acquisire una comprensione più approfondita dell'effettiva esecuzione dei processi attraverso l'analisi dei dati sugli eventi, che catturano il comportamento effettivo dei processi.

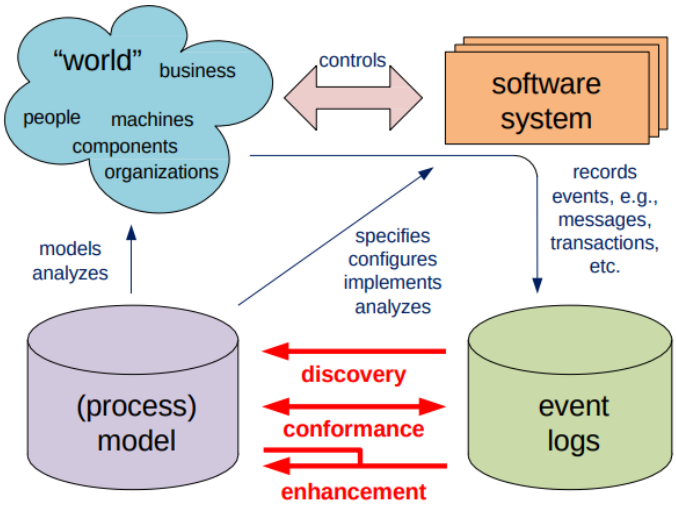


*Fig II.2 I tre principali tipi di Process Mining*[3]

[2] Per *business process* si intende l'esecuzione reale di un processo così come viene registrato dai sistemi dell'organizzazione.

[3] Fonte: Wil van der Aalst, Process Mining Data Science in Action, Second Edition, Springer, 2016

La Fig. II.2 mostra che il sistema informativo registra gli eventi che avvengono all'interno di un database (*event log*), e su di esso si possono effettuare diversi tipi di operazioni. Se si hanno solo i dati, si può generare un modello di processo, ossia lo schema. Mentre se si hanno sia lo schema che i dati si può fare anche conformance checking, che è anche propedeutica per il *process enhancement*.

Il processo di mining evidenzia **tre principali ambiti applicativi:**

Il primo tipo è *process disocvery*. Una tecnica di scoperta che prende un event log e genera un modello senza utilizzare alcuna informazione a priori. Ad esempio, l'algoritmo α è in grado di costruire automaticamente la rete di Petri senza utilizzare alcuna conoscenza aggiuntiva. Se l'event log contiene informazioni sulle risorse, è anche possibile scoprire modelli correlati alle risorse.

Il secondo tipo è *conformance checking*. In questo caso, un modello di processo esistente viene confrontato con l'event log dello stesso processo. La conformance checking può essere utilizzata per verificare se la realtà, come registrata nel log, è conforme al modello e viceversa. Ad esempio, l'analisi dell'event log può mostrare se una regola viene seguita o meno. Un altro esempio è la verifica del *"fuor-eyes" principle*, che stabilisce che determinate attività non dovrebbero essere eseguite dalla stessa persona. Scansionando il registro degli eventi utilizzando un modello che specifica questi requisiti, è possibile individuare potenziali casi di frode. Pertanto, la verifica della conformità può essere utilizzata per rilevare, individuare e spiegare deviazioni.

Il terzo tipo è *enhancement*. In questo caso, l'idea è estendere o migliorare un modello di processo esistente utilizzando informazioni sul processo reale registrato in un event log. Mentre la conformane checking misura l'allineamento tra modello e realtà, questo terzo tipo di process mining mira a modificare o estendere il modello a priori. Un tipo di ottimizzazione è il *repair*, ovvero la modifica del modello per riflettere meglio la realtà. Ad esempio, se due attività sono modellate in sequenza ma nella realtà possono verificarsi in qualsiasi ordine, allora il modello può essere corretto per riflettere ciò. Un altro tipo di ottimizzazione è *l'extension*, ovvero l'aggiunta di una nuova prospettiva al modello di processo mediante la correlazione con il log. Un modo è l'estensione di un modello di processo con dati di prestazione. Ad esempio, utilizzando i timestamp nell'event log è possibile estendere un processo per mostrare vincoli, tempi di throughput e frequenze. Un altro modo può essere estendere con informazioni sulle risorse, regole decisionali, metriche di qualità, ecc.

Considerando le prospettive di analisi del PM molti algoritmi seguono una prospettiva control-flow, ossia cercano di capire l'ordine delle attività svolte. Tuttavia, quando si estendono i modelli di processo, vengono considerate prospettive aggiuntive. Inoltre, le tecniche di discovery e conformance non sono limitate al control-flow.

Altri tipi di approcci:

- *Organizational perspective:*

  La prospettiva organizzativa si focalizza sulle informazioni relative alle risorse interne nel log, come ad esempio quali attori sono coinvolti nell'attività, come ad esempio persone, ruoli, o dipartimenti e come sono correlati. Una volta capito chi ha svolto quale attività, è interessante analizzare il flusso delle relazioni tra i dipendenti e le collaborazioni. Permette di analizzare non solo la struttura del processo ma anche le dinamiche organizzative. L'obiettivo è strutturare l'organizzazione.

- *Case prespective:*

  La prospettiva del caso si concentra sulle proprietà dei casi. Si possono fare delle analisi individuali sui singoli processi.

- *Time prespective:*

  La prospettiva temporale si occupa dei tempi e della frequenza degli eventi. Quando agli eventi corrisponde un timestamp, è possibile scoprire i colli di bottiglia, monitorare l'utilizzo delle risorse e prevedere il tempo di elaborazione rimanente dei casi in corso.

I tre tipi di PM: scoperta, verifica della conformità e ottimizzazione possono essere collegati tre fasi essenziali ("Play-In", "Play-Out" e "Replay") attraverso le quali vengono acquisiti, simulati e confrontati i dati relativi all'esecuzione di un processo rispetto al suo modello teorico.

Si illustra cosa rappresentano ciascuno di questi concetti:

- *Play-In:*

  Il “play-in” rappresenta la raccolta di dati reali riguardanti l'esecuzione di un processo. I dati raccolti durante il play-in costituiscono una rappresentazione concreta e dettagliata dell'esecuzione pratica del processo. Si parte quindi da un log, che descrive la successione degli eventi per poi costruire uno schema che rappresenta tutte le successioni possibili.

- Play-Out:

  Il “play-out” coinvolge l'utilizzo del modello teorico del processo per eseguire una simulazione, il modello può essere una rete di Petri o un altro tipo di modello. Il play-out crea una simulazione che illustra come il processo dovrebbe idealmente svolgersi secondo il modello, partendo dallo schema del processo si possono simulare quali sono le possibili strade. Si possono simulare n esecuzioni del processo.

- *Replay:*

  Il “replay” comporta il confronto tra i dati reali raccolti durante il play-in e la simulazione ottenuta tramite il play-out. L'obiettivo del replay è valutare quanto bene l'esecuzione reale del processo corrisponda al modello teorico. Le discrepanze rilevate durante il replay possono indicare differenze, inefficienze o altre informazioni utili per il miglioramento del processo. In pratica, dallo schema e dalle esecuzioni reali si controlla elemento per

elemento se si trova all'interno dello schema e se è nell'ordine corretto. Si fa il replay della successione degli eventi reali sullo schema fisso per vedere se c'è una corrispondenza.

In sintesi, queste tre fasi costituiscono un ciclo di analisi nel PM, in cui si acquisiscono dati reali, si simulano scenari ideali e si confrontano i dati reali con il modello teorico. Questo ciclo aiuta a identificare opportunità di miglioramento, valutare l'allineamento tra l'esecuzione reale e quella ideale, e prendere decisioni basate su dati concreti per ottimizzare i processi.

Per applicare le tecniche di PM è necessario estrarre le informazioni da un **file log.**

| Case id | Event id | Properties | | | | |
|---|---|---|---|---|---|---|
| | | Timestamp | Activity | Resource | Cost | ... |
| 1 | 35654423 | 30-12-2010:11.02 | register request | Pete | 50 | ... |
| | 35654424 | 31-12-2010:10.06 | examine thoroughly | Sue | 400 | ... |
| | 35654425 | 05-01-2011:15.12 | check ticket | Mike | 100 | ... |
| | 35654426 | 06-01-2011:11.18 | decide | Sara | 200 | ... |
| | 35654427 | 07-01-2011:14.24 | reject request | Pete | 200 | ... |
| 2 | 35654483 | 30-12-2010:11.32 | register request | Mike | 50 | ... |
| | 35654485 | 30-12-2010:12.12 | check ticket | Mike | 100 | ... |
| | 35654487 | 30-12-2010:14.16 | examine casually | Pete | 400 | ... |
| | 35654488 | 05-01-2011:11.22 | decide | Sara | 200 | ... |
| | 35654489 | 08-01-2011:12.05 | pay compensation | Ellen | 200 | ... |
| 3 | 35654521 | 30-12-2010:14.32 | register request | Pete | 50 | ... |
| | 35654522 | 30-12-2010:15.06 | examine casually | Mike | 400 | ... |
| | 35654524 | 30-12-2010:16.34 | check ticket | Ellen | 100 | ... |
| | 35654525 | 06-01-2011:09.18 | decide | Sara | 200 | ... |
| | 35654526 | 06-01-2011:12.18 | reinitiate request | Sara | 200 | ... |
| | 35654527 | 06-01-2011:13.06 | examine thoroughly | Sean | 400 | ... |
| | 35654530 | 08-01-2011:11.43 | check ticket | Pete | 100 | ... |
| | 35654531 | 09-01-2011:09.55 | decide | Sara | 200 | ... |
| | 35654533 | 15-01-2011:10.45 | pay compensation | Ellen | 200 | ... |

*Fig. II.3 Esempio di file log*[4]

[4] Fonte: Wil van der Aalst, Process Mining Data Science in Action, Second Edition, Springer, 2016

La Fig. II.3 propone un esempio di file log: si ha un case id, che identifica il caso e serve per tracciare tutte le attività svolte in quel processo (permette di ricostruire la cronologia delle attività). Tutti gli eventi con lo stesso case id corrispondono a tutte le attività svolte in un determinato processo. Altre informazioni rilevanti sono il nome dell'attività e il timestamp. Il timestamp permette di calcolare la durata del processo: differenza tra quando il processo si chiude e l'istante di tempo quando il processo si è aperto.

All'interno del file log ci possono essere anche altre informazioni, come ad esempio il nome di chi ha svolto l'attività (permette ulteriori analisi sulla rete sociale dell'organizzazione) e il costo dell'attività.

Un processo è costituito da casi, e un caso è costituito da eventi in modo tale che ogni evento sia riconducibile solo ad un caso. Gli eventi all'interno di un caso sono ordinati e possono avere attributi.

I file con cui vengono registrati i log hanno un formato **XES (eXtensible Event Stream)**. Un log consiste in:

- **Traccia**:

  Serie di eventi che riguardano lo stesso caso, una traccia che riguarda un particolare processo **(istanza di processo)**.

- **Evento**:

  Esecuzione di una particolare attività in un certo istante di tempo.

Per concludere il quadro teorico sul PM si evidenza quelle che sono le differenze con il data maining: entrambe partono dai dati, ma il data mining non ha una prospettiva dinamica, le sue tecniche non sono process-centric.

Gli aspetti come il process discovery (scoprire uno schema dai dati), la conformance checking (confrontare i dati con lo schema) e l'analisi dei colli di bottiglia non sono coperte dalle tecniche di data mining tradizionali.

Inoltre, il PM utilizza gli event log, dove gli eventi sono caratterizzati da un timestamp, e si riferiscono ai casi (istanze di processo), mentre gli algoritmi di data mining vengono applicati a dataset eterogenei provenienti da varie fonti.

In sintesi, la differenza chiave sta nel focus dell'analisi: il data mining mira a scoprire pattern e relazioni in dataset eterogenei, mentre il PM si concentra sulla ricostruzione e l'analisi dei processi basandosi sui dati degli eventi collegati a un processo specifico.

Sebbene rappresentino due tipi di analisi differenti, process mining e data mining possono essere combinate per rispondere ad esigenze più complesse.

### 2.1.2 Process Model

L'obiettivo di un modello di processo è quello di decidere quali attività devono essere eseguite e in quale ordine. Le attività possono essere eseguite in sequenza, facoltative o simultanee e l'esecuzione ripetuta della stessa attività può essere possibile.

Nel contesto del PM, vengono utilizzati diversi modelli per rappresentare e analizzare le dinamiche dei processi di un'organizzazione. Questi modelli forniscono una rappresentazione astratta e strutturata dei processi, consentendo di esaminare le attività, le interazioni e le sequenze che si verificano nel corso delle operazioni. Alcuni dei principali modelli utilizzati in questo ambito includono le reti di Petri e le notazioni come BPMN (Business Process Model and Notation). Ognuno di questi modelli offre una prospettiva unica sul modo in cui i processi vengono eseguiti e può essere adattato per soddisfare diverse esigenze analitiche.

Tra questi, le reti di Petri sono utilizzate per visualizzare le relazioni causali tra le attività, offrendo un'immagine chiara delle sequenze possibili e delle condizioni necessarie per il progresso del processo. Allo stesso tempo, le notazioni come BPMN consentono una rappresentazione grafica più intuitiva dei processi, in cui le attività, le decisioni e i flussi possono essere facilmente compresi da stakeholder diversi.

L'utilizzo di questi modelli consente agli esperti di PM di scoprire, valutare la conformità e migliorare i processi in modo più efficiente ed efficace.

Esistono vari modelli per visualizzare un processo: Petri Nets, YAWL, BPMN, Event-Driven Modelling Process Chains (EPCs), Casual Nets e Process Trees.

La visualizzazione **Petri Net** è quella utilizzata per sviluppare questo studio.

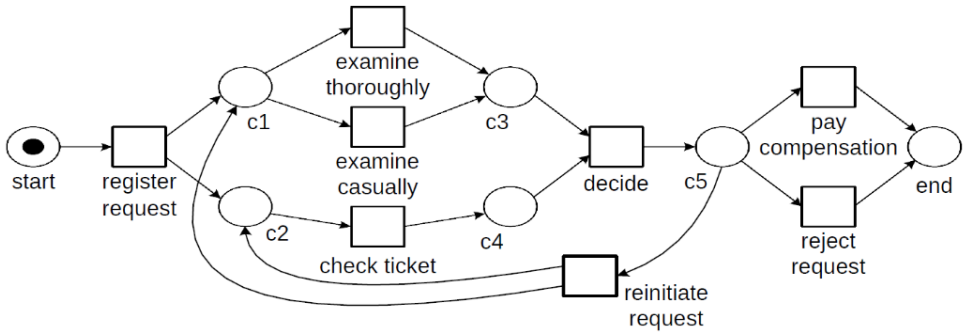


*Fig. II.4 Esempio di visualizzazione Petri Net*[5]

Una rete di Petri è un modello grafico utilizzato per rappresentare e analizzare i sistemi dinamici, in particolare i processi che coinvolgono attività e interazioni tra elementi. Utilizzando le reti di Petri, si può estrarre informazioni su come il sistema si comporta in varie condizioni, oppure si potrebbe identificare potenziali colli di bottiglia e controllare le prestazioni delle attività.

Gli elementi fondamentali di una rete di Petri sono **place**, **transizioni** e **token**.

La rappresentazione grafica è un grafo orientato (digramma) con archi ponderati formati da place e transizioni. I place sono rappresentati da cerchi e rappresentano stati o condizioni che il sistema può assumere. Le transizioni sono rappresentate da quadrati e rappresentano le attività, le azioni o le decisioni che possono verificarsi nel processo. La connessione tra posti e transizioni è data da archi direzionati, i quali collegano place alle transizioni e viceversa. Un place collegato a una

[5] Fonte: Wil van der Aalst, Process Mining Data Science in Action, Second Edition, Springer, 2016

transizione indica che è necessario che una condizione specifica sia soddisfatta affinché la transizione possa essere attivata.

I token si muovono tra place e transizioni, riflettendo i cambiamenti nello stato del sistema quando si verificano gli eventi: permette di capire dove è arrivato il processo, si può spostare tra i diversi place, e permette di individuare tutti i possibili percorsi che il processo può svolgere.

Per l'interpretazione della visualizzazione è importate sottolineare che se una transizione è collegata a due place significa che le due transizioni devono essere effettuate in parallelo. Mentre, se due transizioni si diramano da un place, devono essere scelte esclusivamente.

L'evoluzione di una rete di Petri può essere simulata seguendo le attivazioni delle transizioni e gli spostamenti delle risorse tra i place. Questa simulazione consente di comprendere come il processo si evolve nel tempo e quali sequenze di attività sono possibili.

In sintesi, una rete di Petri è uno strumento visivo che aiuta a comprendere e analizzare le dinamiche di un processo o di un sistema. Attraverso le sue connessioni e le attivazioni delle transizioni, offre un modo intuitivo per esplorare le interazioni tra attività e condizioni e per valutare il comportamento complessivo del sistema.

Questi modelli sono caratterizzati da diverse proprietà fondamentali che consentono di analizzarne il comportamento e le caratteristiche.

Alcune delle proprietà chiave delle reti di Petri includono:

- *Raggiungibilità*:

  Una rete di Petri è raggiungibile se è possibile raggiungere uno specifico stato (marcatura[6]) partendo da uno stato iniziale. Questa proprietà è utile per capire se tutti gli stati possibili possono essere raggiunti nel processo.

- *Limitatezza*:

  Un place $p$ di una rete si dice k-limitato se, in tutte le marcature raggiungibili a partire da quella iniziale, il numero di token presenti del place non supera mai il valore k. Una rete si dice k-limitata se tutti i posti sono k-limitati. Una rete si dice limitata se è k-limitata per almeno un valore finito di k.

  Una rete di Petri è quindi limitata se il numero di risorse (nei place) non può superare un certo limite. Questo può essere importante per assicurarsi che il processo non vada in uno stato di sovraccarico.

- *Binarietà o Sicurezza*:

  Una rete di Petri k-limitata con k=1 si dice binaria o sicura. Ne deriva che, tutte le possibili marcature della rete contengono solo 0 o 1.

[6] Le transizioni rappresentano eventi o azioni che possono verificarsi, mentre i place rappresentano le condizioni o gli stati del sistema. La marcatura di un place indica quanti "gettoni" (o "token") sono presenti in quel place in un dato momento.

- *Vivacità*:

  Una rete di Petri è viva se da qualsiasi stato raggiungibile è sempre possibile attivare almeno una transizione; quindi, una rete si dice viva se tutte le sue transizioni sono vive.

  Questa proprietà delle reti di Petri rappresenta se la rete può continuare ad evolvere a partire da un qualunque stato si trova oppure si blocca (tutta o in parte).

  Questo assicura che il processo possa progredire continuamente senza bloccarsi.

- *Conservatività*:

  Una rete di Petri è conservativa se la somma delle risorse presenti nei posti rimane costante durante l'esecuzione del processo. Questo è importante per garantire che le risorse non siano create o perse nel processo.

### 2.1.2 Process Discovery

Durante la fase di process discovery non si ha ancora lo schema, perché l'organizzazione non ne ha ancora definito uno, ma si hanno le successioni di eventi avvenuti riguardanti una determinata serie di attività. Si osserva ciò che è avvenuto e si estrae lo schema del processo, cercando di generalizzare le n esecuzioni dei vari processi osservati.

La finalità del process discovery è quello di creare un modello basato sui dati reali che vengono analizzati. Questo approccio è particolarmente vantaggioso per le organizzazioni che non dispongono ancora di un modello definito e desiderano generarne uno in base alle effettive esecuzioni del processo.

In questa fase, gli algoritmi di PM lavorano sulle tracce degli eventi registrati nei sistemi delle organizzazioni per identificare schemi di comportamento e sequenze di attività.

L'algoritmo **Alpha Miner** è uno dei primi algoritmi introdotti e tra i più semplici per la scoperta di processi. Esso si basa sull'analisi dell'ordine temporale delle attività nel log, l'obiettivo è l'estrazione dello schema sottoforma di rete di Petri. L'algoritmo identifica l'insieme delle attività e le relazioni tra di esse: inizia col definire le diverse tipologie di relazioni (successioni, casualità, parallelismo, e attività mutualmente esclusive), analizza le casistiche, integra le varie parti e le collega insieme per produrre lo schema finale.

Tuttavia, il suo approccio lineare può portare a sovrastimazioni o sottostimazioni delle relazioni tra le attività se ci sono variazioni significative nei dati.

Il problema fondamentale dell'alpha algorithm è che non tiene conto delle occorrenze rare, questo algoritmo tende a mescolare le esecuzioni tipiche con gli outlier, risultando in una rappresentazione che assegna lo stesso peso sia alle tracce frequenti che agli outlier. In sostanza, se anche una singola esecuzione di un'attività nel log si discosta dalla norma, questa variazione verrà trattata con uguale

importanza rispetto alle tracce più frequenti. Questa situazione è poco desiderabile poiché ciò potrebbe portare a modelli di processo che non riflettono accuratamente la dinamica del flusso di lavoro predominante. Inoltre, questa complessità aggiuntiva potrebbe ostacolare l'identificazione dei modelli di processo più significativi e rilevanti all'interno dell'organizzazione.

Per affrontare questo problema, è essenziale considerare algoritmi più avanzati e sofisticati, come il **Heuristic Miner**, il **Genetic Process Mining** o **l'Inductive Miner**, che integrano euristiche, approcci evolutivi e analisi induttive per ottenere una rappresentazione più accurata e bilanciata dei modelli di processo. Questi algoritmi permettono di distinguere meglio tra le tracce frequenti e le eccezioni, fornendo così una visione più completa e dettagliata dei processi.

L'algoritmo **Heuristic Miner** nasce con l'obiettivo di gestire il rumore all'interno dei dati, è un'estensione dell'alpha miner che cerca di affrontare le limitazioni dell'approccio lineare. Utilizza euristiche per identificare relazioni più complesse tra le attività: la relazione di dipendenza viene pesata per la frequenza della relazione e questo consente di creare uno schema che permette di filtrare tutti gli archi che hanno un punteggio inferiore a determinate soglie, ad esempio, eliminando i percorsi più rari. Questo semplifica anche il processo, che permette di avere un focus sulle attività più frequenti, in quanto tiene conto della frequenza degli eventi, ossia conta il numero di volte all'interno del log avvengono determinate successioni.

L'algoritmo analizza quindi il log degli eventi e costruisce un grafo direzionato, evidenziando le connessioni più significative tra le attività.

L'algoritmo **Genetic Process Mining** invece utilizza un approccio simile all'evoluzione genetica, la scoperta del processo si sviluppa attraverso iterazioni. Inizialmente, vengono generati schemi casuali che rappresentano processi. Ogni schema viene valutato in base a quanto bene si adatta ai dati (fitness). Attraverso mutazioni casuali o ricombinazioni tra schemi considerati buoni secondo la fitness, nuovi schemi emergono. La fitness viene nuovamente calcolata per gli schemi mutati.

Questo processo ripetuto porta gradualmente a modelli di processo più accurati ed efficienti. Si conclude scegliendo lo schema con la fitness più alta.

L'approccio genetico è particolarmente utile quando ci sono molte possibili configurazioni del processo e si vuole trovare la migliore soluzione.

L'algoritmo **Region-Based Mining** si concentra sulla scoperta di regioni all'interno del log degli eventi. Una regione è un sottoinsieme coerente di attività che mantengono un certo ordine tra di loro. L'algoritmo individua queste regioni e le utilizza per costruire modelli di processo più accurati. Questo approccio è particolarmente efficace quando il flusso di lavoro ha una struttura ricorrente e complessa.

Infine, l'algoritmo **Inductive Miner (IM)** analizza l'event log per identificare regole di dipendenza tra le attività. Una volta identificate le regole di dipendenza,

l'IM le organizza in un modello di processo. Questo modello può includere sia sequenze lineari di attività che situazioni di scelta ramificate.

L'IM inizia creando un grafo diretto basato sui dati di log. Poiché il grafo iniziale potrebbe contenere molte relazioni ridondanti o poco significative tra le attività, l'algoritmo applica tecniche di riduzione per semplificare il grafo. Ciò aiuta a identificare le sequenze più rilevanti di attività. L'algoritmo cerca poi modelli di processo all'interno del grafo semplificato, questo viene fatto identificando sequenze di attività comuni, che rappresentano i percorsi di esecuzione più frequenti all'interno del processo. Infine, estrae il modello di processo finale utilizzando le sequenze di attività identificate.

In termini di output, l'IM produce un modello di processo che rappresenta la struttura delle attività e delle relazioni tra di esse all'interno del processo dell'organizzazione.

Questo algoritmo è quindi in grado di gestire dati più complessi rispetto ad altri algoritmi ed è noto per la sua capacità di scoprire anche modelli di processo meno evidenti. Poiché si basa sull'analisi delle relazioni ed è in grado di catturare modelli di processo più dettagliati e flessibili. Questo è particolarmente utile quando le sequenze di attività possono variare notevolmente in base a scenari diversi o ad eccezioni occasionali.

Tuttavia, può generare modelli complessi, il che richiede un'attenta analisi e un'eventuale semplificazione. Inoltre, può essere più lento rispetto ad alcuni algoritmi più semplici, data la sua analisi dettagliata.

Di seguito si cerca di chiarire il funzionamento dell'algoritmo IM.

Per quanto riguarda il funzionamento dato un event log l'algoritmo produrrà in output l'albero di processo equivalente, il quale può essere convertito in un equivalente WF-net, modello BPMN, ecc.

L'algoritmo IM divide iterativamente l'event log iniziale in sublogs più piccoli. Per ogni sublog si crea un *directly-follows graph (dfg)*[7].

Per dividere l'event log si devono trovare i cosiddetti tagli nel dfg del (sub)log che si vuole dividere, considerando tagli a scelta esclusiva, tagli di sequenza, tagli paralleli, e tagli redo-loop. I quattro tipi di tagli corrispondono ai quattro operatori degli alberi di processo (×, →, ∧, e ↺).

[7] Un Directly-Follows Graph (dfg) è una rappresentazione visuale di un event log, che evidenzia le sequenze di attività che si verificano in modo consecutivo all'interno di un processo. Questo tipo di grafico viene utilizzato nel contesto del process mining per visualizzare le relazioni di ordine temporale tra le attività che compongono un processo.

Nel dfg, ogni attività è rappresentata da un nodo, e gli archi tra i nodi rappresentano le transizioni da un'attività all'altra. Un arco diretto da un nodo A a un nodo B indica che l'attività B è stata eseguita immediatamente dopo l'attività A. Questa struttura rende evidenti i flussi sequenziali all'interno del processo.

- *Tagli a scelta esclusiva* (×):

Raggruppa le attività in modo tale che le attività appartenenti a gruppi diversi non abbiano relazioni tra loro.

- *Tagli di sequenza* (→):

Il taglio divide l'insieme delle attività in sottoinsiemi disgiunti in modo tale che gli archi vanno solo da sinistra a destra.

- *Tagli paralleli* (∧):

Divide l'insieme delle attività in sottoinsiemi disgiunti in modo tale che ogni attività in un insieme sia collegata a tutte le attività nell'altro insieme (e viceversa).

- *Tagli redo loop* (↺):

L'operatore del loop redo richiede un albero con almeno due eventi. C'è un evento “do” e il resto sono eventi “redo”. L'evento “do” si alterna agli eventi "redo" e il processo inizia e finisce con l'evento “do”.

Usando i sublogs, si creano nuovi dfg, che rappresentano casi base, cioè sottoprocessi costituiti da una singola attività eseguita una volta per caso; quindi, ogni evento finisce in uno dei sublogs.

In sintesi, l’algoritmo IM funziona come segue: dato un event log, viene costruito il dfg. Se c'è un taglio a scelta esclusiva, viene applicato e si divide l’event log in parti più piccole. Mentre se non c'è un taglio a scelta esclusiva, ma c'è un taglio di sequenza, viene applicato quest’ultimo. Se non ci sono tagli a scelta esclusiva e sequenza, viene applicato un taglio parallelo e allo stesso modo se non ci sono tagli

esclusivi, sequenziali e paralleli, ma c'è un taglio redo-loop, viene applicato un taglio redo-loop. Dopo aver diviso l'event log in sublogs, la procedura viene ripetuta finché non viene raggiunto un caso di base (sublog con una sola attività).

Se non si riesce a rilevare nessuno dei quattro tagli spiegati in precedenza, si applica quello che viene chiamato "*fall through*" o passaggio a cascata. La parte che non può essere suddivisa viene rappresentata dal cosiddetto "*flower model*" (modello a fiore), ad esempio considerando le attività da a ad h, il modello a fiore rappresenta un albero di processo che consente qualsiasi traccia che coinvolge le attività da a ad h. Questo modello funge da risorsa finale per garantire la fitness, ma potrebbe comportare una minore precisone.

L'output è un albero di processo dove i nodi interni corrispondono agli operatori utilizzati per tagliare l'event log in sublogs; quindi, i nodi interni di un albero di processi sono annotati con operatori che definiscono l'ordine in cui le attività possono essere eseguite. I quattro tipi di tagli si basano sulle caratteristiche dei quattro operatori degli alberi di processo presumendo che non vi siano attività duplicate o silenziose.

È importante sottolineare che questo algoritmo produce sempre un modello di processo sound[8] in grado di riprodurre l'intero event log. La soundness (correttezza)

[8] Un Workflow Net (WF-net, è una formalizzazione utilizzata nella teoria delle reti di Petri per modellare processi e sistemi concorrenti) è considerato "sound", ovvero corretto o valido, se

è fondamentale per garantire l'affidabilità e l'accuratezza della rappresentazione del modello del processo sottostante. Inoltre, a differenza di molti altri algoritmi, la fitness è garantita. Poiché i modelli sono strutturati in blocchi e le attività non sono duplicate, i modelli tendono ad essere semplici e generali. Tuttavia, gli alberi di processo costruiti dall'algoritmo IM possono essere inadeguati se il comportamento osservato richiede un albero di processo con attività duplicate o silenziose.

L'algoritmo descritto di base non può astrarre dal comportamento infrequente in quanto quest'ultime non sono prese in considerazione: per garantire una perfetta fitness, non tiene conto delle frequenze. Ciò può portare a modelli di processo di qualità inferiore se nel log sono presenti comportamenti poco frequenti o devianti.

Nel tempo sono state sviluppate una famiglia di tecniche di IM:

- *Inductive Miner-infrequent (IMF)*
- Inductive Miner - incompleteness (IMC)
- Inductive Miner - directly-follows based (IMD),

---

vengono rispettate le seguenti proprietà: sicurezza (quando i place nella rete non possono contenere più di un token contemporaneamente); completamento corretto (una volta che una transizione che porta al place di output (o) viene attivata, non dovrebbero esserci altri token nella rete tranne che nel place di output); raggiungibilità (per ogni marcatura, ovvero distribuzione dei token, nella rete, deve essere possibile raggiungere il luogo di output); assenza di parti inattive (la rete non dovrebbe contenere transizioni inattive, quindi transizioni che non possono essere attivate in nessuna circostanza).

- Inductive Miner - infrequent-directly-follows based (IMFD)
- Inductive Miner - incompleteness-directly-follows based (IMCD).

In sintesi, le tecniche di IM inoltre sono in grado di scoprire una classe molto più ampia di processi e apprendere modelli di processo in situazioni nelle quali altri algoritmi falliscono.

Le funzioni utilizzate nell'algoritmo IMF nel rilevamento del taglio sono le stesse utilizzate nell'IM di base: se viene trovato un taglio, si divide il log e la ricorsione continua sui sublog risultanti, se invece non viene trovato alcun taglio, si filtra dfg in base a un parametro di soglia $f$. Quindi, a differenza dell'algoritmo di base, la versione infrequente non prende solo un event log come input, ma anche un parametro $f$.

Utilizzando questo algoritmo si ha la possibilità di visualizzare negli archi i numeri che indicano le frequenze (ad esempio, l'attività b è stata eseguita 880 volte ed è stata seguita direttamente dall'attività c 620 volte). In questo modo si possono visualizzare casi dove se la relazione tra gli archi c e d è poco frequente rispetto a tutti gli altri archi, questo permette di poter filtrare gli archi. Inoltre, è anche possibile filtrare le attività se due attività sono meno frequenti delle altre.

L'algoritmo IMF utilizza vari tipi di filtraggio con l'obiettivo di mostrare il comportamento mainstream, presenta quindi ottime capacità nel gestire il rumore.

In generale, questo algoritmo permette di far fronte a comportamenti infrequenti, gestire event log di grandi dimensioni garantendo la soundness.

L’IMF richiede di impostare una soglia che determini la quantità di comportamento del processo da filtrare. In questo studio per determinare il threshold di ogni processo estratto è stata testata una gamma di soglie da 0 a 1, ad ognuna delle quali corrisponde un determinato valore di: fitness, precisione, generalizzazione e semplicità, è stato selezionato il valore corrispondente al miglior compromesso tra fitness e precisione.

Concludendo, la scelta dell'algoritmo dipende dalla natura dei dati e dagli obiettivi dell'analisi.

Quando si sviluppa un algoritmo di process discovery, è essenziale considerare quattro aspetti fondamentali che influenzano la qualità del modello ottenuto:

- *Fitness:*

  La “fitness” rappresenta la capacità del modello di descrivere accuratamente le diverse tracce di esecuzione presenti nei dati. Un modello ha un'elevata fitness quando è in grado di rappresentare in modo adeguato le sequenze e le dinamiche delle attività riportate nel log. In altre parole, l'obiettivo è che il modello si adatti bene ai dati reali.

  Si ha un’elevata fitness quando lo schema ottenuto corrisponde bene alle tracce nel log.

- *Generalizzazione:*

  La “generalizzazione” si riferisce all'equilibrio tra un modello troppo generale e uno troppo specifico. L'obiettivo è evitare che il modello sia

troppo generico, il che potrebbe portare a una rappresentazione poco precisa e a una perdita di informazioni importanti (underfitting). In altre parole, il modello dovrebbe essere in grado di catturare le caratteristiche principali senza essere troppo semplificato.

- *Precisione:*

  La "precisione" riguarda la capacità del modello di adattarsi esclusivamente ai dati del log specifico, rischiando però di non generalizzare bene per altre situazioni. Un modello eccessivamente preciso può portare a un'adattabilità eccessiva ai dati di esempio, ma potrebbe mancare di rappresentatività per altre tracce o varianti (overfitting).

- *Semplicità:*

  È desiderabile ottenere un modello il più semplice possibile, cioè una rappresentazione che sia chiara e comprensibile, senza eccessive complessità. Questo favorisce l'interpretazione e la comunicazione del modello a diverse parti interessate.

L'obiettivo è trovare un equilibrio tra questi aspetti, cercando un modello che abbia una buona fitness, generalizzazione e precisione, mantenendo al contempo una semplicità ragionevole. Un algoritmo di process discovery di successo deve affrontare queste sfide e produrre un modello che rappresenti accuratamente le dinamiche del processo sottostante in modo comprensibile ed efficiente.

### 2.1.3 Conformance Checking

In questa fase, si dispone già di uno schema e delle informazioni relative alle esecuzioni reali. Il processo di conformance checking confronta ogni singola esecuzione con lo schema e stabilisce se ciascuna esecuzione è in linea con lo schema stesso. L'obiettivo è capire se il modo in cui le attività vengono svolte nell'organizzazione è coerente con lo schema precedentemente definito.

Gli algoritmi utilizzati per il conformance checking hanno il compito di calcolare quante delle esecuzioni effettivamente realizzate sono conformi allo schema nel suo complesso, e in quali punti specifici le attività si discostano. L'analisi delle deviazioni consente di valutare se tali differenze possono portare a vantaggi o risultati migliori, oppure se si tratta di errori che necessitano correzione.

La pratica del conformance checking descrive come una sequenza di attività deve essere eseguita, in quale ordine e quali alternative sono ammissibili durante l'esecuzione. Questa analisi è preziosa poiché aiuta a valutare l'allineamento tra la realtà operativa e il modello teorico.

Il conformance checking costituisce un passo preliminare essenziale per l'ottimizzazione (process enhancement). Se vengono individuate deviazioni che possono effettivamente portare vantaggi, si può procedere con la modifica finale dello schema per incorporare tali deviazioni e ottenere miglioramenti nel processo.

In sintesi, il conformance checking rappresenta un'importante fase di analisi che mira a valutare la coerenza tra le esecuzioni reali e lo schema concettuale, fornendo informazioni preziose per la correzione e l'ottimizzazione del processo.

Per valutare la differenza tra l'esecuzione reale e il modello, vengono considerate diverse metriche e misurazioni, tra cui la *fitness*: è una metrica chiave nel contesto della verifica della conformità. Rappresenta quanto bene il modello teorico si allinea con le esecuzioni reali. In termini di ProM, uno strumento ampiamente utilizzato per il process mining, la fitness può essere calcolata per ciascun caso o per l'intero log degli eventi. Maggiore è la fitness, maggiore è la coerenza tra il modello e l'esecuzione reale.

ProM, uno strumento di process mining, offre diverse tecniche e metodi per il conformance checking. Attraverso analisi avanzate, ProM è in grado di calcolare la fitness, l'allineamento e altre metriche rilevanti per valutare quanto bene il modello concettuale rispecchia l'esecuzione effettiva. Questo aiuta le organizzazioni a individuare in modo preciso le aree di non conformità, identificando le cause delle deviazioni e suggerendo possibili azioni correttive.

Va inoltre sottolineato che le deviazioni possono essere interpretate in maniera sia positiva che negativa. In molti processi, le persone deviano intenzionalmente dal percorso prestabilito, ma, ciò nonostante, producono risultati positivi. Pertanto, le deviazioni non devono essere sempre considerate come elementi negativi. Quindi, sarebbe sbagliato assumere che il modello di processo sia sempre corretto e che

l’event log sia necessariamente errato. La complessità della realtà può spesso richiedere adattamenti e flessibilità nel seguire il percorso stabilito, senza che ciò rappresenti un'irregolarità o un errore.

In sintesi, l'obiettivo è quello di adattare le tracce dei log a un modello, esaminando le deviazioni e valutando la loro efficacia tramite la metrica di fitness, rappresenta un approccio pragmatico per affrontare le sfide nell'analisi dei processi. È fondamentale mantenere una prospettiva equilibrata sul significato delle deviazioni, riconoscendo che possono contribuire positivamente alla flessibilità e all'efficacia del processo, piuttosto che rappresentare necessariamente errori. La comprensione della complessità e della dinamicità dei processi è essenziale per una valutazione accurata e una migliore gestione delle attività.

In conclusione, la verifica della conformità è un processo critico per garantire l'allineamento tra il modello teorico e l'esecuzione reale dei processi. Strumenti come ProM sono essenziali per calcolare metriche come la fitness e l'allineamento, fornendo un quadro dettagliato delle discrepanze e delle opportunità di miglioramento.

## 2.2 REGRSSIONE LINEARE

La regressione lineare è una tecnica statistica utilizzata per analizzare la relazione tra due o più variabili, di solito con l'obiettivo di prevedere o comprendere il comportamento di una variabile di risposta in base alle variazioni di una o più

variabili indipendenti. In particolare, la regressione lineare si concentra sulla ricerca di una relazione lineare approssimata tra queste variabili.

La formula fondamentale della regressione lineare, rappresentata come $y = mx + b$, esprime una relazione lineare tra una variabile di risposta y e una variabile indipendente x. In questa equazione, "m" rappresenta la pendenza della retta, cioè quanto la variabile di risposta cambia al variare della variabile indipendente, mentre "b" indica l'intercetta, il valore della variabile di risposta quando la variabile indipendente è uguale a zero. Determinare i valori ottimali di "m" e "b" è il core della regressione lineare.

La tecnica della regressione lineare si basa sulla minimizzazione dell'errore residuo, ovvero la differenza tra i valori osservati e quelli predetti dalla retta. Il metodo dei minimi quadrati viene utilizzato per calcolare i coefficienti "m" e "b" che minimizzano la somma dei quadrati degli errori. Questo processo garantisce che la retta si avvicini il più possibile ai dati osservati, rendendo la relazione tra le variabili una stima accurata.

La regressione lineare può essere suddivisa in due categorie principali: la regressione lineare semplice e la regressione lineare multipla. Nella regressione lineare semplice, viene esaminata la relazione tra una variabile indipendente e una variabile di risposta. Nella regressione lineare multipla, invece, si considerano più variabili indipendenti, consentendo di esaminare come più fattori influenzino la variabile di risposta.

La regressione lineare ha alcune assunzioni chiave:

- *Linearità:*

  L'ipotesi fondamentale è che la relazione tra le variabili sia approssimativamente lineare. In caso contrario, la regressione lineare potrebbe produrre risultati inaccurati.

- *Indipendenza degli errori:*

  Gli errori residui devono essere indipendenti tra loro e distribuiti normalmente con una media zero.

- *Omoschedasticità:*

  La varianza degli errori residui deve essere costante lungo tutto il range delle variabili indipendenti.

- *Assenza di multicollinearità:*

  Nella regressione lineare multipla, le variabili indipendenti non devono essere fortemente correlate tra loro.

È importante notare che la regressione lineare funziona meglio quando la relazione tra le variabili può essere approssimata in modo lineare. Se i dati seguono un modello non lineare, la regressione lineare potrebbe non essere la scelta migliore, e potrebbero essere necessari modelli più complessi, come la regressione polinomiale o altri tipi di regressione.

In conclusione, la regressione lineare fornisce una base solida per comprendere e analizzare le relazioni tra variabili. Tuttavia, è importante capire le limitazioni del

metodo e considerare attentamente se una relazione lineare è appropriata per i dati a disposizione.

## 2.3 REGRESSIONE LOGISTICA

La regressione logistica è uno strumento fondamentale nell'ambito dell'analisi statistica e del machine learning, rappresenta un metodo statistico utilizzato per modellare la relazione tra una variabile dipendente binaria (variabile di risposta) e un insieme di variabili indipendenti (variabili predittive).

La regressione logistica deriva dalla regressione lineare e si adatta alla necessità di modellare variabili dipendenti binarie, come ad esempio “successo” o “fallimento”, “sì” o “no”, “1” o “0”. Contrariamente alla regressione lineare, che prevede valori continui, la regressione logistica prevede probabilità di appartenenza a una classe specifica. Essa sfrutta la funzione logistica, nota anche come funzione sigmoide, per trasformare la somma ponderata delle variabili predittive in una probabilità compresa tra 0 e 1.

La funzione sigmoide restringe il suo input in un intervallo compreso tra 0 e 1, consentendo quindi di interpretare il risultato come una probabilità.

Il modello di regressione logistica è espresso come:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x1 + \beta_2 x2 + \cdots + \beta_p xp$$

dove:

- $p$ rappresenta la probabilità della variabile di risposta di appartenere alla classe positiva,
- $\beta_0 + \beta_1 x1 + \beta_2 x2 + \cdots + \beta_p xp$ sono i coefficienti di regressione da stimare per ciascuna variabile predittiva $x1,\ x2,\ \ldots,\ xp$.

La stima dei coefficienti di regressione è spesso effettuata utilizzando il metodo dei minimi quadrati massimi, che cerca di minimizzare la differenza tra i valori predetti e quelli osservati. Tuttavia, nella regressione logistica, la massimizzazione della funzione di verosimiglianza viene spesso preferita alla minimizzazione dei residui. Questo perché la variabile di risposta binaria non segue una distribuzione normale; quindi, i metodi basati sui minimi quadrati non sono appropriati.

L'efficacia di un modello di regressione logistica viene valutata attraverso diversi metodi, tra cui la matrice di confusione, l'accuratezza, la precisione, e l'F1-score. La matrice di confusione mostra il numero di previsioni corrette e errate suddivise per classe, mentre l'accuratezza misura la percentuale di previsioni corrette rispetto al totale. La precisione rappresenta la percentuale di istanze positive correttamente classificate rispetto a tutte le istanze classificate come positive,

Come in molti metodi di machine learning, anche la regressione logistica può incorrere nel problema di overfitting, in cui il modello si adatta eccessivamente ai dati di addestramento e ha difficoltà a generalizzare su nuovi dati. Per affrontare questo problema, si possono utilizzare tecniche come la regolarizzazione, che

aggiunge un termine di penalità alla funzione di costo, o ridurre la complessità del modello selezionando attentamente le variabili predittive.

La regressione logistica è uno strumento utilizzato per la modellazione di variabili di risposta binarie. Grazie alla sua capacità di fornire probabilità di appartenenza a una classe specifica, risulta utile in numerosi scenari applicativi.

## 2.4 SUPPORT VECTOR MACHINES

Il Support Vector Machine (SVM) è un algoritmo di classificazione che si basa sull'idea di trovare il bordo decisionale ottimale tra due classi. L'obiettivo è separare le classi nel modo più efficiente possibile. Il miglior bordo decisionale è quello che massimizza la distanza tra i punti delle classi, evitando così possibili errori di classificazione. Se il bordo è vicino ai punti di una delle classi, c'è una maggiore probabilità di errore.

Per ottenere un buon bordo, l'SVM cerca di massimizzare lo spazio tra le due classi, noto come margine, per gestire eventuali variazioni nei dati.

L'SVM può apprendere bordi decisionali lineari o non lineari nello spazio degli attributi per separare le classi. Questo modello è in grado di controllare la complessità al fine di evitare l'overfitting e offre buone prestazioni di generalizzazione. Inoltre, l'SVM è noto per la sua capacità di regolarizzare il suo apprendimento in modo intrinseco.

Un aspetto unico dell'SVM è l'uso di un sottoinsieme di esempi di training chiamati "vettori di supporto" per definire il bordo decisionale. Questi vettori di supporto sono i punti delle classi più vicini al bordo e sono fondamentali nella costruzione del bordo stesso. Di conseguenza, la funzione decisionale dell'SVM è definita da questi vettori di supporto, distinguendolo da altri approcci di apprendimento automatico.

In sintesi, l'SVM è un algoritmo di classificazione che cerca il bordo ottimale per separare due classi, massimizzando il margine tra di esse e utilizzando solo un sottoinsieme di punti di training, i vettori di supporto, per definire il bordo decisionale. Questa metodologia lo rende efficace in diverse applicazioni di machine learning.

L'approccio dell'SVM comporta diversi principi chiave che lo distinguono da altri metodi di classificazione:

- *Margini e Vettori di Supporto:*

  Come accennato, i vettori di supporto sono i punti dati che giacciono più vicino all'iperpiano di regressione. Gli SVM cercano di massimizzare il margine tra l'iperpiano e i vettori di supporto, consentendo un certo errore all'interno di questo margine. Questo approccio offre robustezza contro il rumore nei dati e contribuisce a evitare l'overfitting.

- *Funzioni Kernel:*

  Un aspetto cruciale delle SVM è l'uso di funzioni kernel. Queste funzioni consentono di mappare i dati in uno spazio dimensionale superiore, rendendo possibile la modellazione di relazioni non lineari. I tipi di funzioni kernel includono il kernel lineare (per relazioni lineari), il kernel polinomiale (per relazioni polinomiali) e il kernel a base radiale (RBF, per relazioni complesse).

- *Parametri di Regolarizzazione:*

  Le SVM incorporano anche un parametro di regolarizzazione (C) che controlla l'equilibrio tra la massimizzazione del margine e la classificazione accurata dei punti di addestramento. Un valore più elevato di C porta a un adattamento più preciso ai dati di addestramento, ma potrebbe aumentare il rischio di overfitting.

L'SVM offre diversi vantaggi: grazie all'uso dei kernel, l'SVR può modellare relazioni complesse tra le variabili, anche se sono non lineari, inoltre grazie all'attenzione posta sui vettori di supporto, è in grado di mitigare l'effetto degli outlier. Infine, la scelta del kernel consente di adattare l'SVM a diverse distribuzioni di dati e relazioni tra variabili.

Tuttavia, ci sono alcune limitazioni che vanno considerate, la scelta dei parametri di kernel e di regolarizzazione può influenzare significativamente le prestazioni dell'SVM.

# 3. METODOLOGIA DELLA RICERCA

Nel presente capitolo, si delinea in modo dettagliato l'approccio metodologico adottato per condurre la ricerca proposta. L'obiettivo principale è quello di esaminare le dinamiche dei processi accademici correlati al tempo di laurea e al numero di esami superati. A tal fine, verranno illustrati i passaggi chiave della metodologia implementata, che comprendono: l'estrazione di modelli di processo tramite l'utilizzo dell'IM, l'applicazione di conformance checking ai processi estratti, l'analisi delle correlazioni tra il tempo di laurea e il numero di esami superati, l'identificazione delle varianti ottimali per un minor tempo di laurea e, infine, l'implementazione di un modello predittivo per stimare il tempo di laurea.

In particolare, in questo capitolo verranno presentati gli obiettivi della ricerca, fornendo una chiara struttura per le domande di studio che hanno guidato l'approccio metodologico.

Successivamente, si discute quali approcci sono stati adottati, e a questo proposito la sezione è stata suddivisa in sottosezioni per illustrare in dettaglio ciascun passaggio chiave dell'approccio implementato.

Si illustra come l'IM è stato impiegato per estrarre i modelli di processo dai dati accademici disponibili. Questo passaggio è fondamentale per comprendere la struttura e i flussi dei processi legati al percorso di studio degli studenti.

Si spiega come il conformance checking è stato applicato ai processi estratti al fine di valutare quanto i dati reali dei percorsi di studio si allineino ai modelli di processo ideali delineati.

Successivamente si mostra il metodo con cui si analizza le correlazioni tra il voto medio, il tempo di laurea e il numero di esami superati dagli studenti, offrendo un'ulteriore prospettiva sull'effetto dell'onere accademico sul completamento degli studi.

Si descrive inoltre come sono state identificate le varianti ottimali dei percorsi di studio degli studenti che hanno conseguito il titolo di laurea nel minor tempo, considerando sia la quantità di esami che la sequenza temporale in cui vengono affrontati.

Infine, si illustra com'è stato sviluppato un modello predittivo al fine di stimare il tempo di laurea degli studenti in base alle loro scelte accademiche.

Complessivamente, questo capitolo fornirà un quadro dettagliato della metodologia adottata per condurre la ricerca e raggiungere gli obiettivi di studio.

Prima di procedere con l'analisi degli obiettivi, è opportuno introdurre una precisazione: considerando la tipologia dei dati trattati in questo studio si parla di *Educational Process Mining (EPM)*, ovvero tecniche di Process Mining che trovano applicazione nel campo dell'educazione.

Nell'era digitale, l'istruzione ha superato i confini tradizionali: le istituzioni educative, dalle scuole alle università, stanno sempre più abbracciando strumenti

tecnologici per migliorare l'insegnamento, personalizzare le esperienze di apprendimento e migliorare i risultati educativi complessivi. Una delle metodologie innovative emerse in questo contesto è l'EPM.

L'EPM sfrutta la potenza dell'analisi dei dati e delle tecniche di mining per esplorare, analizzare e ottimizzare il processo educativo, portando ad ambienti di apprendimento più efficaci e migliori prestazioni degli studenti.

È un approccio multidisciplinare che integra concetti di istruzione, data mining, analisi dei processi e machine learning. Il focus dell'EPM coinvolge l'estrazione, l'analisi e l'interpretazione dei dati educativi per acquisire informazioni su come i processi educativi possono essere migliorati. Questa metodologia ruota attorno all'utilizzo di tracce digitali lasciate dagli studenti mentre interagiscono con piattaforme di apprendimento online, software educativi e altre risorse digitali. Queste tracce possono includere dati sulle interazioni, comportamenti, prestazioni, livelli di coinvolgimento degli studenti e altro ancora.

Quindi, l'EPM essenzialmente adatta i principi del processo di estrazione, una tecnica ampiamente utilizzata nella gestione dei processi aziendali, al settore educativo. Proprio come le aziende scoprono modelli, colli di bottiglia e inefficienze nei processi aziendali, EPM rivela modelli nei percorsi di apprendimento degli studenti, identifica gli ostacoli che devono affrontare e individua le aree in cui i processi educativi possono essere ottimizzati.

## 3.1 OBIETTIVI DELLA RICERCA

Nel contesto di questa ricerca, è emerso un obiettivo di cruciale importanza: lo sviluppo di un approccio mirato a consentire alle istituzioni universitarie di acquisire informazioni approfondite e dettagliate in merito ai percorsi effettivamente intrapresi dagli studenti. Questo obiettivo si pone quale tassello essenziale nel panorama dell'ottimizzazione dell'esperienza accademica.

L'attenzione non si limita solo all'estrazione dei processi, ma si estende anche all'analisi approfondita di tali informazioni. In particolare, l'analisi si concentra sull'individuazione di dissonanze e differenze significative che potrebbero emergere tra i sottogruppi di studenti identificati come: "in tempo", "un anno in ritardo" e "in ritardo". Questo approccio consente di cogliere sfumature e pattern che altrimenti potrebbero sfuggire ad un'osservazione più generale, contribuendo a una comprensione più completa e dettagliata della dinamica accademica.

Un elemento aggiuntivo di questo obiettivo è quello di fornire analisi statistiche. Queste analisi permettono di indagare sui risultati ottenuti dai processi e di estrarre informazioni contestualizzate in base alla natura delle analisi condotte (essenziale per comprenderne appieno l'impatto e la rilevanza di variabili o fattori all'interno del contesto accademico).

Complessivamente, l'obiettivo è quello di mettere a disposizione delle istituzioni universitarie uno strumento che mira a offrire una panoramica dettagliata, approfondita e significativa dei percorsi degli studenti, supportando le decisioni e

le strategie accademiche attraverso una base informativa solida. Parallelamente, questo strumento può anche offrire consigli e orientamento agli studenti stessi che potrebbero prendere decisioni più informate sul loro percorso accademico.

Le domande (**D**) che guidano l'analisi sono le seguenti:

- **D1:** Qual è il processo degli studenti che si sono laureati in tempo? Quali sono le principali differenze con il processo degli studenti che si sono laureati con un anno di ritardo e in ritardo?
- **D2:** Considerando il processo ideale che gli studenti dovrebbero seguire, quanto i processi degli studenti sono veramente conformi al manifesto?
- **D3:** Ci sono correlazioni tra laurearsi in tempo o in ritardo e altre variabili, come il numero di esami che gli studenti riescono a sostenere entro il primo semestre o entro il primo anno e il voto medio?
- **D4:** Quali sono le combinazioni di esami del primo anno per le quali corrisponde un minor tempo di laurea?
- **D5:** Quali sono i principali fattori che influenzano le prestazioni accademiche degli studenti? Il tempo che gli studenti impiegano per laurearsi può essere previsto?

## 3.2 DATASET E PRE-PROCESSING

Le informazioni necessarie per affrontare gli interrogativi posti sono state estratte da un'ampia selezione di dati provenienti dal sistema Esse3. Quest'ultimo

costituisce una piattaforma web utilizzata da alcune istituzioni universitarie italiane, fruibile al fine di agevolare diverse esigenze degli studenti. Tra i servizi forniti figurano: la facoltà di prenotare esami, la visualizzazione delle valutazioni conseguite successivamente al superamento delle prove d'esame, oltre alla possibilità di accettare o respingere tali valutazioni direttamente mediante l'interfaccia telematica. La scelta di questa fonte di dati è fondata sulla sua capacità di consentire l'accesso a tutti gli aspetti delle attività svolte dallo studente in fase di esame, permettendo così l'individuazione degli istanti in cui gli studenti completano specifici esami.

I dati iniziali provenienti dalla piattaforma Esse3 sono stati estratti e resi disponibili in un formato SQL, un linguaggio di interrogazione per database, il dataset iniziale era composto da 55663 righe e 2684 studenti. Prima di essere utilizzati ulteriormente, questi dati sono stati sottoposti a operazioni di pulizia, che hanno coinvolto la rimozione di eventuali informazioni ridondanti.

Successivamente, sono state eseguite operazioni di "*join*" (unione) con altri dataset. Ad esempio, per arricchire il dataset con informazioni come il nome degli esami, è stato eseguito un join utilizzando un codice identificativo presente nei dati. Questo ha consentito di collegare i dati provenienti da fonti diverse e ottenere un dataset più completo e informativo.

Dopo queste manipolazioni, il dataset è stato filtrato per selezionare solo gli studenti laureati. Questa selezione è stata effettuata mediante un ulteriore join con un altro

dataset contenente solo sugli studenti laureati. Questo passaggio ha permesso di focalizzarsi specificamente sugli studenti che hanno completato con successo il loro percorso di studi.

Infine, il dataset risultante è stato esportato in formato CSV, un formato tabellare comunemente utilizzato per l'archiviazione e la condivisione dei dati, per poi importare il dataset in Python, un linguaggio di programmazione, per condurre analisi inerenti al process mining e data mining sui dati.

Il dataset estratto è composto da 410 studenti iscritti nel periodo accademico che si estende dal 2015-2016 al 2019-2020. Va precisato che questo campione è stato frazionato in due distinti segmenti, in considerazione al cambiamento avvenuto nel piano di studi tra gli anni accademici 2016-2017 e 2017-2018. Ne consegue che i risultati delle analisi proposte saranno categorizzati non solo secondo ai sottogruppi di studenti identificati nei paragrafi precedenti, ma anche agli anni accademici di riferimento.

Per ciascuno studente, il dataset si articola come segue: un insieme di caratteri alfanumerici, identificano in maniera univoca lo studente; il corrispondente anno accademico di immatricolazione; il numero di giorni impiegati per conseguire il titolo di laurea; e le valutazioni associate a ciascun esame superato. Inoltre, per ogni singolo studente, sono disponibili tutti gli esami sostenuti e per ciascun esame, cui lo studente si registra, sono presenti ulteriori informazioni correlate alle attività svolte dallo studente. Ciò comprende il momento in cui lo studente ha superato o

non superato l'esame, l'eventualità che lo studente sia stato assente il giorno della prova, e la circostanza in cui lo studente si sia presentato ma abbia deciso di non affrontare la prova stessa. Le suddette categorie di attività sono rispettivamente identificate come “Promosso”, “Bocciato”, “Assente” e “Ritirato”. L'etichetta “Assente” denota il caso in cui lo studente abbia prenotato l'esame senza presentarsi in seguito per sostenere la prova. Invece, l'etichetta “Ritirato” sottintende che lo studente si sia presentato all'esame, ma per varie ragioni abbia scelto di non affrontarlo.

Nella fase di *pre-processing*, sono stati applicati dei cambiamenti al dataset in modo da adeguarlo alle analisi successive.

Innanzitutto, per questo tipo di analisi, dove il focus è il percorso eseguito dagli studenti e il flusso con il quale sostengono gli esami, le attività inerenti agli studenti sono state filtrare considerando solo l’attività “Promosso”.

È importante evidenziare che nel dataset originale le diverse attività erano etichettate con nomi differenti: “Caricato”, “Chiuso”, “Prenotato”, “Verbalizzato” e “Annullato”. Le ultime due attività, “Verbalizzato” e “Annullato”, sono state rimosse a causa della loro scarsa frequenza all'interno del dataset, mentre l'attività “Caricato” è stata trasformata in “Promosso”. Inoltre, dalle attività rimanenti, sono state identificate “Bocciato”, “Assente” e “Ritirato” utilizzando variabili binarie (0 e 1) presenti nel dataset originale. Queste variabili ci hanno consentito di distinguere all'interno dell'attività “Chiuso” coloro che sono stati bocciati, coloro

che non si sono presentati e coloro che si sono ritirati. L'attività “Prenotato” è stata rinominata “Assente”, poiché indica l'azione di prenotare un esame senza poi effettivamente sostenere la prova.

Successivamente all’anno di immatricolazione si è aggiunto per tutti gli studenti lo stesso giorno e mese corrispondente in modo fittizio al 1° ottobre per poter considerare le differenze in giorni tra il giorno di iscrizione e il giorno di superamento dell’esame, questo cambiamento ha così consentito di prendere in considerazione analisi temporali.

Per questa analisi si considera solo studenti laureati, su questa base sono stati tolti dal dataset quegli studenti per i quali non esisteva l’attività “Promosso” per determinati esami, in quanto la laurea prescinde il fatto che tutti gli esami sono stati superati con successo.

Infine, nel dataset sono state aggiunte due variabili: l’anno e il semestre nel quale è stato svolto l’esame in modo tale da poter suddividere i processi in base agli anni e poter ricavare informazioni più dettagliate da analisi statistiche. Ad esempio, considerando quanti esami sono stati dati entro il primo semestre o entro il primo anno.

## 3.3 APPROCCIO

Nel presente paragrafo, verrà delineata l'articolazione metodologica adottata per il raggiungimento degli obiettivi di ricerca e per poter rispondere ai quesiti posti in

questo capitolo. L'approccio complessivo si struttura in diverse fasi di analisi, ciascuna finalizzata a evidenziare aspetti specifici e differenti, seppur collegati da un filo logico all'interno del contesto di studio. Le sezioni seguenti dettagliano le procedure chiave implementate nell'ambito di questa indagine.

### 3.3.1 Estrazione dei Modelli di Processo

Lo scopo principale dell'analisi della carriera degli studenti è quello di scoprire il vero processo che gli studenti seguono durante la triennale di Ingegneria Informatica e dell'Automazione, considerando ogni anno separatamente. Tuttavia, va notato sin da subito che, nonostante la durata triennale, l'analisi si concentrerà principalmente sui primi due anni. Questa scelta è motivata dal fatto che il terzo anno è caratterizzato per lo più da esami a scelta, rendendo il percorso non direttamente confrontabile con un manifesto standard applicabile a tutti gli studenti. Quindi attraverso l'approccio del process mining, si intende esaminare i processi accademici seguiti dagli studenti durante i primi due anni. Tuttavia, ciò non pregiudica la validità e la rilevanza dell'analisi stessa, poiché i primi due anni costituiscono la base strutturale e formativa del percorso.

È importante sottolineare che l'analisi verrà condotta sui tre gruppi distinti di studenti identificati in precedenza. Inoltre, questi processi saranno poi estratti in base all'anno accademico, concentrandosi sui due periodi precedentemente definiti.

Si avrà quindi in output tre processi in base al primo ed al secondo anno che saranno replicati per i due sottogruppi di anni accademici individuati.

Questo ci consentirà di confrontare direttamente i risultati ottenuti dai tre gruppi, evidenziando le differenze e valutando l'impatto del ritardo sulle attività accademiche (**D1**).

In sintesi, l'analisi si concentrerà sulla comprensione dei processi accademici dei primi due anni di un percorso di studio triennale. L'analisi dei due anni iniziali fornisce una visione approfondita e comparabile dei processi seguiti dagli studenti. La replicazione dei risultati per diversi sottogruppi e anni accademici permetterà di trarre conclusioni significative su tali processi.

Per perseguire questo obiettivo si devono applicare tecniche di process discovery, la quale sfida è quella di derivare un modello di processo che rappresenta accuratamente il comportamento osservato da un event log. L'obiettivo è quindi quello di trovare il modello di processo più adatto che descriva l'effettiva esecuzione del processo.

Per estrarre i processi relativi agli esami che gli studenti hanno passato durante il corso triennale in questione è stato utilizzato l'algoritmo IMF, in quanto considerando il percorso degli studenti e la variabilità che caratterizza le scelte che gli studenti possono effettuare nel flusso degli esami che sostengono è importante che l'algoritmo sia in grado di trattare comportamenti poco frequenti o devianti. Questa fase mira a catturare le sequenze di attività intraprese dagli studenti durante

il loro percorso accademico. Attraverso tale approccio, sarà possibile identificare i pattern ricorrenti e le varianti nei percorsi degli studenti, consentendo una visualizzazione comprensibile dei processi.

Inoltre, per poter rispondere alla domanda posta in questa sezione sono state aggiunte al log quattro attività artificiali a scopo temporale, le quali indicano la conclusione di ogni semestre e anno: "*End first semester*", "*End first year"*, "*End third semester*" e "*End second year"*. Queste attività sono state inserite in modo strategico per poter visualizzare chiaramente nell'output prima di quale periodo vengono sostenuti gli esami. Questa disposizione ci consente di avere una panoramica immediata e chiara su quando si svolgono gli esami nel corso dell'anno accademico.

Per questa analisi è stato utilizzato un file log così composto: il case id dato dall'identificativo univoco dello studente, l'attività "Passato" riferito all'esame e la data corrispondente al superamento dell'esame come timestamp.

Il modello ottenuto come risultato dell'analisi è una rete di Petri che rappresenta la sequenza degli esami sostenuti dagli studenti.

Per estrarre i modelli di processo è stato utilizzato il linguaggio di programmazione **Python** e la libreria "*PM4Py*" che è una raccolta di strumenti e algoritmi utilizzati per il PM.

PM4Py è progettata per l'analisi dei processi e la scoperta di modelli a partire dai dati dei log di eventi. Questa libreria offre una serie di strumenti per esplorare,

analizzare e ottimizzare i processi. Con PM4Py, è possibile importare e manipolare i log di eventi, estrarre informazioni chiave dai dati e visualizzare grafici e diagrammi che rappresentano le sequenze di attività, le relazioni temporali e altre dinamiche di processo. La libreria offre anche metodi avanzati di scoperta dei processi che consentono di identificare i modelli sottostanti nei dati dei log di eventi.

Il codice implementato per estrarre i processi è il seguente:

```
tree = pm4py.discovery.discover_process_tree_inductive(log, threshold)

net, initial_marking, final_marking = pm4py.convert_to_petri_net(tree)

parameters = {pn_visualizer.Variants.FREQUENCY.value.Parameters.FORMAT: “png”}

gviz = pn_visualizer.apply(net, initial_marking, final_marking, parameters=parameters, variant=pn_visualizer.Variants.FREQUENCY, log=log)

pt_visualizer.view(gviz)
```

Qui di seguito si spiega brevemente il codice implementato:

1. “tree”:

   Questo passaggio utilizza l'algoritmo Inductive Miner di PM4Py per scoprire un albero di processo a partire da un log di eventi. L'argomento “log” è il log di eventi da cui verrà estratto l'albero, mentre “threshold” rappresenta il parametro di soglia da impostare tra 0 e 1 per filtrare comportamenti infrequenti.

2. “net, initial_marking, final_marking”:

   Questo passaggio converte l'albero di processo scoperto in precedenza in una rete di Petri.

3. “parameters”:

   Qui vengono definiti alcuni parametri per la visualizzazione della rete di Petri. In questo caso, il parametro indica che l'immagine generata sarà in formato PNG.

1. “gviz”:

   Questo passaggio utilizza il modulo `pn_visualizer` di PM4Py per generare una visualizzazione della rete di Petri. La visualizzazione viene creata utilizzando i marcatori iniziali e finali, oltre ai parametri specificati. Successivamente si visualizza l'immagine della rete di Petri generata in questo passaggio tramite “pt_visualizer.view(gviz)”.

In sintesi, il codice esegue l'estrazione di un albero di processo da un log di eventi utilizzando l'algoritmo IM, converte l'albero in una rete di Petri e visualizza questa rete di Petri in un formato immagine PNG.

Inoltre, una delle piattaforme di PM più popolari e comunemente usate è **Disco**, che è un software open source sviluppato dalla società Fluxicon. Disco ha un ambiente user-friendly e semplice, che lo rende comprensibile per la maggior parte degli utenti. Nonostante Disco permetta di estrarre modelli di processo, in realtà la

maggior parte dei ricercatori utilizzano Disco al fine di pre-elaborare, pulire, o filtrare i dati. Un'altra caratteristica popolare di Disco è la capacità di convertire il tipo di dati importati in altri formati come XES e MXML supportati da ProM, che è un'altra popolare piattaforma di PM.

Anche in questo contesto di ricerca il tool Disco è stato utilizzato solo al fine di pre-elaborare e filtrare i dati, infatti il file log dato in input a PM4py è stato creato attraverso questo tool che ha consentito di estrarre il file log in formato XES permettendo di identificare e selezionare le colonne: case id, attività e timestamp, partendo da un dataset in formato CSV.

Il tool Disco è stato inoltre utilizzato anche nella fase di pre-processing per filtrare le attività affinché terminassero con "Promosso". Questo perché, nonostante il dataset estratto da SQL contenesse solo gli studenti laureati, si potrebbero comunque verificare situazioni in cui uno studente ha cambiato università o ha svolto determinati esami durante un periodo di Erasmus. Queste situazioni causano l'assenza di una vera e propria attività all'interno di Esse3 da parte dello studente. Non essendoci quindi un vero riferimento temporale per poter collocare il percorso di questi studenti all'interno di un processo, sono stati eliminati. Nell'analisi è stato utilizzato il filtro "*Endpoints*" in Disco, che permette di scegliere tra le varianti presenti e mantenere solo quelle per le quali l'attività ultima è quella scelta attraverso il filtro.

### 3.3.1 Conformance Checking con i processi estratti

Lo scopo di questa sezione è quello di spiegare qual è il processo ideale che gli studenti dovrebbero seguire e quanto gli studenti sono veramente conformi al manifesto, considerando la differenza tra le tre categorie di studenti “in tempo”, “un anno in ritardo” e “in ritardo” (**D2**). L’obiettivo è anche quello di determinare se l'adesione al manifesto di studio potrebbe aumentare la probabilità che gli studenti si laureeranno in tempo.

Il processo ideale sopra menzionato è il manifesto pubblicato dall'Università Politecnica delle Marche per ogni anno accademico (Fig. I.1, Fig I.2 e Fig. I.3).

Come già spiegato in precedenza il conformance checking viene utilizzato per confrontare un modello di processo con il corrispondente event log. Esso utilizza quindi un modello che rappresenta il processo ideale e i dati provenienti dall'esecuzione reale del processo.

Il processo ideale è l'esecuzione del processo nel modo in cui dovrebbe essere quindi il manifesto dell'anno accademico corrispondente, mentre il processo reale rappresenta il percorso effettivo che gli studenti seguono durante lo stesso anno accademico, e l'obiettivo è quello di stimare con quale grado la realtà si conforma al processo ideale. Il conformance checking consente così di identificare le deviazioni e le variazioni dal percorso ideale.

Per perseguire questo obiettivo, si prende in considerazione la metrica di fitness, con un valore tra 0 e 1, dove una fitness a zero significa che il processo seguito

dagli studenti è molto diverso da quello ideale, mentre una fitness di uno significa che gli esami che gli studenti hanno sostenuto sono esattamente gli stessi indicati nel manifesto.

Per svolgere questa analisi è stato utilizzato il tool **ProM**, un altro strumento di analisi dei processi, open source e ampiamente utilizzato. Creato presso l'Università Tecnologica di Eindhoven (TU/e), ProM offre un ambiente completo per l'estrazione di modelli, l'analisi e la visualizzazione dei dati dei log di eventi. La sua flessibilità consente agli utenti di esplorare e scoprire modelli di processo da dati reali, contribuendo così a migliorare l'efficienza e la comprensione dei flussi di lavoro delle organizzazioni. ProM supporta una varietà di tecniche di PM e offre una vasta gamma di plugin che coprono diverse aree dell'analisi dei processi, dall'estrazione di modelli alla valutazione delle prestazioni. Inoltre, grazie alla sua natura open source, ProM favorisce lo sviluppo e la condivisione di nuove soluzioni e tecniche nel campo del PM.

Per ottenere i risultati si è importato il manifesto in ProM come una rete di Petri ed il file XES attraverso l'import plugin "*Naive*", successivamente dopo aver selezionato sia il log che il modello, è stato utilizzato il plugin "*Replay a log on Petri net for Conformance Analysis*". Nella fase successiva, si è collegato le etichette degli eventi nel log alle etichette delle transizioni nel modello. Per impostazione predefinita, ProM suggerisce di utilizzare il "*MXML Legacy*

*Classifier*" ma per questo log è stato selezionato "*Activity*" come classificatore, ovvero gli esami svolti dagli studenti.

Dopo aver impostato il modello con una marcatura iniziale e finale e la mappatura tra il registro e il modello, si è ottenuto in output una rete di Petri con colori indicanti le deviazioni, ma quello che è stato preso in considerazione in questa analisi è la metrica di fitness.

Questo procedimento è stato ripetuto per ogni file log suddiviso in base agli anni accademici e ai sottogruppi di studenti identificati, utilizzando il manifesto dell'anno accademico di riferimento. Questo ha permesso di confrontare le fitness tra i vari modelli in modo tale da fornire delle conclusioni in merito alle differenze che potrebbero esistere tra gli studenti che si laureano in tempo e quelli che si laureano in ritardo.

### 3.3.2 Correlazioni tra le variabili

In questa sezione l'obiettivo è capire come il tempo di laurea, inteso come riuscire o meno a laurearsi in tempo sia influenzato dal numero di esami che gli studenti riescono a sostenere entro il primo semestre o entro il primo anno e dal voto medio degli studenti (**D3**).

Per poter raggiungere questo obiettivo è stato implementato un modello di regressione logistica con la libreria "*statsmodel*" nel linguaggio di programmazione Python.

Come prima cosa è stata definita una variabile binaria per indicare se lo studente si è laureato in tempo o meno, e questa è stata usata come variabile dipendente della regressione logistica. La variabile è stata generata assegnando il valore 1 fino ai giorni che rappresentano un periodo di tre anni e sei mesi, mentre i giorni che superano questa soglia sono stati contrassegnati con il valore 0, in base all'indicatore iC02 menzionato nei paragrafi precedenti.

Creando un subset del dataset per il primo semestre e un altro per il primo anno si è creato una colonna contenete per ogni studente dei 410 un conteggio di quanti esami hanno sostenuto entro il primo semestre ed entro il primo anno.

Inoltre, si è creata un'ulteriore colonna considerando la media dei voti di ogni studente per capire l'influenza che ha il voto finale sul laurearsi in tempo o meno.

Quindi come variabili dipendenti sono stati utilizzati il numero di esami passati nel primo semestre e il voto medio per una regressione logistica, e per l'altra il numero di esami passati entro il primo anno e il voto medio. Inoltre, le variabili sono state standardizzate, utilizzando la funzione "StandardScaler" della libreria "*scikit-learn*" di Python. Utilizzare StandardScaler aiuta a garantire che le variabili abbiano una media zero e una deviazione standard unitaria, rendendo più facile l'interpretazione e l'addestramento dei modelli.

Per capire se i coefficienti delle variabili indipendenti nella regressione logistica sono statisticamente significativi, si è osservato il valore p (p-value). Questo valore rappresenta la probabilità che l'effetto osservato sia dovuto al caso anziché a una

vera relazione. Valori p bassi indicano che è meno probabile che l'effetto sia casuale. Se il valore p è inferiore al livello di significatività ($p < 0.05$), si può affermare che esiste una relazione statisticamente significativa tra la variabile indipendente e la variabile dipendente. Mentre se il valore p è maggiore del livello di significatività ($p > 0.05$), non ci sono prove sufficienti per affermare che esiste una relazione statisticamente significativa.

Inoltre, si può anche considerare l'intervallo di confidenza del coefficiente: se l'intervallo di confidenza non include zero, è un altro segno di significatività.

La relazione verrà quindi interpretata considerando il coefficiente stesso: se il coefficiente è positivo o negativo e se il valore p associato è basso o alto.

### 3.3.3 Varianti con il minor tempo di laurea

Lo scopo finale di tutta l'analisi è quello di migliorare il tasso di successo degli studenti fornendo loro un'indicazione di quale potrebbe essere la successione migliore di esami per laurearsi nel minor tempo possibile (**D6**).

Per perseguire questo obiettivo si analizzerà il "percorso vincente", osservando la sequenza di esami sostenuti dagli studenti che riescono a laurearsi nel più breve tempo possibile. Si prende in considerazione solo gli esami dati durante il primo anno perché si ritiene che sia il momento più critico per la carriera dello studente.

Partendo dai processi scoperti durante le analisi precedenti riguardanti il percorso seguito dagli studenti durante il primo anno, ci si concentra sulla *process variant analysis* (analisi delle varianti[9] di processo).

L' analisi delle varianti di processo è una tecnica nell'ambito del PM che ha come obiettivo quello di capire quante varianti un processo ha e come differiscono. Questa tecnica si concentra quindi sull'identificazione e l'analisi delle diverse varianti o percorsi che possono emergere all'interno di un processo.

Inoltre, questo tipo di analisi è spesso utilizzata per esaminare come le attività, le decisioni e gli eventi si sviluppano in modo diverso all'interno di un processo, in base a diversi scenari o circostanze: può fornire una comprensione più approfondita delle dinamiche del processo.

Attraverso l'analisi delle varianti di processo, è possibile scoprire quali percorsi sono i più frequenti, quali possono portare a ritardi e come le diverse varianti possono influenzare le prestazioni e l'efficienza del processo, può quindi essere utile per l'ottimizzazione dei processi, l'individuazione di aree problematiche, l'identificazione di opportunità di miglioramento e la comprensione delle dinamiche complesse all'interno di un'organizzazione.

[9] Una variante è un insieme di casi che condividono le stesse attività nello stesso ordine, che può essere distinto da altri sulla base di alcune caratteristiche.

Al fine di condurre questa analisi, è stata impiegata la libreria PM4py di Python. Tramite la funzione "*get_variants*", è stato possibile estrarre le diverse varianti da un file log fornito in input. Successivamente, queste varianti sono state disposte in ordine crescente secondo il tempo di laurea, consentendo così la visualizzazione del percorso seguito dagli studenti che hanno conseguito la laurea nel minor tempo possibile.

### 3.3.4 Predizione del tempo di laurea

In ultimo, in questa analisi ci si è chiesti quali sono i principali fattori (esami) che influenzano le prestazioni accademiche degli studenti e se il tempo che gli studenti impiegano per laurearsi può essere previsto (**D5**).

Per rispondere a questa domanda è stato utilizzato il seguente dataset:

Il dataset è composto da una variabile binaria 0 e1 che indica se lo studente si è laureato in tempo o meno e tutti gli esami del primo anno rappresentati nelle colonne, mentre nelle righe del dataset ci sono tutti gli studenti (410). In dettaglio, per ciascuno studente, i valori corrispondono a una variabile "timedelta", la quale indica, per ogni esame superato dagli studenti il numero di giorni impiegati, rappresentato dall'intervallo temporale in giorni tra la prima attività registrata e l'attività finale "Promosso". Nel caso in cui un esame non sia stato superato entro il primo anno, è stato attribuito un valore di 1000 che indichi questa situazione.

Per effettuare previsioni sui tempi di laurea degli studenti, sono stati utilizzati due algoritmi di machine learning distinti: la regressione logistica e il Support Vector Machine (SVM).

La regressione logistica è stata impiegata per modellare la relazione tra le variabili di input, rappresentate dai tempi di superamento degli esami, e la variabile di output binaria, che indica se lo studente si è laureato in tempo o meno. Questo modello è stato utilizzato per stimare la probabilità di successo, ovvero se lo studente si laurea in tempo in base ai dati di input.

Il SVM è un altro algoritmo utilizzato per la classificazione. Ha lo scopo di trovare un iperpiano ottimale che separa i dati in classi distinte, in questo caso, studenti laureati in tempo e studenti che non lo sono. L'SVM cerca di massimizzare il margine tra le due classi, permettendo così una migliore generalizzazione delle previsioni.

In sintesi, l'obiettivo è quello di prevedere se uno studente si laureerà in tempo o meno in base ai tempi di superamento degli esami. Questi modelli aiutano a comprendere meglio quali fattori influenzino il successo degli studenti nel completare il loro corso di laurea entro il primo anno.

Per valutare l'efficacia dei modelli di regressione logistica e SVM, è stata utilizzata l'F1-score come metrica di valutazione. Questa metrica è fondamentale quando si tratta di problemi di classificazione, poiché fornisce una valutazione completa delle prestazioni di un modello.

L'F1-score è utilizzata per valutare la *precision* e il *recall*.

La *precision* misura quanto è preciso o accurato il modello nella predizione dei positivi. Si calcola come il rapporto tra i veri positivi (le istanze correttamente classificate come positive) e la somma dei veri positivi e dei falsi positivi (le istanze erroneamente classificate come positive).

Una precision elevata indica che il modello ha una bassa probabilità di fare falsi positivi, ossia di classificare erroneamente un'istanza negativa come positiva. In altre parole, il modello è altamente preciso nel rilevare le vere istanze positive.

Il *recall* noto anche come sensitività o ture positive rate rappresenta la frazione di istanze positive predette correttamente dal modello rispetto al totale delle vere istanze positive presenti nei dati. Si calcola come il rapporto tra i veri positivi e la somma dei veri positivi e dei falsi negativi (le istanze positive erroneamente classificate come negative). In altre parole, indica quanto bene il modello riesce a identificare le istanze positive effettive rispetto a quelle che sono effettivamente positive. Un recall elevato indica che il modello riesce a individuare bene le vere istanze positive, evitando falsi negativi, ossia classificazioni erroneamente negative.

In breve, la precision si concentra sulla precisione delle previsioni positive, mentre il recall misura la capacità del modello di identificare correttamente le istanze positive presenti nei dati. Entrambe queste metriche sono importanti e vengono

spesso considerate insieme per ottenere una valutazione completa delle prestazioni di un modello di classificazione.

L'F1-score è una metrica che combina precision e recall in un unico valore. Viene calcolato come la media armonica tra questi due parametri.

Un valore elevato di F1-score indica che il modello ha una buona capacità sia di prevedere accuratamente le istanze positive che di minimizzare i falsi positivi e i falsi negativi. In altre parole, il modello è bilanciato e ha una solida capacità di classificazione.

Un valore basso dell'F1-score suggerisce che il modello ha difficoltà a trovare un equilibrio tra precision e recall, il che potrebbe essere dovuto a vari problemi, come la presenza di falsi positivi o falsi negativi significativi. D'altra parte, un valore alto dell'F1-score indica che il modello è efficace nel combinare precisione e recall, dimostrando una capacità di classificazione robusta.

In sintesi, l'F1-score è una metrica cruciale per valutare la qualità delle previsioni di un modello di classificazione e per capire quanto sia bilanciato tra la precisione delle previsioni positive e la capacità di identificare correttamente le istanze positive.

Inoltre, per valutare l'efficacia dei modelli di regressione logistica e SVM, è stata condotta anche un'analisi della curva ROC (Receiver Operating Characteristic).

La curva ROC è un grafico che viene utilizzato per valutare le prestazioni di un modello di classificazione binaria in base al suo trade-off tra tasso di veri positivi

(True Positive Rate, TPR) e tasso di falsi positivi (False Positive Rate, FPR). La curva ROC è una parte essenziale dell'analisi delle prestazioni dei modelli di classificazione ed è spesso utilizzata insieme all'area sotto la curva ROC (AUC-ROC) per valutare la bontà di un modello.

Di seguito si spiega come interpretare la curva ROC:

L'asse x (FPR - False Positive Rate) rappresenta il tasso di falsi positivi, ovvero la percentuale di casi negativi che vengono erroneamente classificati come positivi dal modello. L'asse y (TPR - True Positive Rate o Sensibilità), invece rappresenta il tasso di veri positivi, ovvero la percentuale di casi positivi che vengono correttamente classificati come positivi dal modello.

La linea diagonale tratteggiata rappresenta il caso in cui il modello effettua una classificazione casuale. Questo significa che il modello non sta facendo alcuna previsione significativa.

La curva ROC del modello mostra come varia il TPR rispetto al FPR quando si modificano i threshold di classificazione. Il punto in alto a sinistra della curva ROC rappresenta il caso ideale in cui il modello ha un alto TPR e un basso FPR.

Inoltre, l'AUC-ROC è l'area sotto la curva ROC ed è una misura della capacità discriminante complessiva del modello. Un modello con un AUC-ROC maggiore è migliore nel distinguere tra le classi.

Alcune considerazioni:

- Un modello con una curva ROC che si avvicina alla parte superiore sinistra è migliore perché ha una migliore capacità di discriminazione tra le classi.
- Se la curva ROC è una linea retta diagonale, il modello non è in grado di fare previsioni migliori di un lancio di moneta casuale.
- L'AUC-ROC varia da 0 a 1, dove un valore di 0,5 indica una previsione casuale e un valore di 1 indica una previsione perfetta.

In sintesi, la curva ROC e l'AUC-ROC forniscono una valutazione visiva e quantitativa delle prestazioni del tuo modello di classificazione binaria. Un modello con una curva ROC vicina all'angolo superiore sinistro e un'alta AUC-ROC è desiderabile.

Per sviluppare queste analisi è stata utilizzata la libreria "*scikit-learn*" di Python.

# 4. RISULTATI

Nella sezione seguente, si presentano in dettaglio i risultati dello studio sulla durata degli anni accademici e le variabili correlate. La sezione sarà suddivisa come segue: inizialmente si esaminano i processi relativi agli anni accademici 2015-2017 e i processi degli anni accademici successivi, comprendendo il periodo 2017-2020. Inoltre, si confronta i risultati con il manifesto per valutare la conformità tra il processo reale e quello ideale.

Successivamente, si esaminano le correlazioni tra le diverse variabili prese in considerazione, analizzando come queste interagiscano tra loro e come influenzino la durata del percorso di laurea degli studenti.

Si presentano poi le varianti estratte dal file log così da poter esaminare i profili degli studenti che hanno completato il percorso accademico in tempi più brevi e identificare informazioni che potrebbero aver favorito questi risultati. In modo da fornire agli studenti una linea guida.

Infine, si discute i risultati dell'analisi predittiva del tempo di laurea.

## 4.1 PROCESSI ANNI ACCADEMICI 2015-2017

In questa sezione si analizzano i processi estratti per gli anni accademici 2015-2016 e 2016-2017 per le categorie di studenti: laureati in tempo, un anno in ritardo e in ritardo. I risultati rappresentano separatamente il primo e il secondo anno.

Per ogni processo estratto è stato definito un threshold che filtra i comportamenti infrequenti nel processo in output, in base al miglior compromesso tra *Fitness* e *Precision*.

Il primo processo (Fig. IV.1) si riferisce agli studenti classificati come “in tempo” del primo anno: in generale, come era logico aspettarsi, sono loro che dovrebbero seguire più fedelmente il manifesto.

All'inizio del processo ci sono tutte le attività parallele che possono anche essere saltate, tranne “End First Semester” in quanto è un’attività che viene svolta da tutti per definizione.

Si nota un'alta frequenza di studenti che passano “Analisi Matematica I” e “Algebra Lineare e Geometria” I, mentre meno studenti riescono a passare “Fisica generale I”.

Un comportamento simile si nota nel secondo semestre, l’esame che viene fatto da meno studenti è “Fisica Generale II”, in quanto più della metà degli studenti saltano l'esame, come ci si potrebbe aspettare considerando che anche al primo semestre “Fisica Generale I” è l’esame sostenuto da meno studenti.

Mentre, un’alta frequenza di studenti passano gli esami “Fondamenti di Informatica” ed “Economia dell’Impresa”.

Per questo processo è stato settato un threshold di 0.3 con una fitness di 0.99 e una precsion di 0.76.

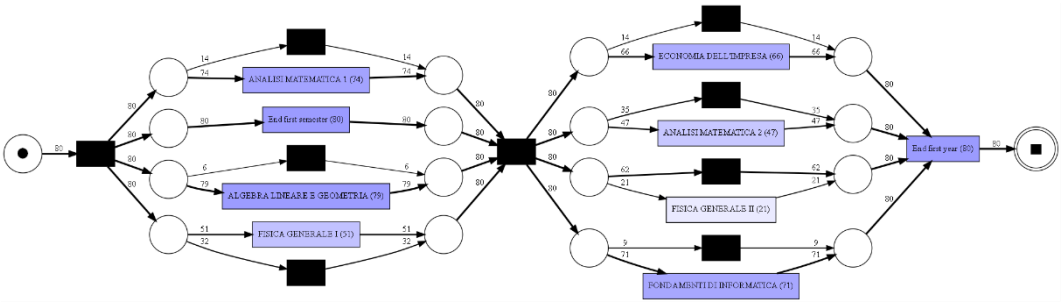


*Fig. IV.1 Percorso degli studenti "in tempo" – primo anno*

Osservando la figura (Fig. IV.2) che rappresenta gli studenti classificati come "un anno in ritardo" si nota che il primo anno meno della metà degli studenti riesce a passare "Analisi Matematica I" e "Fisica Generale I" mentre quasi tutti passano "Algebra Lineare e Geometria".

Di conseguenza, solo pochi studenti nel secondo semestre riescono a passare "Analisi Matematica II" mentre nessuno prova a sostenere l'esame di "Fisica Generale II". Si nota infatti che questo esame non è presente nemmeno come attività nel processo.

Quindi gli esami che vengono passati maggiormente durante il secondo semestre dagli studenti in questo caso sono "Fondamenti di Informatica" ed "Economia dell'Imprese".

Inoltre, tutte le attività del primo semestre sono da svolgere in parallelo, tutte le attività possono anche essere saltate, tranne "End First Semester" in quanto è un'attività che viene svolta da tutti per definizione.

Per quanto riguarda il secondo semestre invece l'attività "Analisi Matematica II" può essere saltata mentre le attività "Economia dell'Impresa" e "Fondamenti di Informatica" sono svolte in parallelo e nessun studente le ha saltate; quindi, probabilmente sono gli unici due esami del secondo semestre che tutti riescono a passare.

Per questo processo è stato settato un threshold di 0.4 con una fitness di 0.95 e una precsion di 0.91.

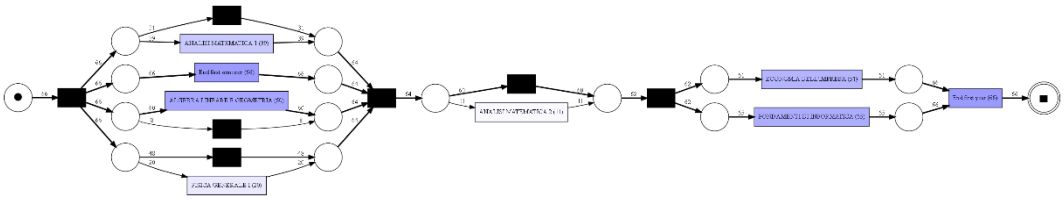

*Fig. IV.2 Percorso degli studenti "un anno in ritardo" – primo anno*

Ora si sottolinea alcune differenze del primo anno tra gli studenti che si laureano in tempo e chi si laurea in ritardo.

La prima cosa che si nota come differenza tra il processo degli studenti in tempo e quello relativo agli studenti in ritardo (Fig. IV.3) è che nel secondo caso "Fisica Generale II" non appare, il che significa che nessuno degli studenti ha sostenuto questo esame, il quale è obbligatorio nel secondo semestre, similmente agli studenti che si sono laureati con un anno di ritardo.

Inoltre, la maggior parte degli studenti passa "Algebra Lineare e Geometria", ma molti di loro non passano "Fisica Generale I" e "Analisi Matematica I" nel primo semestre.

Nel secondo semestre l'esame "Analisi Matematica II" è sostenuto da pochissimi studenti, mentre "Fondamenti di Informatica" e "Economia dell'Imprese" sono passati almeno da più della metà degli studenti.

Inoltre, è interessante notare a differenza dei processi precedenti che alcuni studenti saltano completamente il primo ed il secondo semestre, in quanto il processo rappresenta la possibilità di saltare tutte le attività.

Per questo processo è stato settato un threshold di 0.1 con una fitness di 0.99 e una precsion di 0.86.

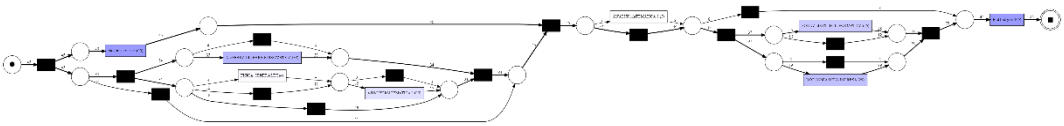

*Fig. IV.3 Percorso degli studenti "in ritardo" – primo anno*

Di seguito si analizzano i processi relativi al secondo anno per le stesse categorie di studenti identificate sopra.

Nel secondo anno, il manifesto contiene anche esami non obbligatori, lo studente può scegliere di sostenere un esame tra questi quattro esami: "Calcolo delle Probabilità e Statistica matematica", "Meccanica Razionale" e "Algebra e Logica" e "Analisi Numerica".

Il processo degli studenti che si sono laureati in tempo (Fig. IV.4) mostra che gli studenti all'inizio cercano di recuperare gli esami che non sono riusciti a sostenere al primo anno, in particolare tra questi gli esami con più studenti sono "Fisica

Generale I" e "Fisica Generale II", poiché probabilmente gli studenti trovano più difficili questi corsi rispetto agli altri del primo anno.

Inoltre, più della metà degli studenti riesce a sostenere gli esami del secondo anno al primo semestre, compresi gli eventuali esami a scelta. Stessa situazione si presenta per il secondo semestre, unica eccezione per l'esame "Elettromagnetismo per la Trasmissione dell'Informazione" che lo passano meno della metà degli studenti.

Per questo processo è stato settato un threshold di 0.3 con una fitness di 0.91 e una precsion di 0.82.

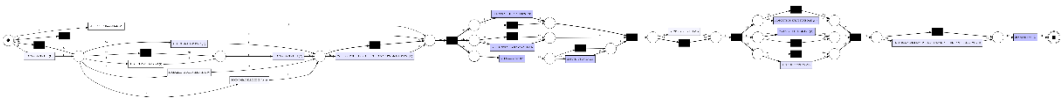

*Fig. IV.4 Percorso degli studenti "in tempo" – secondo anno*

Osservando il processo degli studenti laureati con un anno di ritardo (Fig. IV.5) si nota che all'inizio sono presenti gli esami del primo anno che gli studenti devono ancora sostenere. A differenza del processo precedente meno studenti passano gli esami "Fisica Generale I" e "Fisica Generale II", probabilmente perché alcuni lo fanno dopo il secondo anno.

Per quanto riguarda gli esami del primo semestre del secondo anno meno della metà degli studenti passano l'esame "Fondamenti di Automatica", a differenza degli esami "Elettrotecnica" ed "Elementi di Elettronica" che sono stati superati dalla maggior parte degli studenti. Una situazione simile si presenta nel secondo semestre

dove pochi studenti passano “Algoritmi e Strutture Dati” e “Controlli Automatici”, e quasi nessuno passa “Elettromagnetismo per la Trasmissione dell’Informazione”. Inoltre, è interessante notare che non tutti gli esami a scelta sono presenti, ad esempio “Meccanica Razionale” non è presente come attività nel processo mentre gli altri esami a scelta sono presenti ma li passano pochi studenti prima della fine del secondo anno. Questo perché probabilmente nel secondo anno gli studenti sono concentrati nel recuperare gli esami del primo anno e cercare di finire anche quelli del secondo anno obbligatori.

Questo suggerisce che non completare il primo anno senza accumulare un numero significativo di esami arretrati, specialmente quelli considerati più difficili dagli studenti, potrebbe avere ripercussioni nel secondo anno. Ciò potrebbe portare a un accumulo di ulteriori esami nel secondo anno e così via, aumentando il rischio di ritardare la laurea.

Per questo processo è stato settato un threshold di 0.2 con una fitness di 0.93 e una precsion di 0.72.

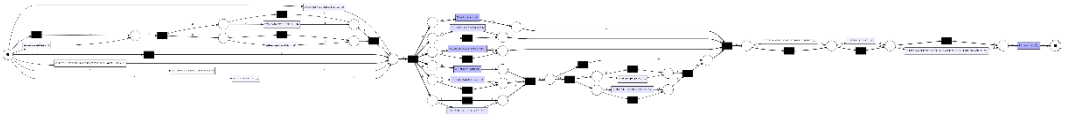

*Fig. IV.5 Percorso degli studenti “un anno in ritardo” – secondo anno*

Per quanto riguarda il processo degli studenti laureati in ritardo (Fig. IV.6), le differenze con il processo degli studenti laureati in tempo sono meno evidenti: in entrambi i processi ci sono ancora esami che gli studenti non sono riusciti a passare

il primo anno. Inoltre, gli studenti che si laureano in tempo mostrano una tendenza a sostenere più esami obbligatori del secondo anno, oltre a finire quelli del primo.

Nel processo degli studenti che si laureano in tempo, si può notare che sono presenti tutti gli esami a scelta mentre nel processo degli studenti in ritardo solo gli esami "Calcolo delle Probabilità e Statistica Matematica" sono stati scelti.

Come descritto sopra, gli studenti che si laureano in tempo sembrano sostenere più esami obbligatori del secondo anno rispetto a quelli che si laureano in ritardo: infatti, nel processo degli studenti che si laureano in ritardo l'esame "Elettromagnetismo per la Trasmissione dell'Informazione" non è presente come attività nel processo, questo significa che nessuno nel secondo anno per questa categoria di studenti riesce a sostenere questo esame. Si deve comunque considerare che nel processo degli studenti classificati come "in tempo" meno della metà riesce a passare questo esame.

Inoltre, nel processo degli studenti laureati in ritardo si nota che ci sono pochissimi studenti che riescono a superare gli esami "Fondamenti di Automatica", "Controlli Automatici" e "Algoritmi e Strutture Dati". D'altra parte, per gli esami "Elementi di Elettronica" ed "Elettrotecnica" si nota che riescono a passarlo circa la metà degli studenti.

In questo processo, all'inizio sono ancora presenti esami del primo anno e oltre la metà degli studenti incontra difficoltà nel superare gli esami obbligatori del secondo anno. In generale, questa situazione riflette la stessa problematica riscontrata nel

processo precedente: gli studenti appartenenti alle categorie “un anno in ritardo” e “in ritardo” continuano a incontrare difficoltà nel superare tutti gli esami del secondo anno, a differenza degli studenti "in tempo" per i quali tutte le attività sono pianificate nel processo e più della metà di essi riesce a superare con successo tutti gli esami, ad eccezione dell'esame “Elettromagnetismo per la Trasmissione dell'Informazione”.

Per questo processo è stato settato un threshold di 0.3 con una fitness di 0.94 e una precsion di 0.76.

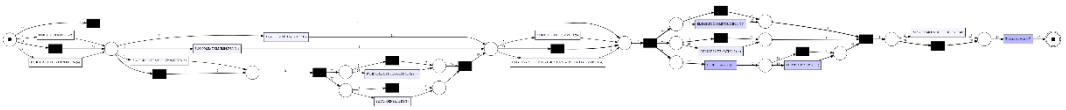

*Fig. IV.6 Percorso degli studenti “in ritardo” – secondo anno*

Di seguito si valuta la conformità dei processi con il manifesto:

| *Anno* | *Fitness studenti “in tempo”* | *Fitness studenti “in ritardo”* |
|---|---|---|
| *Primo anno* | **0.74** | **0.59** |
| *Secondo anno* | **0.70** | **0.44** |

*Tab. IV.1 Indicatori studenti iscritti negli anni accademici 2015-2016 e 2016-2017*

Per valutare quanto gli studenti sono conformi al manifesto, il processo ideale e il file di log sui processi sopra descritti sono stati confrontati. Per perseguire questo obiettivo, si prende in considerazione una metrica di fitness, con un valore tra zero e uno, dove una idoneità a zero significa che il processo seguito dagli studenti è

molto diverso da quello ideale e una idoneità di uno significa che gli esami che gli studenti hanno sostenuto sono esattamente gli stessi indicati nel manifesto.

Nella Tabella (Tab. IV.1) vengono presentati i risultati relativi alla conformità degli studenti iscritti negli anni accademici 2015-2016 e 2016-2017 riguardo al conseguimento della laurea nei tempi previsti, in confronto agli studenti che si laureano in ritardo. Questa analisi riguarda sia il primo che il secondo anno di laurea.

Un'osservazione iniziale rivela una notevole differenza tra gli studenti che rispettano il piano di studi previsto e quelli che invece accumulano ritardi. Nel primo anno, gli studenti che completano il percorso di studi secondo le linee guida presentate nel manifesto mostrano un livello di conformità del 74%. D'altra parte, gli studenti che si laureano in ritardo dimostrano una minore aderenza al manifesto, registrando una percentuale di conformità del 59%.

La tendenza si ripete anche nel secondo anno: gli studenti che non riescono a rispettare il piano di studi sono meno in linea con il manifesto rispetto a coloro che riescono a farlo entro i tempi prestabiliti. Mentre i primi presentano una percentuale di conformità del 44%, i secondi mantengono una percentuale di conformità del 70%. Questi dati indicano che la conformità al manifesto potrebbe svolgere un ruolo cruciale nella possibilità per gli studenti di conseguire la laurea nei tempi previsti.

In base a questi risultati si può supporre che la conformità al manifesto influisca sulla possibilità per gli studenti di laurearsi in tempo o meno.

Pertanto, potrebbe essere fondamentale esaminare attentamente il manifesto per identificare eventuali colli di bottiglia che potrebbero compromettere la conformità degli studenti. Risolvere tali problematiche potrebbe risultare essenziale per garantire che gli studenti abbiano una maggiore probabilità di conseguire la laurea entro i tempi stabiliti e ridurre il numero di laureati in ritardo.

## 4.2 PROCESSI ANNI ACCADEMICI 2017-2020

In questa sezione si analizzano i processi estratti per gli anni accademici 2017-2018, 2018-2019 e 2019-2020 per le categorie di studenti: laureati in tempo, un anno in ritardo e in ritardo. I risultati rappresentano separatamente il primo e il secondo anno.

In modo analogo al precedente paragrafo per ogni processo estratto è stato definito un threshold che filtra i comportamenti infrequenti nel processo in output, in base al miglior compromesso tra "*Fitness"* e "*Precision"*.

Il primo processo (Fig. IV.7) riguarda gli esami sostenuti dai laureati in tempo durante il loro primo anno di studio. Come descritto nel paragrafo precedente per questa categoria di studenti ci si aspetta una fedeltà maggiore al manifesto.

All'inizio del processo ci sono tutte le attività parallele che possono anche essere saltate, tranne "End First Semester" in quanto è un'attività che viene svolta da tutti per definizione.

Si nota un'alta frequenza di studenti che passano gli esami del primo semestre: gli studenti che saltano gli esami “Analisi Matematica I” e “Algebra Lineare e Geometria” sono pochi, si nota una frequenza maggiore per l’esame “Fisica generale I”, ma comunque circa due terzi degli studenti riesce a sostenere l’esame.
Un comportamento simile si nota nel secondo semestre, l’esame che viene fatto da meno studenti è “Fisica Generale II”, e solo un terzo degli studenti salta l’esame di “Analisi Matematica I”. Mentre quasi tutti gli studenti passano gli esami “Fondamenti di Informatica” ed “Economia dell’Impresa”.
In questo contesto, si può affermare che gli studenti mostrano una notevole aderenza al manifesto, poiché la maggior parte di loro supera gli esami, ad eccezione di alcuni esami specifici, che sembrano rappresentare una sfida comune: “Fisica Generale I”, “Fisica Generale II” e “Analisi Matematica II”. È comunque degno di nota che un buon numero di studenti riesce a superare questi esami, anche se “Fisica Generale II” è l'esame meno frequentemente superato.
Per questo processo è stato settato un threshold di 0.4 con fitness 0.99 e precsion 0.81.

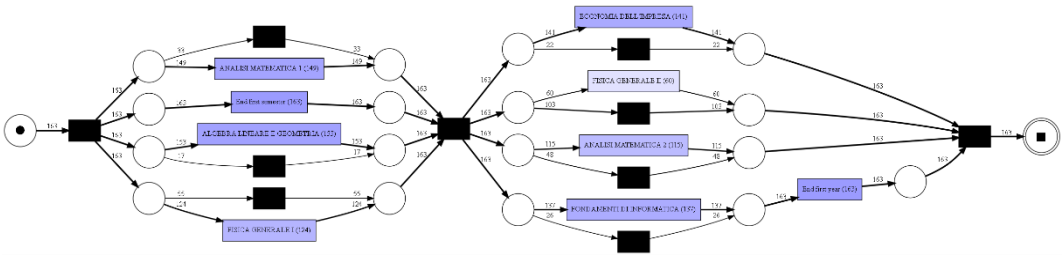


*Fig. IV.7 Percorso degli studenti “in tempo” – primo anno*

Il processo che rappresenta il flusso degli esami sostenuto dagli studenti che si sono laureati con un anno di ritardo (Fig. IV.8) è molto simile al processo precedente.

Le attività sono da svolgere il parallelo e possono essere saltate, tranne l'attività "End First Semester" che per definizione è svolta da tutti gli studenti.

Quello che è interessante notare in questo processo è che nel primo semestre solo un numero limitato di studenti supera "Fisica Generale I" e solamente uno studente passa nel secondo semestre "Fisica Genarle II". Questo esame è quindi quello che rappresenta maggiormente una sfida per gli studenti.

Inoltre, va notato che anche gli esami di "Analisi Matematica I" e "Analisi Matematica II" risultano essere delle sfide per gli studenti, dato che solo la metà di loro riesce a superarli con successo.

Mentre Gli esami con la frequenza più elevata di studenti che li superano sono: "Algebra Lineare e Geometria", "Fondamenti di Informatica" ed "Economia dell'Impresa".

Per questo processo è stato settato un threshold di 0.3 con fitness 0.97 e precsion 0.78.

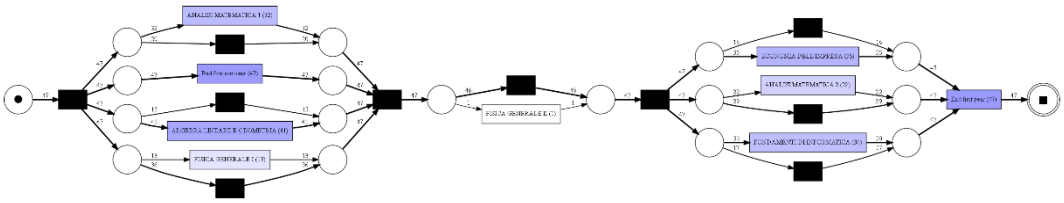

*Fig. IV.8 Percorso degli studenti "un anno in ritardo" – primo anno*

Ora si evidenzia le differenze dei processi del primo anno tra gli studenti che si laureano in tempo e chi si laurea in ritardo.

È importante notare innanzitutto che questo processo coinvolge solo sette studenti. Questa limitazione deriva dal fatto che il dataset copre i dati fino al 2023. Di conseguenza, non sono disponibili dati per gli studenti immatricolati dopo l'anno accademico 2018-2019 che rientrano nella categoria di laureati "in ritardo" secondo gli indicatori definiti nei capitoli introduttivi.

Si può comunque osservare il comportamento dei pochi studenti che si sono laureati in ritardo (Fig. IV.9) e notare come prima differenza dal processo di chi si laurea in tempo che alcuni di loro concludono il primo semestre senza aver sostenuto nemmeno un esame, mentre una piccola parte supera solo gli esami "Fisica Generale I" e "Analisi matematica I". Nel secondo semestre si nota l'esame "Algebra Lineare e Geometria" che probabilmente provano a recuperare dal primo semestre e passano gli esami "Fondamenti di Informatica" ed "Economia dell'Impresa".

Un'altra differenza che si nota è che nel caso degli studenti che si laureano in ritardo "Fisica Generale II" non appare nemmeno come attività nel processo, questo significa che nessuno degli studenti ha sostenuto questo esame obbligatorio nel secondo semestre, similmente agli studenti che si sono laureati con un anno di ritardo, in quanto solo uno di loro è riuscito a passare l'esame. In aggiunta, durante

il secondo semestre, l'esame “Analisi Matematica II” è stato superato da un solo studente.

Per questo processo è stato settato un threshold di 0.4 con fitness 0.97 e precsion 0.57.

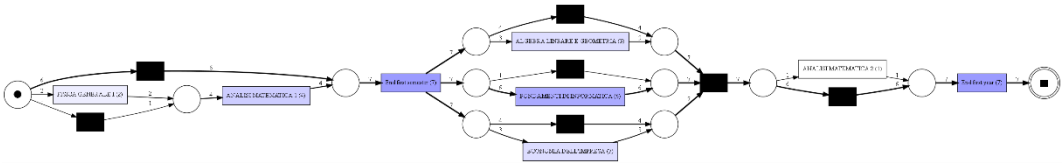

*Fig. IV.9 Percorso degli studenti “in ritardo” – primo anno*

Di seguito si analizzano i processi relativi al secondo anno per le stesse categorie di studenti identificate sopra.

Nel secondo anno, il manifesto contiene anche esami non obbligatori, lo studente può scegliere di sostenere un esame tra questi quattro esami: “Meccanica Razionale”, “Calcolo delle Probabilità e Statistica matematica” e “Algebra e Logica” e “Analisi Numerica”.

Il processo degli studenti che si sono laureati in tempo (Fig. IV.10) mostra che gli studenti all’inizio cercano di recuperare gli esami che non sono riusciti a sostenere durante il primo anno o scelgono di dare come primi esami quelli a scelta del secondo anno.

Per quanto riguarda il primo semestre del secondo anno più della metà degli studenti riesce a passare gli esami “Fondamenti di Automatica” ed “Elettrotecnica”, con una frequenza di poco minore passano “Elementi di elettronica”. Situazione simile per

il secondo semestre, l'esame "Controlli Automatici" viene superato con successo da quasi tutti gli studenti. Inoltre, è interessante notare che l'esame "Elettromagnetismo per la Trasmissione dell'Informazione", come evidenziato nei processi precedenti degli anni accademici dal 2015 al 2017, era superato da meno della metà degli studenti. Tuttavia, in questo specifico processo, almeno la metà degli studenti riesce a superarlo con successo. In generale, non emergono esami con una frequenza significativa di studenti che li saltano o non li superano.

Per questo processo è stato settato un threshold di 0.3 con fitness 0.96 e precsion 0.90.

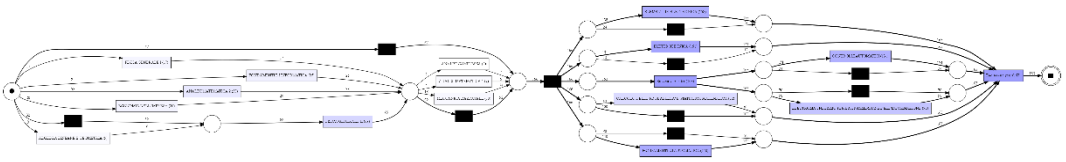

*Fig. IV.10 Percorso degli studenti "in tempo" – secondo anno*

Anche nel processo degli studenti laureati un anno in ritardo (Fig. IV.11) sono presenti gli esami del primo anno che gli studenti devono ancora sostenere.

Per quanto riguarda gli esami del primo semestre del secondo anno meno della metà degli studenti passa l'esame "Fondamenti di Automatica", a differenza degli esami "Elettrotecnica" ed "Elementi di Elettronica" che sono stati superati con successo dalla maggior parte degli studenti. Una situazione simile si presenta nel secondo semestre dove pochi studenti superano gli esami "Controlli Automatici" ed "Elettromagnetismo per la Trasmissione dell'Informazione".

Inoltre, è interessante notare che non tutti gli esami a scelta sono inclusi nel processo; ad esempio, “Analisi numerica” non compare tra le attività, mentre per gli altri esami a scelta, che solo pochi studenti riescono a superarli entro la fine del secondo anno.

In questo caso “End Thrid Semester” si trova in parallelo all’inizio del processo, non separando quindi effettivamente gli esami tra i due semestri, questo significa che l’algoritmo non ha individuato relazioni d’ordine abbastanza forti. Infatti, esaminando i casi registrati su Disco, è evidente che gli studenti non seguono rigorosamente il manifesto e che c'è una notevole variabilità nell’ordine in cui scelgono di sostenere gli esami.

Probabilmente gli studenti si concentrano nel cercare di recuperare gli esami del primo anno e superare almeno gli esami obbligatori del primo semestre del secondo anno, di conseguenza in pochi riescono a passare anche gli esami obbligatori del secondo semestre. Come si è visto, questa categoria di studenti ha la tendenza a lasciare indietro più esami del primo anno e questo sembra avere delle conseguenze nelle performance del secondo anno.

Si conferma quindi quanto visto nel paragrafo precedente: non superare con successo il primo anno di studio, accumulando soprattutto esami considerati complessi dagli studenti, potrebbe avere conseguenze che si ripercuotono sul secondo anno e oltre. Questo può tradursi nell'accumulo di ulteriori esami e, di conseguenza, aumentare la probabilità di ritardi nella laurea.

Per questo processo è stato settato un threshold di 0.3 con fitness 0.95 e precsion 0.56.

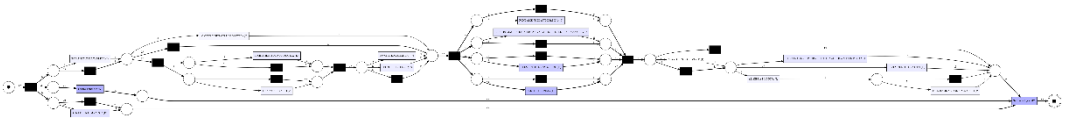

*Fig. IV.11 Percorso degli studenti "un anno in ritardo" – secondo anno*

Nel processo del secondo anno riferito agli studenti che si laureano in ritardo (Fig. IV.12) come si è motivato sopra sono presenti pochi studenti. Si può notare un comportamento comune tra gli studenti, che sembrano concentrarsi principalmente sul recupero degli esami del primo anno, trascurando gran parte degli esami obbligatori del secondo anno. Nel processo in esame, ad esempio, si rileva solo la presenza degli esami di "Elementi di Elettronica" ed "Elettrotecnica", mentre gli altri esami obbligatori del secondo anno non compaiono tra le attività di questo processo.

In generale, il secondo anno sembra essere focalizzato sugli esami del primo anno, sebbene sia interessante notare che gli esami "Fisica Generale I" e "Fisica Generale II" non compaiano nemmeno in questa fase. Ciò suggerisce che gli studenti di questa categoria affronteranno questi esami solo dopo la conclusione del secondo anno.

Per questo processo è stato settato un threshold di 0.2 con fitness 0.98 e precsion 0.83.

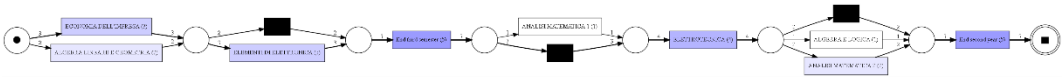

*Fig. IV.12 Percorso degli studenti "in ritardo" – secondo anno*

Di seguito si valuta la conformità dei processi con il manifesto:

| *Anno* | *Fitness studenti "in tempo"* | *Fitness studenti "in ritardo"* |
|---|---|---|
| *Primo anno* | **0.76** | **0.60** |
| *Secondo anno* | **0.60** | **0.50** |

*Tab. IV.2 Indicatori studenti iscritti negli anni accademici 2017-2018 e 2019-2020*

Similmente al paragrafo precedente, anche per questi processi è stata valutata la conformità al manifesto, effettuando un confronto tra il processo ideale e il file di log utilizzato per estrarre i processi. Anche in questo caso si utilizza quindi una metrica di fitness con un valore compreso tra zero e uno. Un valore pari a zero indica una notevole discrepanza tra il percorso seguito dagli studenti e quello ideale, mentre un valore di uno indica che gli esami sostenuti dagli studenti corrispondono esattamente a quelli indicati nel manifesto.

Nella tabella (Tab. IV.2) vengono presentati i dati relativi alla conformità degli studenti iscritti negli anni accademici 2017-2018, 2018-2019 e 2019-2020 riguardo al conseguimento della laurea nei tempi previsti, in confronto agli studenti che si laureano in ritardo. Questa analisi riguarda sia il primo che il secondo anno di laurea.

Si evidenzia una differenza tra gli studenti che aderiscono al piano di studi previsto e quelli che accumulano ritardi. Nel primo anno, gli studenti che seguono il percorso di studi in linea con le linee guida fornite dal manifesto raggiungono una percentuale di conformità del 76%. Al contrario, gli studenti che si laureano in ritardo dimostrano un minor rispetto del manifesto, con una percentuale di conformità del 60%.

Questa tendenza si ripresenta anche nel secondo anno: gli studenti che non riescono a seguire il piano di studi predefinito mostrano una minore aderenza al manifesto rispetto a coloro che riescono a farlo nei tempi stabiliti. Mentre i primi presentano una percentuale di conformità del 50%, i secondi mantengono una percentuale di conformità del 60%. Questi dati rafforzano ulteriormente il ruolo cruciale del rispetto del manifesto accademico, come già evidenziato tramite l'analisi di conformance checking nel paragrafo precedente. Emerge chiaramente che gli studenti che completano la loro laurea in tempo sono maggiormente in linea con le disposizioni del manifesto rispetto a coloro che si laureano in ritardo.

In base a questi risultati, si può concludere che la conformità al manifesto potrebbe influire significativamente sulla capacità degli studenti di laurearsi nei tempi previsti o subire ritardi. Per questo motivo, è cruciale condurre un'analisi approfondita del manifesto al fine di identificare potenziali colli di bottiglia che potrebbero minare la conformità degli studenti. Come già accennato nel paragrafo precedente questo potrebbe risultare fondamentale per aumentare le probabilità di

conseguire la laurea nei tempi prestabiliti e ridurre il numero di studenti che accumulano ritardi.

## 4.3 CORRELAZIONI TRA LE VARIABILI

Nelle sezioni precedenti, è emerso che il ritardo nel superare gli esami del primo anno può compromettere la capacità degli studenti di seguire il programma previsto nel manifesto del secondo anno e di mantenere la conformità con il flusso di esami pianificati per tale anno.

In questo contesto, si è deciso di approfondire l'analisi esaminando l'influenza del numero di esami superati con successo durante il primo anno di studio e della media finale di laurea sul tempo necessario per conseguire la laurea.

Per questo scopo, è stata creata una variabile binaria che indica se lo studente ha conseguito la laurea nei tempi previsti. Questa variabile è stata utilizzata come variabile dipendente in una regressione logistica. Le variabili indipendenti includono il numero di esami sostenuti durante il primo semestre dell'anno e il voto medio, poi è stata svolta un’ulteriore regressione logistica considerando gli esami passati dagli studenti durante il primo anno.

Oltre al numero degli esami si è deciso di inserire anche il voto medio finale tra le variabili indipendenti in quanto un punteggio elevato può essere un indicatore di competenza dello studente, ma potrebbe altresì riflettere un considerevole impegno e molto tempo per studiare.

Di seguito un estratto del dataset utilizzato per l'analisi di regressione logistica:

| ***STU_ID*** | ***N° Esami*** | ***Voto Medio*** | ***In corso*** |
|---|---|---|---|
| 000B927451412FC8224AFC15FE7AC952 | 2 | 27.5 | 1 |
| 00E6107921BADE3261DC993542DA816D | 2 | 26.8 | 1 |
| 0436AB2610AE7C96A0E86EAEF12D2BD9 | 1 | 22.8 | 0 |
| 13E0DD73FA674F372D789FD2678B1146 | 3 | 29.6 | 1 |

*Tab. IV.3 Estratto del dataset utilizzato per la regressione logistica*

La tabella (Tab IV.3) è un estratto del dataset utilizzato per la regressione logistica che tiene in considerazione gli esami superati del primo semestre. La struttura del dataset è la stessa per la regressione logistica che considera gli esami superati entro il primo anno.

Le variabili del dataset sono state standardizzate prima di applicare l'algoritmo.

Inoltre, è importante sottolineare che la variabile “In corso” assume i valori 0 e 1 per indicare rispettivamente se lo studente ha conseguito la laurea nei tempi previsti o meno.

Di seguito i risultati della regressione logistica:

| *Anno* | *Modello 1° SEMSTRE* | *Modello 1° ANNO* |
|---|---|---|
| *Numero di esami* | **2.50** | **3.74** |
| *Voto medio* | **3.73** | **2.97** |

*Tab. IV.4 Risultati regressione logistica*

Esaminando i coefficienti nella tabella (Tab. IV.4), emergono differenze significative tra il primo semestre e l'intero anno accademico nell'incidenza della media dei voti rispetto al numero di esami superati.

Nel dettaglio, l'analisi ha rivelato dinamiche interessanti nel corso dell'anno accademico. Durante il primo semestre, sembra che la media dei voti abbia un peso maggiore rispetto al numero di esami passati con successo. Questo potrebbe essere dovuto alla composizione del primo semestre, che comprende solamente tre esami. In questo contesto, anche se uno studente non riuscisse a superare tutti e tre gli esami, avrebbe ancora la possibilità di recuperare e completare il corso di laurea nei tempi previsti.

Tuttavia, quando si esamina l'arco dell'intero anno accademico, emerge che superare un numero maggiore di esami assume un'importanza maggiore rispetto al

conseguimento di voti elevati. Questo fatto è di particolare rilevanza in quanto indica che una strategia di studio efficace degli studenti dovrebbe non solo mirare al raggiungimento di buoni voti, ma anche a passare con successo la maggior parte degli esami del primo anno.

Questi risultati evidenziano l'importanza di una pianificazione accurata nel corso del primo anno di studio, poiché perdere un numero eccessivo di esami durante questo periodo potrebbe portare all'accumulo di esami da recuperare nei due anni successivi, che, come si è visto nei processi precedenti, rallenta la capacità degli studenti di passare egli esami del secondo anno entro il termine previsto.

Questa situazione può creare una sfida significativa per il conseguimento della laurea entro i tempi previsti e mettere a rischio il successo accademico complessivo dello studente.

I coefficienti vengono considerati significativi dal punto di vista statistico in quanto il valore p associato a ciascun coefficiente è inferiore a 0,05. Questo indica che le associazioni tra le variabili indipendenti e la variabile dipendente sono statisticamente rilevanti, con un alto grado di confidenza.

In conclusione, questi risultati enfatizzano l'importanza di un approccio equilibrato nel corso del primo anno accademico, che tenga conto sia del raggiungimento di buoni voti che del superamento di un numero significativo di esami. Una gestione non attenta di questa fase iniziale può avere un impatto decisivo sulla carriera

accademica complessiva dello studente e sulla sua capacità di conseguire la laurea nei tempi previsti

### 4.4 VARIANTI CON IL MINOR TEMPO DI LAUREA

L'obiettivo di questa analisi è migliorare il tasso di successo degli studenti nel completare i loro programmi accademici, compresi i corsi più impegnativi, entro i tempi per poter esser classificati come studenti laureati in tempo. Al fine di fornire ulteriori informazioni per perseguire tale scopo, ci si concentra sull'analisi del "percorso vincente", definito come la sequenza di esami che gli studenti dovrebbero seguire per ottenere il titolo laurea nel minor tempo possibile.

L'indagine si focalizza esclusivamente sugli esami sostenuti durante il primo anno del percorso di studio, in quanto si ritiene che questo periodo rivesta una fondamentale importanza per il futuro successo degli studenti. Questa affermazione è confermata da due importanti fonti di evidenza: in primo luogo, i risultati dei processi analizzati in precedenza mettono in luce l'impatto negativo sul secondo anno di studio causato dalla necessità di recuperare esami del primo anno; in secondo luogo, una correlazione significativa che dimostra l'importanza cruciale del superamento degli esami del primo anno per la buona riuscita del percorso accademico.

Quindi, basandoci sui risultati emersi dalle precedenti analisi, ora ci si concentra sull'analisi delle varianti di processo. L'obiettivo principale dell'analisi delle

varianti di processo è comprendere quanti tipi differenti di percorsi esistono e le ragioni sottostanti alle loro differenze.

Una variante di processo è definita come un insieme di casi che condividono la stessa sequenza di attività svolte nello stesso ordine, ma che possono essere distinti da altre varianti sulla base di specifiche caratteristiche o attributi. Durante questa analisi, ci si focalizza sulle varianti che corrispondono al minor tempo impiegato dagli studenti per il conseguimento della laurea. Questo approccio ci consentirà di fornire consigli basati sui dati agli studenti, aiutandoli a compiere scelte accademiche più informate e migliorando, di conseguenza, il loro tasso di successo nel percorso di studio.

Di seguito la tabella (Tab IV.5) mostra in una colonna la sequenza di esami che lo studente ha superato con successo entro il primo anno e l'altra colonna il tempo che ha impiegato in giorni per laurearsi.

| ***Esami*** | ***Tempo Laurea*** |
|---|---|
| Analisi Matematica I, Fisica Generale I, Algebra Lineare e Geometria, Fisica Generale II, Analisi Matematica II | **953** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica II, Economia dell'Impresa, Fondamenti di Informatica, Fisica Generale II | **986** |

| | |
|---|---|
| Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica I, Economia dell'Impresa, Analisi Matematica II, Fondamenti di Informatica, Fisica Generale II | **989** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Fondamenti di Informatica, Fisica Generale II, Economia dell'Impresa, Analisi Matematica II | **990** |
| Algebra Lineare e Geometria, Analisi Matematica I, Fisica Generale I, Economia dell'Impresa, Fisica Generale II, Fondamenti di Informatica | **990** |
| Analisi Matematica I, Algebra Lineare e Geometria, Analisi Matematica II, Economia dell'Impresa, Fisica Generale II, Fisica Generale I, Fondamenti di Informatica | **990** |
| Algebra Lineare e Geometria, Fisica Generale I, Fondamenti di Informatica, Analisi Matematica I, Economia dell'Impresa, Analisi Matematica II | **1000** |
| Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica I, Fondamenti Di Informatica, Analisi Matematica II, Economia dell'Impresa | **1002** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica II, Fondamenti di Informatica | **1002** |

| | |
|---|---|
| Analisi Matematica I, Fisica Generale I, Algebra Lineare e Geometria, Fondamenti di Informatica, Analisi Matematica II, Fisica Generale II, Economia dell'Impresa | **1010** |
| Analisi Matematica I, Fisica Generale I, Algebra Lineare e Geometria, Fondamenti di Informatica, Economia dell'Impresa, Analisi Matematica II | **1013** |
| Fisica Generale I, Algebra Lineare e Geometria, Analisi Matematica I, Analisi Matematica II, Economia dell'Impresa, Fondamenti di Informatica | **1014** |
| Algebra Lineare e Geometria, Analisi Matematica I, Fisica Generale I, Fondamenti di Informatica, Economia dell'Impresa, Fisica Generale II, Analisi Matematica II | **1016** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Economia dell'Impresa, Analisi Matematica II | **1017** |
| Algebra Lineare e Geometria, Analisi Matematica I, Fisica Generale I, Economia dell'Impresa, Analisi Matematica II, Fondamenti di Informatica, Fisica Generale II | **1018** |
| Algebra Lineare e Geometria, Analisi Matematica I, Fisica Generale I, Economia dell'Impresa, Analisi Matematica II, Fisica Generale II, Fondamenti di Informatica | **1024** |

| | |
|---|---|
| Algebra Lineare e Geometria, Analisi Matematica I, Fisica Generale I, Analisi Matematica II, Economia dell'Impresa, Fondamenti di Informatica, Fisica Generale II | **1045** |
| Fisica Generale I, Algebra Lineare e Geometria, Analisi Matematica II, Economia dell'Impresa, Analisi Matematica I | **1092** |
| Analisi Matematica 1, Algebra Lineare e Geometria, Analisi Matematica II, Economia dell'Impresa, Fisica Generale I, Fondamenti di Informatica, Fisica Generale II | **1105** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica II, Fisica Generale II | **1120** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Fondamenti Di Informatica, Analisi Matematica II, Fisica Generale II, Economia dell'Impresa | **1189** |
| Analisi Matematica I, Algebra Lineare e Geometria, Economia dell'Impresa, Fondamenti di Informatica, Fisica Generale I | **1225** |
| Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica I, Fondamenti di Informatica, Economia dell'Impresa, Fisica Generale II | **1233** |

| | |
|---|---|
| Algebra Lineare e Geometria, Fisica Generale I, Fondamenti di Informatica, Economia dell'Impresa, Fisica Generale II, Analisi Matematica II | **1239** |
| Analisi Matematica I, Algebra Lineare e Geometria, Fisica Generale I, Analisi Matematica II, Fisica Generale II, Economia dell'Impresa, Fondamenti di Informatica | **1302** |

*Tab. IV.5 Sequenza degli esami e relativo tempo di laurea*

È importante evidenziare che nella tabella ciascuna variante comprende un numero diverso di studenti, e il valore della variabile "Tempo Laurea" rappresenta la media dei tempi di laurea di tutti gli studenti appartenenti a quella specifica variante.

Interpretando questa analisi si deve tenere in considerazione che uno studente per essere classificate come laureato in tempo, deve terminare i suoi studi e quindi conseguire il titolo di laurea entro 1309 giorni. Ovvero gli studenti che hanno conseguito la laurea entro 3 anni e 6 mesi in base all'indicatore iC02 che tiene conto della percentuale di laureati entro la durata normale del corso.

L'analisi condotta rivela che non è strettamente necessario affrontare tutti gli esami programmati durante il primo anno per conseguire la laurea nel minor tempo possibile. Al contrario, osservando il primo studente: ha superato con successo 5 esami su 7 il primo anno e ha conseguito il titolo di laurea in soli 990 giorni. Tale constatazione non rappresenta un caso isolato, poiché vi sono altre varianti in cui

gli studenti non affrontano tutti e sette gli esami ma riescono comunque a laurearsi nei tempi previsti.

Tuttavia, è importante notare che tutte queste varianti di percorso condividono la caratteristica comune di includere esami precedentemente identificati come sfide significative per gli studenti, soprattutto per quelle categorie di studenti che completano il loro percorso con un anno di ritardo o in ritardo.

Dall'analisi condotta in precedenza, emergono le sfide rappresentate dagli esami di "Fisica Generale I", "Analisi Matematica I", "Fisica Generale II", e "Analisi Matematica II". È stato osservato che questi specifici esami sembrano particolarmente impegnativi per gli studenti, e ciò si riflette nel fatto che gli studenti tendono a concentrare i loro sforzi nel recuperarli principalmente durante il secondo anno del percorso di studio.

Questa tendenza comporta conseguenze significative poiché, mentre gli studenti si dedicano al recupero di questi esami, incontrano delle difficoltà nell'affrontare gli esami obbligatori del secondo anno. Questa situazione genera un effetto a catena che li porta a rimanere costantemente indietro rispetto al programma accademico previsto e di conseguenza li porta ad un ritardo di almeno un anno dal conseguimento del titolo di laurea.

Inoltre, va sottolineato che le traiettorie accademiche degli studenti che seguono tali varianti si mantengono abbastanza fedeli al manifesto del corso di studio. Questa osservazione conferma ulteriormente che il rispetto del manifesto rappresenta un

elemento cruciale per conseguire la laurea nei tempi prestabiliti e rientrare nella categoria di studenti laureati in tempo.

Da quest'analisi emerge quindi che ciò che contraddistingue gli studenti che completano il loro percorso di laurea nel minor tempo possibile dagli studenti che si laureano con almeno un anno di ritardo è la loro capacità di aderire al manifesto accademico e superare gli esami comunemente considerati impegnativi, o almeno alcuni di essi, entro il primo anno. Questo porta all'individuazione di due prerequisiti principali: rispetto del manifesto e superamento degli esami considerati dagli studenti difficili, come elementi chiave per il successo nel completamento del percorso di laurea nei tempi previsti.

## 4.1 PREDIZIONI DEL TEMPO DI LAUREA

In ultimo, in questo paragrafo, l'obiettivo è quello di prevedere un aspetto cruciale per gli studenti universitari: il tempo di laurea. La predizione del tempo necessario per completare un corso di laurea rappresenta un obiettivo di grande rilevanza nel contesto accademico. Un'analisi accurata e informata di questa variabile può fornire importanti indicazioni sia per gli studenti che per le istituzioni accademiche.

Per creare il dataset utilizzato per le previsioni, è stata generata una variabile dipendente in cui è stato assegnato il valore 1 ai casi in cui il tempo necessario per il conseguimento della laurea era inferiore a 1309 giorni, mentre è stato assegnato il valore 0 ai casi in cui il tempo superava tale soglia. Questa soglia corrisponde a

un periodo di 3 anni e 6 mesi, come stabilito dall'indicatore iC02 definito dall'Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR), per poter considerare uno stendete laureato in tempo.

Mentre come variabili dipendenti si considera gli esami che compongono il primo anno accademico. Questa scelta è motivata da una serie di evidenze finora emerse, che hanno rivelato la centralità del primo anno nel percorso di studio. Tale centralità si è manifestata chiaramente sia attraverso l'analisi dei processi nel secondo anno che tramite l'esame delle correlazioni tra il numero di esami sostenuti al primo anno e la durata complessiva del percorso di laurea. Inoltre, il ruolo centrale del primo anno è ulteriormente confermato dall'osservazione delle varianti con il minor tempo di conseguimento della laurea.

Dalle precedenti analisi è quindi emerso che non superare con successo alcuni esami durante il primo anno accademico comporta una diminuzione della probabilità di riuscire a completare il piano di studi entro i tempi prestabiliti.

Nel processo di attribuzione dei valori alle variabili indipendenti, è stato adottato un approccio che si basa sulla misurazione del tempo impiegato per completare ciascun esame. Per ottenere questo valore, si è calcolata la differenza in giorni tra la prima attività registrata in relazione all'esame considerato e l'ultima attività rilevata, che coincide con “Promosso”. Questo approccio ci fornisce una stima del tempo trascorso dallo studente nel tentativo di superare l'esame a partire dalla sua prima decisione di affrontarlo. Un valore di 0 indica che lo studente ha prenotato e

superato l'esame al primo tentativo, mentre un valore di 1000 è stato assegnato per rappresentare il caso in cui lo studente non ha conseguito con successo quell'esame entro il primo anno.

Di seguito un estratto del dataset che è stato utilizzato per gli algoritmi di predizione:

| ***In Corso*** | ***A*** | ***B*** | ***C*** | ***D*** | ***E*** | ***F*** | ***G*** |
|---|---|---|---|---|---|---|---|
| 1 | 1000 | 1000 | 20 | 244 | 1000 | 0 | 77 |
| 0 | 1000 | 1000 | 1000 | 0 | 1000 | 44 | 1000 |
| 1 | 0 | 0 | 176 | 0 | 0 | 0 | 22 |
| 1 | 128 | 0 | 39 | 36 | 0 | 27 | 93 |

*Tab. IV.6 Estratto del dataset utilizzato per le predizioni*

La tabella (Tab. IV.6) rappresenta la struttura del dataset utilizzato per effettuare le previsioni. Va notato che le lettere dell'alfabeto sono utilizzate esclusivamente a fini rappresentativi; nel dataset, queste lettere corrispondono agli esami del primo anno.

Il dataset non è stato sottoposto a standardizzazione poiché tutte le variabili indipendenti hanno la stessa unità di misura, ossia il tempo in giorni che gli studenti hanno impiegato per superare un esame. Inoltre, è stato utilizzato un valore di 1000 per indicare chi non ha sostenuto un determinato esame. Standardizzare in questo contesto avrebbe comportato la perdita di questa distinzione tra coloro che hanno

superato l'esame e coloro che non l'hanno fatto, appiattendo così questa differenza nel dataset.

Per questo dataset sono stati utilizzati due algoritmi: regressione logistica e SVM.

Si inizia commentando i risultati per la regressione logistica: l'algoritmo è stato iterato con l'obiettivo di identificare la configurazione ottimale per “test size” e “random state” che massimizzasse l’F1-score.

Questa scelta è motivata da tre aspetti fondamentali:

- *Stabilità dei risultati:*

  L'efficacia di molti algoritmi di machine learning può variare notevolmente in base alla suddivisione dei dati in training set e test set e al seed del generatore casuale (“random state”). Eseguendo un ciclo for con diverse combinazioni di questi parametri, è possibile valutare la stabilità delle prestazioni del modello e identificare quelle che producono risultati più coerenti e affidabili.

- *Ottimizzazione delle prestazioni*:

  L’F1 score è una metrica comune per la classificazione, ed è importante massimizzarlo per ottenere un modello che bilanci precision e recall. La ricerca di combinazioni di “random state” e “test size” che massimizzano l’F1 score può aiutare a ottenere il modello migliore possibile per i dati a disposizione.

- *Evitare il bias del dataset di test:*

  Divisioni casuali dei dati in training e test set possono portare a dataset di test che non sono rappresentativi dei dati reali, il che può influenzare negativamente le prestazioni del modello. Utilizzando diverse suddivisioni con diverse combinazioni di “random state” e “test size”, si può cercare di mitigare questo problema e ottenere una stima più accurata delle prestazioni del modello.

Inoltre, l'esecuzione di un ciclo for su una gamma di valori per “random state” e “test size” consente di esplorare in modo completo lo spazio dei parametri e individuare le configurazioni ottimali.

In sostanza, questo approccio aiuta a garantire che le prestazioni del modello siano robuste, generalizzabili e ottimali per il dataset specifico, contribuendo così a migliorare la qualità delle previsioni.

È stata utilizzata l'implementazione della regressione logistica fornita dalla libreria *scikit-learn*, con una divisione ottimizzata dei dati. Questa suddivisione ha assegnato l'80% dei dati all'addestramento del modello, riservando il restante 20% per il test.

Di seguito si presenta i risultati emersi dall’analisi predittiva svolta considerando come variabile dipendente il tempo di laurea.

Si osserva le performance del modello:

| | *Precision* | *Recall* | *F1-score* | *Support* |
|---|---|---|---|---|
| *0* | **0.79** | **0.69** | **0.73** | **35** |
| *1* | **0.81** | **0.88** | **0.85** | **50** |
| *weighted avg* | **0.80** | **0.80** | **0.80** | **82** |

*Tab. IV.7 Performance della regressione logistica*

La tabella (Tab. IV.7) riporta le misurazioni delle prestazioni del modello di classificazione di regressione logistica:

- *Precision:*

  Per la classe 0, il modello ha una precisione del 79%, il che significa che su 100 previsioni positive fatte per questa classe, circa il 79% sono effettivamente corrette. Per la classe 1, la precisione è ancora migliore, pari all'81%. In generale, il modello ha dimostrato di essere preciso nella previsione di entrambe le classi.

- *Recall:*

  Per la classe 0, il modello ha un tasso del 69%, il che indica che è in grado di individuare il 69% dei veri positivi rispetto al totale dei casi positivi per questa classe. Per la classe 1, è ancora migliore, raggiungendo l'88%. Il modello ha dimostrato di avere un'elevata capacità nell'individuare i casi positivi per la classe 1.

- *F1-score:*

  L'F1-Score combina precision e recall in un'unica misura. Per la classe 0, lo score è del 73%, mentre per la classe 1 è dell'85%. Questi F1-score riflettono un buon equilibrio tra precisione e richiamo per entrambe le classi.

- *Support:*

  Questa colonna rappresenta il numero totale di campioni appartenenti a ciascuna classe nel dataset di test. In questo caso, ci sono 35 campioni della classe 0 e 50 campioni della classe 1 nel dataset di test.

In generale, questi risultati indicano che il modello ha prestazioni soddisfacenti, con una buona precision, un buon recall e un equilibrio tra le due metriche per entrambe le classi.

- *La curva di ROC:*

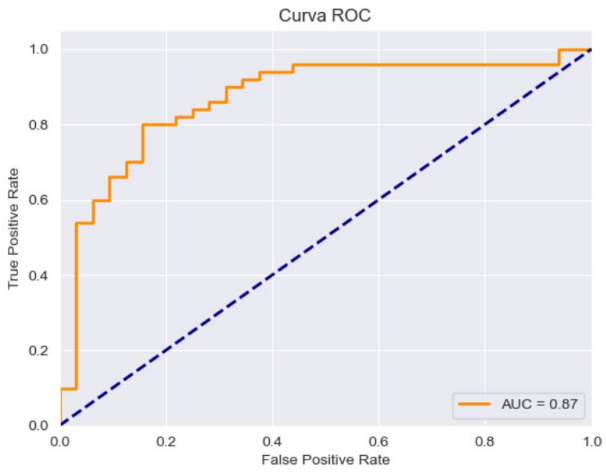


*Fig. IV.13 Curva di ROC (modello di regressione logistica)*

Una curva ROC che si estende verso l'alto a sinistra e un valore AUC-ROC di 0,87 indicano che il modello ha una buona capacità discriminante tra le classi. In altre parole, quando la curva ROC si estende verso l'alto a sinistra suggerisce che il modello è in grado di identificare correttamente un alto numero di veri positivi (True Positive) mentre limita i falsi positivi (False Positive) al minimo. Questo è un segno positivo delle buone prestazioni del modello.

Inoltre, un AUC-ROC di 0,87 suggerisce che il modello ha una capacità discriminante significativamente migliore rispetto a una previsione casuale.

Di seguito l'interpretazione dei coefficienti del modello:

| *Esami* | *Coefficienti* |
|---|---|
| *Fisica Generale II* | **-0.00173** |
| *Fondamenti di Informatica* | **-0.00026** |
| *Fisica Generale I* | **-0.00138** |
| *Analisi Matematica I* | **-0.00202** |
| *Analisi Matematica II* | **-0.00133** |
| *Algebra Lineare e Geometria* | **-0.00093** |
| *Economia dell'Impresa* | **-0.00069** |

*Tab. IV.8 Coefficienti della regressione logistica*

I risultati ottenuti dall'output della regressione logistica suggeriscono l'importanza relativa delle diverse variabili (esami) nell'influenzare la probabilità che uno studente si laurei in tempo (y = 1) o non si laurei in tempo (y = 0).

I coefficienti associati ai vari esami (Tab. IV.8) indicano quanto ciascun esame contribuisce o influisce sulla probabilità di laurearsi in tempo.

Qui di seguito un commento:

- *Fisica Generale II*: un coefficiente negativo di -0.00173 suggerisce che un aumento dei giorni necessari per superare questo esame è associato a una diminuzione della probabilità di laurearsi in tempo. In altre parole, se uno studente impiega più tempo per superare Fisica Generale II, c'è una tendenza a ritardare la laurea.
- *Fondamenti di Informatica*: il coefficiente negativo di -0.00026 indica che un aumento dei giorni impiegati per passare questo esame ha un impatto negativo ma relativamente modesto sulla probabilità di laurearsi in tempo. In questo caso, il ritardo dovuto a Fondamenti di Informatica è meno significativo rispetto ad altri esami.
- *Fisica Generale I*: con un coefficiente negativo di -0.00138, una prolungata preparazione per superare questo esame influisce negativamente sulla probabilità di laurearsi in tempo. Un ritardo nel superamento di Fisica Generale I può comportare un ritardo complessivo nella laurea.

- *Analisi Matematica I*: il coefficiente negativo di -0.00202 suggerisce che un prolungato periodo di preparazione per Analisi Matematica I ha un forte impatto sulla probabilità di laurearsi in tempo. Questo esame sembra essere particolarmente critico per il conseguimento della laurea nei tempi previsti.
- *Analisi Matematica II*: con un coefficiente negativo di -0.00133, Analisi Matematica II ha un impatto negativo significativo sulla probabilità di laurearsi in tempo se lo studente impiega più tempo per superarlo.
- *Algebra Lineare e Geometria*: un coefficiente negativo di -0.00093 indica che un prolungato periodo di studio per questo esame influisce in modo negativo, sebbene in modo meno marcato rispetto ad altri esami, sulla probabilità di laurearsi in tempo.
- *Economia dell'Impresa*: con un coefficiente negativo di -0.00069, una preparazione prolungata per questo esame ha un impatto negativo ma moderato sulla probabilità di laurearsi in tempo.

Tuttavia, i coefficienti degli esami “Fondamenti di Informatica”, “Algebra Lineare e Geometria” e “Economia dell’impresa” mostrano coefficienti molto bassi, infatti i p-value associati sono maggiori di 0.05, si può quindi dire che non sono statisticamente significativi e che non superare questi esami entro il primo anno di studi non sembra avere un reale impatto sulla durata complessiva del percorso di laurea.

In sintesi, un aumento del tempo dedicato a superare ciascuno gli esami con coefficienti significativi è associato a un ritardo nel conseguimento della laurea. Gli esami di “Analisi Matematica I” e “Analisi Matematica II” sembrano avere l'effetto più significativo sulla probabilità di ritardare la laurea, seguiti da “Fisica Generale I” e “Fisica Generale II”. Ciò rafforza quanto emerso nei paragrafi precedenti, poiché questi esami sono frequentemente non superati da molti studenti classificati come “un anno in ritardo” o “in ritardo” nel loro percorso di studi. Di conseguenza, nel secondo anno, gli studenti cercano di recuperare questi esami, ma spesso non riescono a laurearsi nei tempi previsti. Come evidenziato anche dall'analisi delle varianti, il mancato superamento di questi esami è correlato all'aumento del tempo necessario per conseguire la laurea.

Utilizzando il modello di regressione logistica ottenuto si mostrano degli esempi di studenti che non fanno parte del dataset originale per valutare il comportamento del modello nel classificarli, ossia se prevede che si laureino in tempo o meno.

Esempio 1:

| *Esami* | *Tempo* |
|---|---|
| *Fisica Generale II* | **20** |
| *Fondamenti di Informatica* | **1000** |
| *Fisica Generale I* | **20** |
| *Analisi Matematica I* | **20** |
| *Analisi Matematica II* | **20** |
| *Algebra Lineare e Geometria* | **1000** |
| *Economia dell'Impresa* | **1000** |

*Tab. IV.9 Esempio studente 1 (regressione logistica)*

Nel primo esempio lo studente supera gli esami identificati come i più rilevanti in base ai coefficienti della regressione logistica, ma non supera comunque tre esami che, come si è visto, hanno un minor impatto sulla tempistica della laurea.

Sono stati assegnati 20 giorni come intervallo temporale per indicare l'intervallo di tempo tra quando lo studente inizia a prepararsi per l'esame e quando effettivamente supera l'esame. Mentre è stato assegnato un valore di1000 per indicare indica, come precedentemente menzionato, che lo studente non è ancora riuscito a superare l'esame entro il primo anno. In questa situazione, il modello predice un esito positivo, ovvero che lo studente riesce a laurearsi in tempo.

Esempio 2:

| *Esami* | *Tempo* |
|---|---|
| *Fisica Generale II* | **1000** |
| *Fondamenti di Informatica* | **20** |
| *Fisica Generale I* | **20** |
| *Analisi Matematica I* | **20** |
| *Analisi Matematica II* | **1000** |
| *Algebra Lineare e Geometria* | **20** |
| *Economia dell'Impresa* | **20** |

*Tab. IV.10 Esempio studente 2 (regressione logistica)*

Nel secondo esempio, lo studente ha superato la maggior parte degli esami, ma conclude il primo anno senza aver ancora superato due degli esami identificati dalla regressione logistica come particolarmente influenti: “Fisica Generale II” e “Analisi Matematica II”. In questa circostanza, il modello restituisce un valore di 0, indicando che lo studente non è riuscito a laurearsi entro il periodo previsto.

Questo conferma quanto emerso anche dall'analisi delle varianti, ovvero che non è essenziale superare tutti gli esami del primo anno, ma è essenziale concentrarsi sui quattro esami che sembrano critici nel primo anno. Ciò perché nel caso in cui alcuni degli esami meno influenti vengano rimandati al secondo anno, gli studenti sembrano avere maggiori probabilità di recuperarli senza subire ritardi significativi

nella laurea. Tuttavia, lasciare indietro anche due dei quattro esami critici sembra avere un impatto notevole sulla tempistica della laurea.

Si passa ora a discutere i risultati ottenuti utilizzando l'algoritmo SVM sullo stesso dataset (Tab. IV.7). Anche in questo caso, è stata utilizzata la libreria *scikit-learn* ed il dataset è stato suddiviso in due parti, in moco coerente al modello precedente: l'80% dei dati è stato utilizzato per addestrare il modello (train), mentre il restante 20% è stato utilizzato per testarlo (test). Per questa analisi, si è scelto di utilizzare il kernel lineare come metodo di base per la SVM: questa scelta consente di ottenere risultati interpretabili e di valutare le prestazioni della SVM nel contesto di questo dataset specifico.

La scelta di utilizzare un kernel lineare per un modello SVM può essere giustificata sulla base di alcune considerazioni:

- *Interpretabilità:*

  Il kernel lineare è il più semplice tra le opzioni di kernel disponibili per le SVM. Crea decisioni di separazione lineare, il che significa che è più interpretabile rispetto a kernel più complessi come il kernel polinomiale o il kernel gaussiano (RBF). Questo rende più facile comprendere come il modello sta facendo le sue previsioni.

- *Visualizzazione dei coefficienti:*

  Con il kernel lineare, è possibile ottenere direttamente i coefficienti di peso associati a ciascuna feature nel modello. Questi coefficienti rappresentano

l'importanza relativa delle feature nella decisione del modello, rendendo possibile l'interpretazione dei fattori che influenzano le previsioni.

In sintesi, l'uso di un kernel lineare in una SVM può essere utile quando si cerca un buon equilibrio tra prestazioni del modello, interpretabilità dei risultati e velocità di addestramento, specialmente se l'obiettivo è analizzare i coefficienti delle feature per comprendere meglio il comportamento del modello.

Il modello è stato scelto eseguendo un ciclo per la scelta del parametro di regolarizzazione “C” considerando l’F1-score pesato corrispondente per ognuno di questi valori.

Di seguito, una spiegazione: è stato creato un Data Frame con due colonne: “C” per i valori del parametro di regolarizzazione e “F1 score” per l’F1-score pesato. Il ciclo for è stato iterato attraverso un intervallo di valori per il parametro di regolarizzazione “C”. L’intervallo di valori va da 1 a 1000 con step di 10.

Alla fine del ciclo, il Data Frame contiene tutte le combinazioni di valori “C” e l’F1-score pesato corrispondente per ogni iterazione del ciclo.

Questo fornisce un'idea di come le prestazioni del modello SVM variano al variare del parametro di regolarizzazione “C” ed è stato scelto il parametro “C” per il quale corrispondeva un F1-score maggiore.

La scelta del parametro “C” in una SVM è cruciale poiché influisce sulla regolarizzazione del modello e, quindi, sulle sue prestazioni. L'obiettivo principale in questa analisi era trovare il valore di “C” che massimizzasse l’F1-score.

Questa procedura è stata svolta per varie ragioni:

- *Bilancio tra bias e varianza:*

  Il parametro di regolarizzazione “C” controlla il trade-off tra la complessità del modello e la sua capacità di adattarsi ai dati di addestramento. Un valore basso di “C” porta a un modello con una maggiore capacità di regolarizzazione, riducendo la varianza ma potenzialmente aumentando il bias. Al contrario, un valore alto di “C” consente al modello di adattarsi meglio ai dati di addestramento, riducendo il bias ma potenzialmente aumentando la varianza.

- *Scelta ottimale:*

  Trovare il valore di “C” che massimizza l’F1-score significa che il modello è in grado di ottenere un buon equilibrio tra la capacità di adattamento ai dati di addestramento e la generalizzazione ai dati di test.

- *Evitare l'overfitting e l'underfitting:*

  La scelta di un valore di “C” troppo alto può portare a un modello che soffre di overfitting, dove si adatta troppo ai dati di addestramento ma generalizza male. D'altra parte, un valore di “C” troppo basso può portare a un modello che soffre di underfitting, che non riesce a catturare le complessità nei dati

di addestramento. Trovare il valore di "C" ottimale aiuta a evitare questi problemi e a ottenere prestazioni ottimali sul dataset di test.

Quindi utilizzando un kernel lineare ed un parametro di regolarizzazione pari a 11 sono state ottenute le seguenti performance:

| | *Precision* | *Recall* | *F1-score* | *Support* |
|---|---|---|---|---|
| *0* | **0.81** | **0.66** | **0.72** | **32** |
| *1* | **0.80** | **0.90** | **0.85** | **50** |
| *weighted avg* | **0.81** | **0.80** | **0.80** | **82** |

*Tab. IV.11 Performance SVM*

Di seguito si visualizza la curva di ROC.

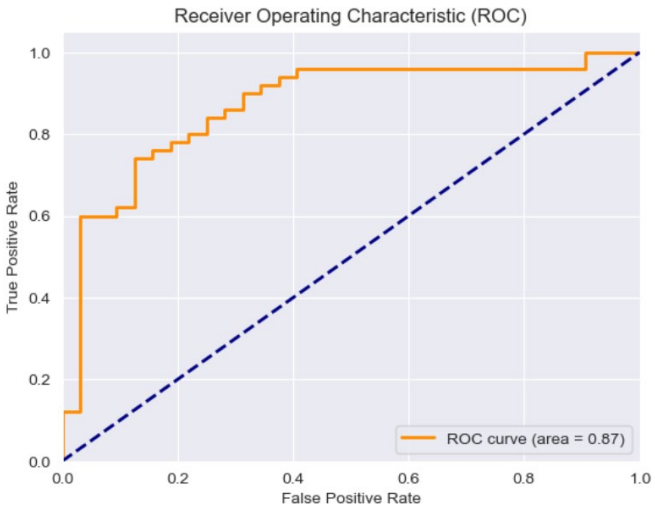


*Fig. IV.14 Curva di ROC (modello SVM)*

Una curva ROC che si estende verso l'alto a sinistra e un valore AUC-ROC di 0,87 indicano che il modello ha una buona capacità discriminante tra le classi. In altre parole, quando la curva ROC si estende verso l'alto a sinistra suggerisce che il modello è in grado di identificare correttamente un alto numero di veri positivi (True Positive) mentre limita i falsi positivi (False Positive) al minimo. Questo è un segno positivo delle buone prestazioni del modello.

Inoltre, un AUC-ROC di 0,87 suggerisce che il modello ha una capacità discriminante significativamente migliore rispetto a una previsione casuale.

Di seguito si interpreta i risultati del modello:

| *Esami* | *Coefficienti* |
|---|---|
| *Fisica Generale II* | **-0.00136** |
| *Fondamenti di Informatica* | **-0.00042** |
| *Fisica Generale I* | **-0.00375** |
| *Analisi Matematica I* | **-0.00162** |
| *Analisi Matematica II* | **-0.00181** |
| *Algebra Lineare e Geometria* | **-0.00109** |
| *Economia dell'Impresa* | **-0.00085** |

*Tab. IV.12 Coefficienti del modello SVM*

I coefficienti della tabella (Tab. IV.12) indicano l'effetto delle variabili predittive (gli esami) sulla variabile dipendente (la laurea in tempo o meno, che è codificata come 0 o 1).

In particolare, questi coefficienti rappresentano quanto cambia la probabilità di laurearsi in tempo (1) rispetto a non laurearsi in tempo (0) per ciascun corso, tenendo conto degli altri fattori nel modello. I coefficienti negativi indicano che, all'incremento del tempo impiegato per passare un esame, la probabilità di laurearsi in tempo tende a diminuire.

Di seguito una breve interpretazione dei coefficienti:

In questo caso, i coefficienti indicano quanto il numero di giorni aggiunti per dare l'esame influisce sulla probabilità di laurearsi in tempo.

- *Fisica Generale II*: un coefficiente negativo di -0.00136 suggerisce che un aumento dei giorni necessari per superare questo esame è associato a una diminuzione della probabilità di laurearsi in tempo. In altre parole, se uno studente impiega più tempo per superare Fisica Generale II, c'è una tendenza a ritardare la laurea.
- *Fondamenti di Informatica*: il coefficiente negativo di -0.00042 indica che un aumento dei giorni impiegati per passare questo esame ha un impatto negativo ma relativamente modesto sulla probabilità di laurearsi in tempo. In questo caso, il ritardo dovuto a Fondamenti di Informatica è meno significativo rispetto ad altri esami.

- *Fisica Generale I*: con un coefficiente negativo di -0.00375, una prolungata preparazione per superare questo esame ha un forte impatto sulla probabilità di laurearsi in tempo. Un ritardo nel superamento di Fisica Generale I può comportare un ritardo complessivo nella laurea.
- *Analisi Matematica I*: il coefficiente negativo di -0.00162 suggerisce che un prolungato periodo di preparazione per Analisi Matematica I influisce negativamente sulla probabilità di laurearsi in tempo. Questo esame sembra essere particolarmente critico per il conseguimento della laurea nei tempi previsti.
- *Analisi Matematica II*: con un coefficiente negativo di -0.00181, Analisi Matematica II ha un impatto negativo significativo sulla probabilità di laurearsi in tempo se lo studente impiega più tempo per superarlo. Aumentare il numero di giorni impiegati per dare questo esame riduce leggermente la probabilità di laurearsi in tempo, ma incide comunque più dell'esame Analisi Matematica I.
- *Algebra Lineare e Geometria*: un coefficiente negativo di -0.00109 indica che un prolungato periodo di studio per questo esame influisce in modo negativo, sebbene in modo meno marcato rispetto ad altri esami, sulla probabilità di laurearsi in tempo.

- *Economia dell'Impresa*: con un coefficiente negativo di -0.00085, una preparazione prolungata per questo esame ha un impatto negativo ma moderato sulla probabilità di laurearsi in tempo.

Tuttavia, gli esami “Fondamenti di Informatica” ed “Economia dell’impresa” mostrano dei coefficienti molto bassi, considerando che nei modelli SVM lineari, i coefficienti rappresentano direttamente i pesi delle feature nel modello, si può dire che non passare questi due esami entro la fine del primo anno sembra non avere un reale impatto sulla durata complessiva del percorso di laurea.

I coefficienti evidenziano che alcuni corsi hanno un impatto maggiore sulla probabilità di laurearsi in tempo rispetto ad altri. Ad esempio, il coefficiente negativo per “Fisica Generale I” suggerisce che ritardare l'esame di questa materia ha un forte impatto negativo sulla probabilità di laurearsi in tempo. Questo potrebbe indicare che “Fisica Generale I” è un corso cruciale per il programma di studio e richiede un’attenzione prioritaria.

I risultati di questo modello confermano quanto osservato fino ad ora e sono in linea con le performance del modello precedente, sebbene non siano identici.

Gli esami che sembrano avere un impatto maggiore anche in questo caso sono “Fisica Generale I”, “Fisica Generale II”, “Analisi Matematica I” e “Analisi Matematica II”. Inoltre, a differenza del modello precedente sembra aver acquisito

più importanza “Algebra Lineare e Geometria” ma il coefficiente è comunque molto basso.

Anche in questo caso si fornisce un esempio per studenti che non sono inclusi nel dataset originale così da poter valutare le capacità del modello SVM nel classificarli, ossia se prevede che si laureano in tempo o meno.

Esempio 1:

| *Esami* | *Tempo* |
|---|---|
| *Fisica Generale II* | **20** |
| *Fondamenti di Informatica* | **1000** |
| *Fisica Generale I* | **20** |
| *Analisi Matematica I* | **20** |
| *Analisi Matematica II* | **20** |
| *Algebra Lineare e Geometria* | **20** |
| *Economia dell'Impresa* | **1000** |

*Tab. IV.13 Esempio studente 1 (SVM)*

Nel primo esempio lo studente supera gli esami identificati come i più rilevanti in base ai coefficienti dal modello SVM, ma non supera comunque due esami che, come si è visto, hanno un minor impatto sulla tempistica della laurea.

Anche in questo caso sono stati assegnati 20 giorni come intervallo temporale per indicare il momento in cui lo studente inizia a prepararsi per l'esame e quando

effettivamente supera l'esame. Mentre è stato assegnato un valore di1000 per indicare, come precedentemente menzionato, che lo studente non è ancora riuscito a superare l'esame entro il primo anno. In questa situazione, il modello predice un esito positivo, ovvero che lo studente si laureerà in tempo.

Esempio 2:

| *Esami* | *Tempo* |
|---|---|
| *Fisica Generale II* | **1000** |
| *Fondamenti di Informatica* | **20** |
| *Fisica Generale I* | **1000** |
| *Analisi Matematica I* | **20** |
| *Analisi Matematica II* | **1000** |
| *Algebra Lineare e Geometria* | **20** |
| *Economia dell'Impresa* | **20** |

*Tab. IV.14 Esempio studente 2 (SVM)*

Nel secondo esempio, lo studente ha superato la maggior parte degli esami, ma conclude il primo anno senza aver ancora passato tre degli esami identificati dal modello SVM come particolarmente influenti: “Fisica Generale II” e “Analisi Matematica II”. In questa circostanza, il modello restituisce un valore di 0, indicando che lo studente non riuscirà a laurearsi in tempo.

Questo conferma i risultati ottenuti precedentemente dal modello di regressione logistica: non è essenziale superare tutti gli esami del primo anno, ma è essenziale concentrarsi sugli esami che sembrano avere un impatto maggiore sul tempo di laurea: posticipare gli esami meno rilevanti al secondo anno non sembra causare ritardi significativi, mentre posticipare quelli che risultano critici ha un impatto notevole sulla tempistica per conseguire la laurea.

Gli studenti che stanno affrontando difficoltà in corsi con coefficienti negativi dovrebbero cercare supporto accademico aggiuntivo. Questo potrebbe includere l'aiuto di tutor, corsi di recupero o l'assistenza da parte dei docenti per superare tali esami in modo più efficiente. In generale, questi coefficienti forniscono informazioni preziose per gli studenti, le istituzioni accademiche e gli educatori per identificare aree in cui possono concentrarsi.

## 4.2 LIMITAZIONI E SVILUPPI FUTURI

Le limitazioni e gli sviluppi futuri di questa tesi sono legati a diversi fattori, primo fra tutti è la mancanza di dati esaustivi riguardanti la categoria degli studenti laureati in ritardo, in particolare per gli iscritti nell'anno accademico 2018-2019. Questa carenza di dati ha una serie di conseguenze che influenzano la validità e l'applicabilità dei risultati ottenuti nell'analisi.

Quindi, uno degli ostacoli affrontati in questa tesi è la mancanza di dati rappresentativi per gli studenti che hanno conseguito la laurea con ritardo. Questo

può comportare un'analisi distorta e limitata per questa specifica categoria di studenti, rendendo difficile trarre conclusioni generalizzabili soprattutto nell'analisi dei processi in quanto si è visto che la categoria degli studenti classificati come "in ritardo" comprende solo sette studenti.

La mancanza di dati per gli studenti laureati in ritardo si traduce in una ridotta dimensione del campione di studenti da cui estrarre informazioni. Nel caso in cui si desideri sviluppare modelli predittivi o effettuare analisi statistiche più complesse, questa limitazione può compromettere la robustezza e la rappresentatività dei risultati.

Poiché i dati si fermano al 2023 per gli iscritti nel 2018-2019, non è possibile valutare come gli studenti laureati in ritardo si comportino nel lungo termine o se apportino modifiche al loro percorso di studio.

Inoltre, in questa analisi sono stati utilizzati dati accademici tradizionali, è possibile ampliare le fonti di dati utilizzate per l'analisi. Ad esempio, l'integrazione di dati sociodemografici, sondaggi sugli studenti e informazioni sulle loro attività extracurriculari potrebbero fornire una panoramica più completa dei fattori che influenzano il ritardo nella laurea.

Sarebbe interessante per il futuro creare interventi basati sui risultati ottenuti in questo studio: una volta identificati i fattori chiave che contribuiscono al ritardo nella laurea, è possibile sviluppare interventi mirati per aiutare gli studenti a mantenere uno studio più efficiente. Questi interventi possono includere programmi

di tutoraggio, servizi di consulenza accademica o modifiche nei programmi di studio.

Successivamente, una prospettiva futura potrebbe essere quella di continuare a monitorare gli studenti nel corso degli anni successivi per vedere se mantengono lo stesso andamento o se adottano nuovi comportamenti in risposta a interventi o suggerimenti derivati da questa analisi. Questo offre l'opportunità di valutare l'efficacia delle azioni intraprese.

In conclusione, le limitazioni principali di questa tesi sono la mancanza di dati esaustivi sugli studenti laureati in ritardo (per gli anni accademici 2018-2019 e successivi) e l'uso di dati accademici tradizionali. Tuttavia, ci sono opportunità future per ampliare le fonti di dati, sviluppare interventi mirati per migliorare l'efficienza degli studenti e monitorare nel tempo l'efficacia delle azioni intraprese.

# 5. CONCLUSIONI

In questa tesi, l'obiettivo principale è stato quello di condurre un'analisi dettagliata del percorso accademico seguito dagli studenti all'interno di un corso di laurea triennale presso l'Università Politecnica delle Marche. L'obiettivo era comprendere appieno le scelte fatte dagli studenti durante i loro studi, nonché identificare i fattori che potrebbero contribuire a ritardi nella conclusione del loro percorso di laurea. Lo scopo ultimo di questa ricerca è fornire informazioni preziose sia agli studenti interessati che all'istituzione accademica stessa al fine di sviluppare strategie e interventi volti a migliorare l'efficienza del percorso di studio e a ridurre il tempo medio necessario per conseguire la laurea.

L'analisi dei processi ha chiaramente evidenziato come il dover recuperare esami del primo anno abbia un impatto negativo sulla progressione accademica degli studenti, contribuendo a ritardi nell'andamento del percorso di laurea. Infatti, osservando i processi degli studenti laureati in ritardo a partire dal secondo anno, emerge un trend in cui gran parte del loro impegno è concentrato nel recupero degli esami mancati, rendendo difficile il progresso secondo il piano di studio previsto.

Questo fenomeno sottolinea l'importanza di una gestione efficace degli esami del primo anno per garantire una transizione agevole al secondo anno di studio. Inoltre, l’analisi di conformance checking ha confermato quanto emerso dai processi: gli

studenti che si laureano in tempo sono più conformi al manifesto del primo anno rispetto agli studenti che si laureano in ritardo.

A rafforzare questa tesi le correlazioni emerse hanno dimostrato inequivocabilmente che il superamento degli esami del primo anno è un fattore determinante per il successo generale del percorso accademico. Questa constatazione sottolinea l'importanza di concentrare sforzi e risorse sulla fase iniziale del percorso di studio al fine di garantire una base solida per il futuro successo degli studenti.

Successivamente anche l'analisi delle varianti ha confermato quanto detto, mettendo in luce l'importanza del superare con successo alcuni esami del primo anno, oltre all'essere conformi al manifesto e non lasciare indietro esami: non riuscire a superare con successo alcuni esami entro il primo anno potrebbe avere un impatto maggiore sul tempo di laurea rispetto ad altri.

Gli studenti dovrebbero prestare particolare attenzione alla gestione del tempo quando si tratta di corsi con coefficienti negativi significativi, i risultati delle predizioni hanno confermato che non superare entro il primo anno esami come "Fisica Genarle I", "Fisica generale II", "Analisi Matematica I" e "Analisi Matematica II" comporta conseguenze maggiori sul tempo di laurea. Questi corsi potrebbero richiedere una pianificazione più accurata o uno studio più intensivo da parte degli studenti per evitare ritardi nella laurea.

In sintesi, tutte le analisi condotte hanno ribadito l'importanza cruciale del primo anno accademico per il conseguimento tempestivo della laurea. Un inizio difficile nel primo anno accademico può avere ripercussioni negative negli anni successivi e rendere difficile il recupero.

Questi risultati dovrebbero essere utilizzati come base per lo sviluppo di strategie e interventi volti a migliorare l'efficienza del percorso di studio e a ridurre il tempo medio necessario per conseguire la laurea. Gli studenti dovrebbero essere informati in modo appropriato sull'importanza di un buon inizio accademico e sulla gestione efficace del tempo per affrontare gli esami più critici.

In definitiva, questa ricerca offre una guida preziosa sia agli studenti che vogliono affrontare il loro percorso accademico con successo che all'istituzione accademica stessa, invitandola a implementare politiche e supporti mirati per garantire una formazione più efficiente.

# 6. BIBLIOGRAFIA E SITOGRAFIA


1. Indicatori Cds, https://www.anvur.it/attivita/ava/indicatori-di-monitoraggio-autovalutazione-e-valutazione-periodica/indicatori-cds/
2. Wil van der Aalst, Process Mining Data Science in Action, Second Edition, Springer, 2016
3. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, Introduction to Data Mining, Addison-Wesley
4. E. Alpaydin. Introduction to Machine Learning. MIT Press, Cambridge, MA, 2010.
5. M. A. Ghazal, O. Ibrahim and M. A. Salama, "Educational Process Mining: A Systematic Literature Review," 2017 European Conference on Electrical Engineering and Computer Science (EECS), Bern, 2017, pp. 198-203
6. A. Adriansyah, B. van Dongen, and W.M.P. van der Aalst. Conformance Checking using Cost-Based Fitness Analysis. In C.H. Chi and P. Johnson, editors, IEEE International Enterprise Computing Conference (EDOC 2011), pages 55–64. IEEE Computer Society, 2011.
7. M. Dumas, M. La Rosa, J. Mendling, and H. Reijers. Fundamentals of Business Process Management. Springer, Berlin, 2013.
8. IEEE Task Force on Process Mining. Process Mining Manifesto. In F. Daniel, K. Barkaoui, and S. Dustdar, editors, Business Process Management

Workshops, volume 99 of Lecture Notes in Business Information Processing, pages 169–194. Springer, Berlin, 2012.

9. S.J.J. Leemans, D. Fahland, andW.M.P. van der Aalst. Discovering Block-structured Process Models from Incomplete Event Logs. In G. Ciardo and E. Kindler, editors, Applications and Theory of Petri Nets 2014, volume 8489 of Lecture Notes in Computer Science, pages 91–110. Springer, Berlin, 2014.
10. W.M.P. van der Aalst, A.J.M.M. Weijters, and L. Maruster. Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, 2004.
11. E. M. Real, E. Pinheiro Pimentel, L. V. de Oliveira, J. Cristina Braga and I. Stiubiener, "Educational Process Mining for Verifying Student Learning Paths in an Introductory Programming Course," 2020 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 2020, pp 1-9
12. W. Intayoad, C. Kamyod and P. Temdee, "Process mining application for discovering student learning paths," 2018 International Conference on Digital Arts, Media and Technology (ICDAMT), Phayao, Thailand, 2018, pp. 220-224
13. X. Ma and Z. Zhou, "Student pass rates prediction using optimized support vector machine and decision tree," 2018 IEEE 8th Annual Computing and

Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2018, pp. 209-215, doi: 10.1109/CCWC.2018.8301756.

14. W. Intayoad, C. Kamyod and P. Temdee, "Process mining application for discovering student learning paths," 2018 International Conference on Digital Arts, Media and Technology (ICDAMT), Phayao, Thailand, 2018, pp. 220-224
15. A. G. Costa, E. Queiroga, T. T. Primo, J. C. B. Mattos and C. Cechinel, "Prediction analysis of student dropout in a Computer Science course using Educational Data Mining," 2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO), Loja, Ecuador, 2020, pp. 1-6
16. Pm4py documentation, https://pm4py.fit.fraunhofer.de/docs
17. Scikit-learn documentation, https://pypi.org/project/scikit-learn/
18. Statsmodels documentation, https://pypi.org/project/statsmodels/

# 7. RINGRAZIAMENTI

Un grazie alla mia famiglia, con la quale ho condiviso tutti i momenti di gioia e difficoltà vissuti in questo percorso, permettendomi con tanti sacrifici di concluderlo nelle condizioni migliori.

Un grazie speciale a mio fratello Michele, che ha creduto in me fin dal primo giorno, a volte più di quanto non lo facessi io, grazie per avermi insegnato i valori della perseveranza ed ambizione.

Un grazie a Francesco che è rimasto sempre al mio fianco, mi ha supportata e incoraggiata di fronte ad ogni dubbio e imprevisto.

Un grazie ad Annalisa: amica e compagna di studio con la quale ho preparato ogni singolo esame e sviluppato il progetto di tesi. Il supporto che ci siamo date è stata la nostra forza.

Un grazie alle mie coinquiline Alessia, Rosita e Chiara che, che in questi due anni hanno condiviso con me la loro quotidianità e mi hanno insegnato più di quanto loro credano.

Un grazie ai miei compagni di studio e amici, Simone, Damiano, Daniele, e Marco per tutti i momenti condivisi ed il supporto reciproco.

Un grazie a tutti gli amici che hanno condiviso con me questo traguardo.

Ed infine, un grazie ai professori Domenico Potena e Laura Genga per la disponibilità e la professionalità dimostratami in questi mesi.