UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI ECONOMIA "GIORGIO FUÀ"

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

Tecniche di Process Mining applicate all'analisi dei modelli di comportamento degli studenti nell'affrontare esami universitari

Process Mining techniques applied to the analysis of students' behavior patterns in taking university exams

Relatore: Prof. Domenico Potena

Tesi di Laurea di:
Annalisa Basta

Correlatrice: Prof.ssa Laura Genga

Anno Accademico 2022 – 2023

INDICE

INDICE DELLE FIGURE

INDICE DELLE TABELLE

I. INTRODUZIONE

1.1 OPPORTUNITÀ DELL'EDUCATIONAL PROCESS MINING

I dati raccolti dai sistemi informatici delle Università possono essere considerati dei punti di partenza promettenti per affrontare questioni difficili da sempre oggetto di studio delle scienze dell'apprendimento. L'istruzione moderna fa sempre più affidamento su questi sistemi per implementare nuovi strumenti in grado di assistere l'insegnamento e la valutazione, gestire i contenuti educativi pubblicati e studiare le interazioni degli studenti, facilitare il lavoro di gruppo e la discussione costruttiva alunno-alunno e alunno-insegnante.

Durante la pandemia dovuta al COVID-19, l'insegnamento e l'apprendimento online sono diventati imprescindibili sia per gli studenti che per gli insegnanti. Ciò ha creato la possibilità per gli studenti di apprendere in maniera flessibile e da qualsiasi luogo, ma il potenziale del processo educativo online non è stato completamente esplorato, non riuscendo a colmare e prevenire le difficoltà che gli studenti potrebbero incontrare durante il processo di apprendimento.

Estrarre ed analizzare i dati degli studenti, tuttavia, non permette soltanto di migliorare la produttività dell'apprendimento online ma anche l'apprendimento tradizionale in classe.

I suddetti sistemi informatici offrono opportunità senza precedenti per la raccolta di dati su larga scala sulle abitudini di studio degli studenti: questa è solo una delle motivazioni alla base della nascita di una nuova disciplina.

I campi dell'educazione e dell'apprendimento, infatti, si sono rivelati essere delle applicazioni molto interessanti per il data-mining: perciò nasce un nuovo campo di ricerca che è l'Educational Data Mining.

L'Educational Data Mining (EDM) si propone di trovare pattern e fare predizioni riguardo ai comportamenti di apprendimento e su quando e come gli studenti raggiungono gli obiettivi. EDM è una disciplina emergente, che si occupa di sviluppare metodi per esplorare dati sempre più su larga scala che provengono da contesti educativi e utilizzare tali metodi per comprendere meglio gli studenti e le loro abitudini di apprendimento, oltre che la pianificazione del curriculum e la strutturazione dei corsi. Tutte le analisi svolte nel campo dell'Educational Data Mining hanno l'obiettivo finale di migliorare l'apprendimento degli studenti, oltre che perseguire obiettivi aggiuntivi come la riduzione dei costi educativi.

In questo contesto, la valutazione del rendimento scolastico degli studenti assume un ruolo importante: l'obiettivo di quantificare il rendimento degli studenti è quello di aiutare tutte le parti interessate nel processo educativo. Per gli studenti, è importante nel momento della scelta dei corsi di laurea e per la pianificazione degli esami da svolgere. Per gli insegnanti è utile per adattare i materiali di apprendimento in base alle capacità dello studente, ma soprattutto per individuare

preventivamente gli studenti a rischio. Infine, per i responsabili dell'istruzione, la quantificazione del rendimento degli studenti potrebbe aiutare nella definizione dei curricula.

Tuttavia, è interessante analizzare il percorso che uno studente compie durante il corso dei suoi studi nella sua interezza, ossia anche da una prospettiva "process-centric": perciò si utilizza l'Educational Process Mining per estrarre e comprendere il processo educativo.

L'Educational Process Mining (EPM) è un nuovo ambito di ricerca utilizzato nell'Educational Data Mining (EDM), il cui fulcro dell'analisi è il processo, che svolge un ruolo centrale nella scoperta e visualizzazione del percorso di apprendimento degli studenti.

Le differenze tra gli approcci dell'Educational Data Mining e dell'Educational Process Mining sono le stesse che caratterizzano il Data Mining e il Process Mining: entrambe partono dai dati, ma il Process Mining ha una prospettiva dinamica. L'Educational Process Mining si basa sull'utilizzo delle tecniche di process mining per identificare le cause che risiedono dietro i ritardi nel completamento degli studi e quali esami sono un blocco per la maggior parte degli studenti che non riescono a laurearsi entro l'anno previsto.

1.2 CASO DI STUDIO

In un caso di studio reale, sono stati implementati i principi dell'Educational Process Mining per analizzare i percorsi degli studenti. Ciò consente di determinare i fattori che influenzano il completamento degli studi in modo tale da migliorare l'efficacia dell'apprendimento.

L'obiettivo di questa tesi è identificare eventuali colli di bottiglia incontrati dagli studenti durante il loro percorso di studi, analizzando le loro carriere, ossia la sequenza delle attività registrate dall'Ateneo.

I dati che sono stati presi in considerazione per questo studio provengono dalla piattaforma online "Esse3", utilizzata da molti atenei italiani per la gestione dei dati dello studente. La piattaforma consente di gestire online molte operazioni, a partire dall'immatricolazione, ma soprattutto permette allo studente di iscriversi agli esami, visualizzare il relativo voto e verbalizzarlo. I dati estratti da questa piattaforma, perciò, sono un ottimo punto di partenza per visualizzare il processo effettuato dagli studenti, osservando quando si prenotano ad un determinato esame, se vengono bocciati entro quando lo ritentano, se rifiutano un voto e qual è il voto finale.

Il focus di questa tesi sarà principalmente sulle tempistiche con le quali gli studenti effettuano gli esami, quando iniziano a studiare e se scelgono di fare gli esami nel periodo ti tempo prestabilito, ossia nel semestre per cui un determinato esame è inserito nel piano di studi, o meno. Quest'analisi è volta, perciò, ad indagare su

come gli studenti studiano ed affrontano ogni esame, per individuare quali sono i problemi incontrati durante questo percorso.

L'Agenzia per la valutazione del sistema Universitario e della ricerca (ANVUR) si occupa della valutazione della qualità degli atenei italiani, ed ha definito alcuni criteri e metodologie per il monitoraggio dei corsi di studio.

Ciascuno studente del campione considerato è stato classificato in base ai seguenti indicatori[1]:

- iC02, nel quale rientra la percentuale di laureati entro la durata normale del corso.
- iC17, che considera la percentuale di immatricolati che si laureano entro un anno oltre la durata normale del corso.

Partendo dai suddetti indicatori, sono considerati "in tempo" gli studenti che rientrano nell'indicatore iC02, ossia coloro i quali hanno conseguito la laurea entro 3 anni e 6 mesi; sono considerati "un anno in ritardo" gli studenti che si sono laureati un anno dopo il suddetto periodo e perciò rientrano nell'indicatore iC17; mentre gli studenti classificati "in ritardo" hanno impiegato più di 4 anni e 6 mesi per laurearsi, perciò si tratta di tutti gli studenti che non rientrano nei due indicatori sopracitati.

Questa suddivisione è finalizzata ad analizzare più nel dettaglio i processi degli studenti per evidenziare le differenze che intercorrono tra coloro i quali si laureano

[1] ANVUR: https://www.anvur.it/attivita/ava/indicatori-di-monitoraggio-autovalutazione-e-valutazione-periodica/indicatori-cds/#

entro la durata normale del corso e chi invece si laurea in ritardo, in modo da trovare gli ostacoli nei quali incorrono questi ultimi, per prevenirli in futuro ed estrarre considerazioni relative al miglior percorso da seguire per consigliarlo agli studenti che stanno iniziando ora i loro studi.

1.3 PANORAMICA DEL DATASET

I dati utilizzati sono relativi agli studenti iscritti al corso di laurea triennale di Ingegneria Informatica e dell'Automazione, presso l'Università Politecnica delle Marche.

In particolare, sono stati presi in considerazione i dati relativi agli studenti iscritti dall'anno accademico 2015/2016 al 2019/2020. Tutte le analisi svolte sono state condotte solo sulla base dei dati relativi agli studenti già laureati, in modo tale da poter avere informazioni sul processo completo.

Il dataset così costruito è formato da 410 studenti ed ognuno di essi è caratterizzato da un identificativo univoco in forma anonima, dalla data di iscrizione al corso di laurea, dal giorno della laurea e dalla durata del percorso di studio espresso in giorni.

Inoltre, sono presenti informazioni relative agli esami dati da ciascuno studente, anch'essi caratterizzati da un identificativo univoco, il relativo voto, oltre che dal nome del corso, il professore ad esso assegnato, e la data relativa all'iscrizione

all’esame per ogni studente, il giorno dell’appello, la data del caricamento del voto, della verbalizzazione e dell’inserimento.

Inoltre, per ogni studente sono presenti delle attività che fanno riferimento ad ogni esame dato, ciò anche nell’ottica di studiare il relativo processo: “Promosso”, “Bocciato”, “Assente”, “Ritirato”.

Precisamente, per ogni esame sostenuto da uno studente si ha l’attività “Promosso”, e la data del corrispettivo appello, se lo studente considerato ha superato l’esame; “Bocciato” nel momento in cui non ha superato la prova; “Assente” significa che lo studente ha effettuato solo l'azione di prenotazione all’appello ma non ha poi sostenuto l'esame; mentre “Ritirato” significa che lo studente si è presentato ma per vari motivi ha deciso di non sostenere l'esame.

Sulla base di queste informazioni, le analisi svolte riguardano il processo di ogni studente relativamente a ciascun esame, e nel prosieguo di questa tesi si andranno a visualizzarne i risultati.

1.3 STRUTTURA DELLA TESI

I capitoli successivi della tesi tratteranno dei seguenti temi.

Il Capitolo 2 è dedicato a fornire un quadro teorico in maniera generale sul Process Mining, ed in particolare sugli strumenti utilizzati nelle successive analisi, come l’algoritmo Inductive Miner per l’estrazione dei processi e le Reti di Petri per la

loro visualizzazione. Inoltre, nel Capitolo 2 verrà presentata la teoria sottostante le tecniche di machine learning utilizzate per la predizione del voto degli studenti: in particolare saranno delineati accenni teorici sulla Regressione Logistica e su Support Vector Machine SVM.

Il Capitolo 3 contiene la descrizione dettagliata sulla metodologia e sull'approccio utilizzato per le analisi contenute in questa tesi. In particolare, descrive il trattamento e il pre-processing dei dati preliminare alla costruzione del file log.

Nel Capitolo 4 vengono presentate le motivazioni alla base del successivo approfondimento dei processi degli studenti relativamente ad ogni materia, a partire dall'analisi di alcune statistiche.

Il Capitolo 5 è dedicato alla descrizione dettagliata sui processi estratti.

Nel Capitolo 6 sono descritti i modelli finalizzati alla predizione delle performance degli studenti, ossia della media finale da loro ottenuta, sulla base di quanti e quali esami hanno svolto durante il loro primo anno accademico.

Nel Capitolo 7 verranno illustrati i limiti di questo approccio e verranno delineati possibili sviluppi futuri, e verranno discusse le conclusioni tratte dalle analisi sopra descritte.

II. CENNI TEORICI

In questo capitolo verranno fornite, a grandi linee, le conoscenze teoriche necessarie per poter comprendere al meglio l'ambito in cui si sono sviluppate le analisi su cui si basa questo lavoro di tesi.

Di seguito verranno descritte le teorie alla base del Process Mining, l'algoritmo Inductive Miner, una tecnica di visualizzazione dei processi chiamata Rete di Petri, e successivamente accenni teorici sulla Regressione Logistica e Support Vector Machine.

2.1 PROCESS MINING

Il Process Mining è una disciplina che si concentra nell'analizzare, con tecniche nuove, i processi relativi a qualunque tipo di organizzazione. Per processi di business si intende l'esecuzione reale di un processo così come viene registrato dai sistemi dell'azienda.

Una delle principali sfide delle organizzazioni odierne è estrarre informazioni di valore dai dati archiviati nei loro sistemi informativi. Il Process Mining si concentra sulla traccia che l'esecuzione di un processo lascia in tali sistemi.

Il Process Mining è sia una branca della Data Science, il quale aggiunge una prospettiva di processo al machine learning e al data mining, che della Process

Science; infatti, le tecniche di process mining possono essere utilizzate per scoprire i modelli di processo dai dati degli eventi.

Il Process Mining, perciò, può essere considerato un ponte tra la data science e la process science[2].

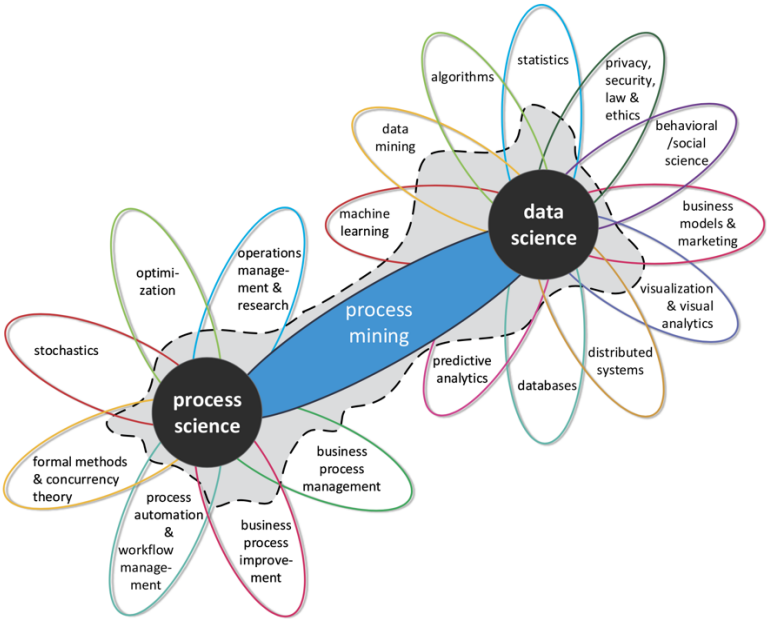

Figura II.1 - Ponte tra la Data Science e la Process Science

Perciò, il Process Mining è considerato una sotto-disciplina sia della data science che della Process Science, e le sue applicazioni sono molteplici.

Il focus del Process Mining è legato all'aspetto della generazione di eventi e i dati ad essi correlati: qualsiasi tipo di evento che interagisce con un dispositivo

[2] Wil van der Aalst, Process Mining – Data Science in Action

elettronico produce informazione. L'obiettivo del Process Mining è estrarre valore dall'evolversi degli eventi e dai dati che gli eventi generano.

Gli eventi considerati possono essere i più svariati: possono svolgersi all'interno di un sistema informativo aziendale, all'interno di un ospedale, all'interno di un social network o anche di un sistema di trasporto.

Tutte le attività aziendali possono essere sintetizzate in processi, ad esempio nel momento in cui un'azienda riceve un ordine da un cliente si attivano diverse procedure, composte da varie fasi: dal cliente che contatta l'azienda e registra la richiesta d'acquisto del prodotto, all'azienda che provvede al reperimento del prodotto, l'invio della fattura al cliente ed il pagamento della stessa.

In queste fasi i vari dipendenti all'interno dell'organizzazione registreranno dei dati su diversi database. Inoltre, sempre in ambito aziendale, il Process Mining si può utilizzare per analizzare la successione degli eventi sia relativi ad un'intera organizzazione che ad un singolo macchinario, ad esempio per la predizione della rottura dei componenti.

Un esempio di processo può anche essere la cura di un paziente in un ospedale, la quale si può considerare come una serie di attività in successione.

In sintesi, il Process Mining si concentra sull'analisi dinamica del processo, ossia guarda ad una serie di attività nella loro successione temporale.

L'obiettivo è quello di estrarre conoscenza utile dall'analisi dei processi, non è finalizzata, perciò, ad una ricerca puntale su un singolo evento.

L’analisi del Process Mining è basata, invece, su aspetti dinamici: mira a sapere se, in tutti i processi, ci sono dei colli di bottiglia, quale fase richiede più tempo e come si può ottimizzare, oltre a verificare la conformità del processo, confrontare le varianti dei processi e suggerire miglioramenti.

Questo tipo di analisi si può fare solo considerando numerose diverse esecuzioni reali di un processo.

È una sfida importante non soltanto nell’ambito dei Big Data, ma anche riguardo ai dati di una dimensione non troppo elevata, i quali sono complessi da analizzare nel loro carattere dinamico, perché ci sono gli aspetti temporali da considerare.

Dall’analisi di processo è possibile ottenere risposte a domande come “Qual è il processo che le persone seguono realmente?”, “Quali sono i colli di bottiglia nel processo? Da quali fattori sono influenzati?”, “Dove avvengono le deviazioni?”, “Si possono predire eventuali malfunzionamenti?”, “Come ottimizzare il processo?”.

2.1.1 Tecniche di process mining

Gli event log possono essere utilizzati per condurre tre diversi tipi di analisi di process mining: Process Discovery, Conformance Checking ed Enhancement.

Una tecnica di Process Discovery mira ad estrarre il processo dai dati, ossia prende in input un event log e produce un modello per spiegare il comportamento catturato

nel log, senza utilizzare alcuna informazione a priori. Se l'event log contiene anche informazioni sulla risorsa, il modello ottenuto riesce anche a spiegare come le persone interagiscono tra loro all'interno di un'organizzazione.

La Conformance Checking, invece, utilizza un modello di processo che rappresenta il processo ideale, ossia l'esecuzione del processo nel modo in cui idealmente dovrebbe essere, e i dati che provengono dalla reale esecuzione del processo, che rappresentano il processo effettivo. L'obiettivo della Conformance Checking è stimare con quale grado la realtà è conforme al processo ideale.

Perciò, questa tecnica di analisi può essere utilizzata per controllare se la realtà registrata dal file log sia conforme al modello e viceversa, oltre che per verificare se le procedure stabilite siano rispettate o meno, oppure per individuare e spiegare possibili deviazioni del processo.

L'Enhancement viene utilizzato per migliorare un modello di processo già esistente, utilizzando le informazioni riguardo il processo registrato dall'event log.

Esistono due tipologie di Enhancement: Repair, che permette di modificare il modello in modo che esso rifletta meglio la realtà; ed Extension per aggiungere una nuova prospettiva al modello, per esempio aggiungendo dati relativi alle performance o informazioni sulle risorse. La conformance checking è propedeutica per l'enhancement, ossia si può modificare lo schema nel caso in cui si trovi qualche deviazione che porta effettivamente ad un vantaggio.

In definitiva, si può affermare che l'elemento chiave del process mining sia stabilire una relazione tra il modello di processo e la "realtà" catturata dai dati.

Si parla di "Play-in" nel momento in cui partendo dall'event log si estrae il modello. Si parte da un log, che descrive la successione degli eventi, e da ciò si costruisce uno schema che rappresenta tutte le successioni possibili.

Il "Play-out", invece, si riferisce all'utilizzo classico dei modelli di processo: partendo dal modello (es. Petri net) si genera un "comportamento". È utilizzato spesso per la simulazione, ossia per simulare diverse possibili esecuzioni del processo, estraendo statistiche ed intervalli di confidenza.

Il termine "Replay" si riferisce, invece, all'individuazione delle differenze tra la realtà ed il modello estratto. Il Replay utilizza come input l'event log e il modello del processo, e verifica per ogni elemento se c'è corrispondenza tra il processo ed i dati.

2.1.2 File log

Tutte le analisi di Process Mining hanno la stessa base di partenza, ossia l'event log: registri elettronici sulla quale si annotano dettagli sulle diverse azioni svolte in un contesto aziendale. Questi registri contengono informazioni riguardo a chi ha eseguito ciascuna azione, il suo ruolo o dipartimento, l'istante in cui l'azione è avvenuta e altro ancora. Ogni azione segue un determinato percorso, tipicamente

composto da un inizio e una conclusione, che segnano l'avvio e il completamento di un'attività.

Perciò, un event log è composto da una sequenza di voci registrate in ordine cronologico. Gli elementi fondamentali per la costruzione di un file log sono: il "Case ID", l' "Event ID", il "Timestamp" e l' "Activity".

Il "Case ID" funge da chiave per collegare le diverse istanze di un processo all'interno del registro, in modo tale da riuscire a tracciare tutte le attività svolte (permette di ricostruire la cronologia delle attività). Tutti gli eventi registrati con lo stesso Case ID corrispondono alle attività svolte in un determinato processo.

Gli altri campi forniscono informazioni come il nome dell'attività, il tipo di evento associato (ad esempio, inizio o fine dell'attività), la data e l'ora dell'esecuzione, oltre che informazioni aggiuntive come la persona responsabile dell'azione.

Il Timestamp, inoltre, permette di calcolare la durata del processo, tramite la differenza tra quando il processo è terminato e l'istante di tempo in cui è cominciato.

Un processo è composto da casi, che a loro volta consistono in una sequenza di eventi.[3] Ogni evento consiste in un'attività svolta all'interno di un caso, ed il loro ordine fornisce un quadro cronologico delle azioni svolte all'interno del processo.

[3] Wil van der Aalst, Process Mining – Data Science in Action

I file con cui vengono registrati i log hanno un formato XES (eXtensible Event Stream).

Un log è costituito da tracce ed eventi. Una traccia è una serie di eventi unici che riguardano lo stesso caso, anche detta “istanza di processo”. Un evento è un’esecuzione di una particolare attività in un certo istante di tempo.

2.2 RETI DI PETRI

Una rete di Petri è un modello formale utilizzato per descrivere e analizzare sistemi con concorrenza, sincronizzazione e condivisione di risorse. È stato introdotto da Carl Adam Petri negli anni '60 ed è ampiamente utilizzato in diversi campi, come l'informatica e l'ingegneria.

Una rete di Petri è un grafo bipartito costituito da *places* e transizioni. La struttura della rete è statica, sulla quale i token possono essere “accesi” attraverso la “*firing rule*”. Lo stato di una rete di Petri è determinato dalla distribuzione dei *token* sui place e viene chiamato *marking*.

Formalizzando la definizione, una rete di Petri è una tripletta $N = (P,T,F)$ dove P è un insieme finito di places, T è un insieme finito di transizioni tali che $P \cap T = \emptyset$, e $F \subseteq (P \times T) \cup (T \times P)$ è un insieme di archi orientati, chiamato relazione di flusso. Una rete di Petri *marked* è una coppia (N,M), dove $N = (P,T,F)$ è una rete di Petri e dove $M \in B(P)$ è un multi-set su P, la quale fa si che la rete sia marcata.

L'insieme di tutte le reti di Petri contrassegnate è indicato con $\mathcal{N}$.[4]

Le Reti di Petri, perciò, sono composte da places, transizioni, token e archi direzionati che collegano i places alle transizioni e viceversa.

I places rappresentano lo stato dell'attività. Possono contenere un certo numero di tokens, che possono rappresentare risorse, dati, o qualsiasi altra entità significativa.

Le transizioni rappresentano il fatto che l'attività è stata svolta. Quando una transizione è abilitata può consumare tokens dai place di ingresso e produrre tokens nei place di uscita.

I tokens permettono di capire dove è arrivato il processo spostandosi tra i diversi places.

Gli archi rappresentano la relazione tra place e transizioni, indicando come i token possono spostarsi attraverso la rete.

[4] Wil van der Aalst, Process Mining – Data Science in Action

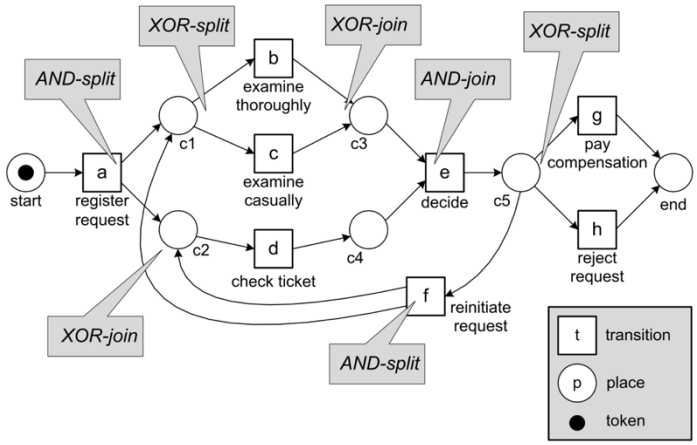

Figura II.2 - Struttura di una Rete di Petri[5]

Questa rappresentazione permette di individuare tutti i possibili percorsi che il processo può svolgere.

Se un'attività è collegata a due rami di uscita significa che le due attività devono essere svolte in parallelo. Mentre, se da un place si diramano due attività, sono da scegliere in maniera esclusiva, cioè o l'una o l'altra.

Lo stato di una rete di Petri è definito dalla distribuzione dei tokens nei place, che è chiamato marking.

Il marking è fondamentalmente un multi-set di places con tokens, che indica quali places contengono tokens in un determinato momento. Il marking iniziale rappresenta lo stato iniziale della rete di Petri, ovvero quali places contengono tokens quando il sistema inizia la sua operazione.

[5] Susana Isabel Do Nascimento Santos - Application of process mining techniques to invoice process

Il comportamento dinamico di una rete di Petri marked è definito dalla cosiddetta "firing rule", o regola di accensione. Una transizione è abilitata se ciascuno dei suoi places di input contiene almeno un token. Una transizione abilitata può attivarsi consumando un token da ciascun place di input e producendo un token per ciascun place di output.

La definizione formale della regola di accensione è la seguente.

Sia (N,M) una rete di Petri marcata con N = (P,T,F) e M ∈ B(P). La transizione t ∈ T è abilitata, denotata come (N,M)[t⟩, se e solo se •t ≤ M. La regola di accensione _ [_⟩ _ ⊆ N × T × N è la più piccola relazione che soddisfa qualsiasi (N, M) ∈ N e qualsiasi t ∈ T, (N,M) [t⟩ ⇒ (N,M) [t⟩ (N,(M\•t) ⊎ t•).[6]

Una rete di Petri marcata gode di alcune proprietà desiderabili:

Una rete di Petri marcata (N, M_0) è limitata a k se nessun place contiene mai più di k tokens.

Formalmente, per ogni p ∈ P e ogni M ∈ [N, M_0⟩: M(p) ≤ k.

Una rete di Petri marcata è *safe* se e solo se ha un limite di 1.

Una rete di Petri marcata è limitata se e solo se esiste un k ∈ N tale che sia k-limitata.

Una rete di Petri marcata (N, M_0) è priva di deadlock se per ogni marking raggiungibile è abilitata almeno una transizione.

Formalmente, per ogni M ∈ [N, M_0⟩ esiste una transizione t ∈ T tale che (N, M) [t⟩.

[6] Wil van der Aalst, Process Mining – Data Science in Action

Una transizione $t \in T$ in una rete di Petri marcata (N, M_0) è attiva se da ogni marking raggiungibile è possibile abilitare t. Formalmente, per ogni $M \in [N, M_0\rangle$ esiste una marcatura $M' \in [N, M\rangle$ tale che (N, M') $[t\rangle$. Una rete di Petri marcata è attiva se ciascuna delle sue transizioni è attiva.

In conclusione, le reti di Petri sono uno strumento potente per la modellazione e l'analisi dei processi e sono ampiamente utilizzate grazie alle loro forti basi teoriche, oltre che per la loro capacità di catturare bene la concorrenza.

2.3 ALGORITMI DI DISCOVERY

Gli algoritmi di discovery sono fondamentali nel process mining in quanto aiutano ad estrarre automaticamente modelli di processo dai dati di log o dagli eventi registrati dai sistemi informativi.

Esempi di algoritmi per l'estrazione di processi sono l'Alpha algorithm, l'algoritmo Heuristic Miner, l'algoritmo Inductive Miner, il Genetic Process Mining e l'algortimo Fuzzy Miner.

L'Alpha Algorithm è uno dei primi algoritmi di discovery sviluppati.

L'obiettivo è l'estrazione dello schema sottoforma di rete di Petri. Il punto di partenza dell'Alpha algorithm è l'estrazione di relazioni, quali la dipendenza temporale, l'indipendenza temporale e l'indipendenza.

La dipendenza temporale si ha quando, ad esempio, l'attività a è seguita da b ma b non è mai seguita da a; perciò, questo significa che b dipende da a.

Nel caso di indipendenza temporale, vi sono delle tracce dove a è seguita da b ma anche altre tracce dove b è seguita da a; ciò significa che a e b possono essere eseguite in parallelo.

L'indipendenza implica, invece, che non ci sono tracce dove a è seguita da b o viceversa; perciò, a e b sono indipendenti.

Queste relazioni vengono utilizzate per apprendere i comportamenti nei modelli nel processo.

Il problema fondamentale dell'alpha algorithm è che non tiene conto delle occorrenze rare e mescola le esecuzioni tipiche con gli outliers: basta che ci sia solo un'esecuzione nel log di un'attività che devia dalla norma per far sì che nello schema vengano rappresentate con lo stesso peso sia le tracce frequenti che gli outliers. Questo caso non è molto desiderabile, perché produrrà schemi complessi che non potranno essere filtrati per eventi più frequenti. Gli algoritmi successivi, dunque, hanno cercato di implementare il concetto di frequenza.

L'Heuristic Miner, quindi, nasce con l'obiettivo di gestire il rumore all'interno dei dati. Utilizza un approccio basato su euristiche per scoprire i modelli di processo, tenendo conto della frequenza degli eventi, ossia il numero di volte in cui all'interno del log avvengono determinate successioni.

Le relazioni di dipendenza pesate per la frequenza della relazione permettono 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 che rappresentano outliers: introdurre questa misura numerica permette anche di semplificare il processo, in modo da potersi concentrare solo sulle attività più frequenti.

L'algoritmo Inductive Miner è basato su un approccio di apprendimento induttivo. La particolarità di questo algoritmo è la sua capacità di gestire al meglio il rumore e di estrarre processi che sono *sound*[7] per costruzione; infatti, è un miglioramento sia dell'Alpha Miner che dell'Heuristic Miner.

Un altro algoritmo di discovery è il Genetic Process Mining, una metodologia che si ispira al mondo naturale. Questo approccio si propone di riprodurre i meccanismi di selezione naturale per adattare i modelli di processo ai dati del file log.

L'algoritmo Fuzzy Miner, infine, viene utilizzato per estrarre un modello di processo gerarchico che riflette la realtà dei dati, anche quando essi sono non strutturati.

[7] Si definisce "sound" una rete che è priva di anomalie, ossia in ogni place non vi possono essere più tokens, ed ognuno di essi può raggiungere tutti i places; perciò, non ci sono punti definiti "deadlock".

Nel paragrafo seguente si andrà ad approfondire una delle tecniche sopracitate, ossia l'algoritmo Inductive Miner, perché sarà utilizzato per l'estrazione dei processi nelle analisi successive.

2.3.1 Inductive Miner

L'algoritmo Inductive Miner è attualmente uno dei principali approcci utilizzati per l'estrazione dei processi grazie alla sua flessibilità, garanzie formali e scalabilità.
L'idea di base che risiede dietro l'algoritmo Inductive Miner è, come per tutti gli algoritmi di process discovery, quella di estrarre un processo dai dati degli eventi.
Ciò che contraddistingue l'Inductive Miner è che produce una visualizzazione ad albero, la quale permette una rappresentazione gerarchica di un modello di processo.
Per una maggiore chiarezza riguardo la semantica, di seguito sarà illustrato il significato dei principali operatori: composizione sequenziale, scelta esclusiva, composizione parallela, ciclo, attività silenziosa.

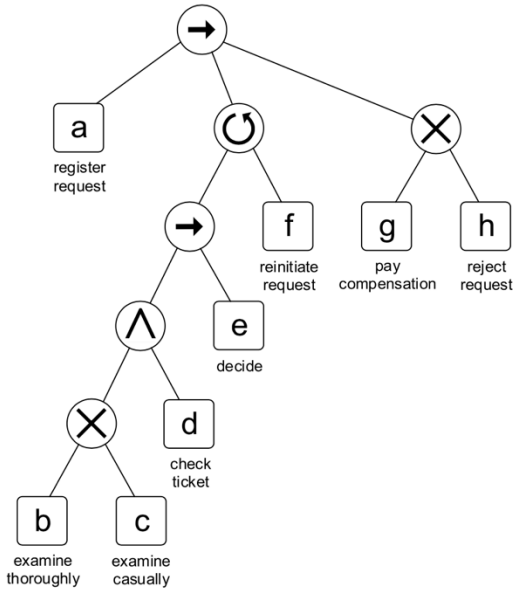

Figura II.3 - Process Tree[8]

L'operatore sequenziale impone di eseguire prima le attività di sinistra e successivamente quelle di destra. Nel caso riportato, come prima attività deve essere eseguita "a", e successivamente le attività collegate al loop.

La scelta esclusiva è l'equivalente dell'operatore "or", ossia permette di scegliere tra una delle attività figlie.

La composizione parallela indica che le attività ad esso collegate devono essere eseguite in parallelo.

L'operatore che indica la presenza di un loop si utilizza se vi sono almeno due attività figlie e vengono eseguite prima le attività nella parte sinistra, e

[8] Slides: Wil van der Aalst (www.vdaalst.com)

successivamente si sceglie se eseguire il ramo di destra o meno. Perciò, si eseguono le attività a sinistra almeno una volta, e poi si sceglie se eseguire le altre attività o interrompere il loop.

L'attività silenziosa, invece, permette di bypassare alcune attività.

Per l'interpretazione successiva dei risultati, spesso l'albero ottenuto dall'algoritmo inductive Miner viene poi convertito in Rete di Petri. La traduzione della visualizzazione dall'albero alla rete di Petri è abbastanza semplice.

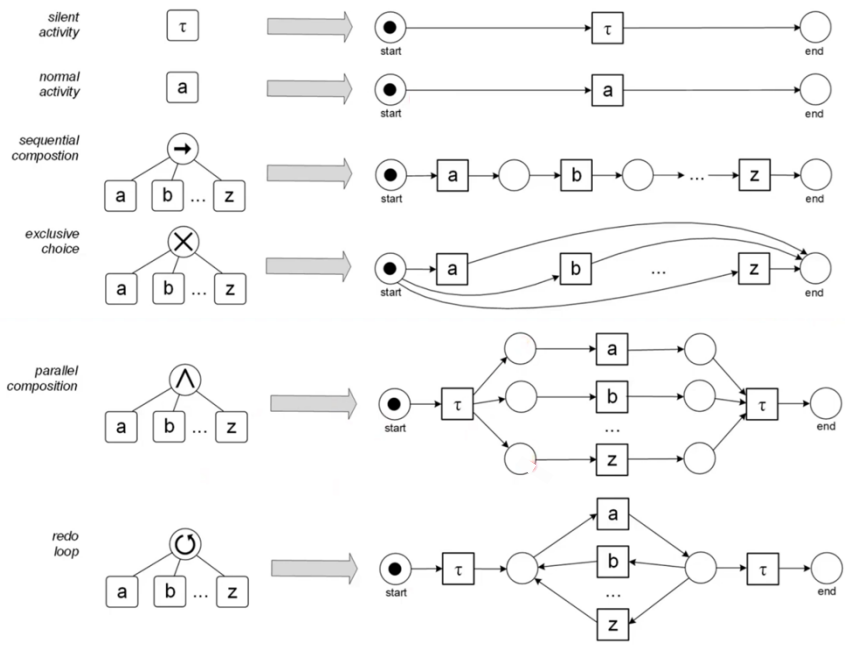

Figura II.4 - Dall'albero di processo alla Rete di Petri[9]

[9] Slides: Wil van der Aalst (www.vdaalst.com)

L’attività silenziosa può essere ricondotta alla transazione silenziosa, che non lascia nessuna traccia nell’event log, mentre l’attività normale corrisponde ad una semplice transazione.

La composizione sequenziale si traduce con delle attività in sequenza. In questo caso le attività sono singole, ma potrebbero anche corrispondere ad un sottoalbero di attività.

La scelta esclusiva, come nel process tree, corrisponde all’esecuzione di una sola delle attività figlie.

Nel caso dell’esecuzione parallela, delle attività atomiche, o sotto-processi, sono eseguiti in parallelo.

Il loop contiene come minimo due attività, in cui la prima attività ha un significato speciale. Il numero di volte che si esegue la prima attività è una in più rispetto alla somma delle volte che tutte le altre attività vengono eseguite.

L’albero di processo viene estratto dall’iterazione della computazione di Directly-follows graphs. Un Directly-Follows Graph (DFG) è la rappresentazione più semplice dei modelli di processo. In un DFG ogni nodo rappresenta un'attività e gli archi rappresentano la relazione tra le varie attività. Tipicamente in un modello di processo, il directly-follows graph ha un'origine, “source” e und una fine, “sink”, le quali rappresentano le attività iniziali e finali. Un arco tra due attività indica che l'attività di origine è seguita direttamente dall'attività finale nell’event log.

Formalizzando, considerando un event log L, dove $L \in \mathbb{B}(A^*)$, il directly-follows graph di L è $G(L) = A_L, \rightarrow_L, A_L^{start}, A_L^{end})$ dove:

$A_L = \{a \in s \mid s \in L\}$ è il set di attività in L,

$\rightarrow_L = \{(a,b) \in A \times A \mid a >_L b\}$ è la relazione di successione,

$A_L^{start} = \{a \in A \mid \exists_{s \in L} a = first(\sigma)\}$ è il set delle attività iniziali,

$A_L^{end} = \{a \in A \mid \exists_{s \in L} a = last(\sigma)\}$ è il set delle attività finali[10].

L'Inductive Miner, perciò, divide iterativamente il log iniziale in vari sub-logs. Per ogni sub-log crea un directly-follow graph G(L) e cerca di estrarre determinati pattern partendo dal DFG.

Per trovare il miglior modo di suddividere l'event log in ogni passaggio, si devono prima individuare i "tagli" da effettuare nel directly-follows graph in ogni sub-log che si dovranno splittare. Vi sono differenti tipi di tagli che si possono effettuare: a scelta esclusiva, tagli sequenziali, tagli paralleli e tagli a ciclo ripetuto, corrispondenti ai quattro operatori dell'albero dei processi (x, →, ∧ e ↺).

[10] Wil van der Aalst, Process Mining – Data Science in Action

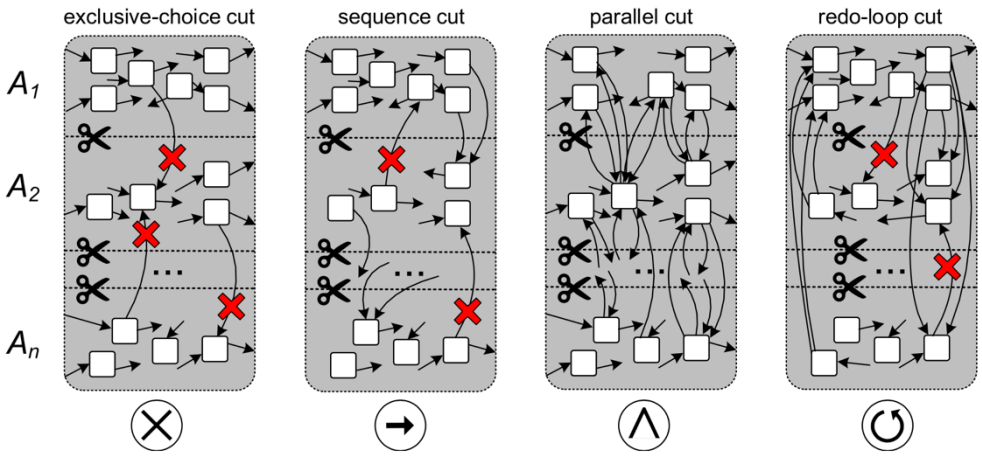

Figura II.5- Tipologie di tagli[11]

Se si è in grado di dividere il DFG in parti che sono totalmente disconnesse tra di loro, allora si parla di "exclusive-choice cut", ossia taglio a scelta esclusiva, nel caso in cui si hanno due attività che sono mutuamente esclusive, perciò nessuna delle due verrà eseguita subito dopo l'altra.

La proprietà che deve essere rispettata nel caso di "exclusive-choice cut" è che se ci sono delle partizioni in diverse attività, A_1, A_2, …, A_n, allora non ci dovrebbero essere relazioni; quindi, non ci dovrebbero essere relazioni di successione tra le attività.

Formalmente, un taglio a scelta esclusiva di G(L) è un taglio (x, A_1, A_2, …, A_n) tale che $\forall_{i,j \in \{1,\ldots,n\}} \forall_{a \in Ai} \forall_{b \in Aj}\ i \neq j \Longrightarrow a \not\rightarrow_L b$.

[11] Wil van der Aalst, Process Mining – Data Science in Action

Il “sequence cut”, o taglio sequenziale, cerca di partizionare il DFG in modo tale che vi siano delle attività che vanno solo in una direzione.

Il taglio sequenziale richiede che due sotto-parti del DFG, A_i e A_j, siano costruite in modo tale che ogni elemento di A_i preceda ogni elemento di A_j.

Perciò, un taglio sequenziale di G(L) è un taglio ($\rightarrow$, A_1, A_2, …, A_n) tale che $\forall_{i,j \in \{1,\dots,n\}} \forall_{a \in Ai} \forall_{b \in Aj}\ i < j \Longrightarrow (a \longmapsto^{+}_{L} b \wedge b \nrightarrow^{+}_{L} a)$.

Viene, invece, identificato come taglio parallelo un taglio effettuato partizionando il DFG in modo tale che si ha un’attività in una sotto-parte del grafo, e l’altra attività in un’altra sotto-parte, ed entrambe possono essere seguite dall’altra in entrambe le direzioni. Ogni attività in un subset deve essere seguita da ogni attività dell’altro subset. Inoltre, ogni attività nel subset deve poter fungere da attività iniziale e finale. Perciò, tra tutti gli elementi di questi due subset ci dovrebbe essere una relazione di successione in entrambe le direzioni. Inoltre, ogni set di attività dovrebbe avere un inizio ed una fine.

Precisamente, un taglio parallelo di G(L) è un taglio ($\wedge$, A_1, A_2, …, A_n) tale che:

$\forall_{i \in \{1,\dots,n\}} A_i \cap A_L{}^{start} \neq \emptyset \wedge A_i \cap A_L{}^{end} \neq \emptyset$,

$\forall_{i,j \in \{1,\dots,n\}} \forall_{a \in Ai} \forall_{b \in Aj}\ i \neq j \Longrightarrow a \longmapsto_{L} b$.

Il “taglio” più complesso è il “redo-loop cut”, ossia il taglio a ciclo ripetuto, in cui viene eseguita in primo luogo la parte superiore del grafo, e successivamente vengono eseguite le altre attività figlie, con la possibilità di ritornare alla prima attività per ricominciare il loop.

Una differenza con il taglio parallelo è che nel taglio a ciclo ripetuto solo la prima sotto-parte del DFG deve poter fungere da attività iniziale e finale, mentre nel taglio parallelo questa proprietà deve valere per entrambe le parti.

Il taglio a ciclo ripetuto può essere formalizzato nel seguente modo.

Un taglio a ciclo ripetuto di G(L) è un taglio ($\circlearrowleft, A_1, A_2, \ldots, A_n$) tale che:

- $n \geq 2$,
- $A_L^{start} \cup A_L^{end} \subseteq A_1$,
- $\{a \in A_1 \mid \exists_{i \in \{2,\ldots,n\}} \exists_{b \in A_i} a \longmapsto_L b\} \subseteq A_L^{end}$,
- $\{a \in A_1 \mid \exists_{i \in \{2,\ldots,n\}} \exists_{b \in A_i} b \longmapsto_L a\} \subseteq A_L^{start}$,
- $\forall_{i,j \in \{2,\ldots,n\}} \forall_{a \in A_i} \forall_{b \in A_j} i \neq j \Longrightarrow a \not\longmapsto_L b$,
- $\forall_{i \in \{2,\ldots,n\}} \forall_{b \in A_i} \exists_{a \in A_L^{end}} a \longmapsto_L b \Longrightarrow \forall_{a' \in A_L^{end}} a' \longmapsto_L b$,
- $\forall_{i \in \{2,\ldots,n\}} \forall_{b \in A_i} \exists_{a \in A_L^{start}} b \longmapsto_L a \Longrightarrow \forall_{a' \in A_L^{start}} b \longmapsto_L a'$. [12]

La prima riga indica che ci dovrebbero essere almeno due attività figlie. Inoltre, ci devono essere sia archi in entrata che in uscita nel primo set di attività A_1.

Se c'è una connessione tra un'attività del primo set ed un'altra appartenente ad un altro set, la prima dovrebbe avere anche un altro arco in uscita, in modo tale che si è in grado di uscire dal loop. Simmetricamente, per ogni attività di qualsiasi altro set connessa ad un'attività del primo set, quest'ultima dovrebbe avere anche un arco in entrata.

[12] Wil van der Aalst, Process Mining – Data Science in Action

Il quinto punto nella lista delle proprietà che dovrebbero essere rispettate per far sì che si tratti di un taglio a ciclo ripetuto tratta di scelta mutuamente esclusiva, ossia attività appartenenti a set diversi non possono essere seguite una dall'altra.

Se l'attività b appartiene ad un set diverso dal primo, e questa ha una relazione di successione con una delle attività finali, allora tutte le attività finali dovrebbero avere una connessione con b. Simmetricamente, se da un'attività che appartiene ad un set diverso dal primo si può arrivare ad un'attività del primo set, dalla prima attività dovrebbe essere possibile ricollegarsi ad ogni attività iniziale.

Le quattro tipologie di tagli sopra descritti si basano sulle caratteristiche dei quattro operatori dell'albero dei processi, presupponendo che non vi siano attività duplicate o silenziose.

In sintesi, l'algoritmo Inductive Miner è caratterizzato dal seguente funzionamento. Dato un event log, viene costruito il directly-follows graph e vengono eseguiti i tagli in base alle caratteristiche del DFG.

Dopo aver suddiviso l'event log in sotto-logs, la procedura viene ripetuta fino a raggiungere un sotto-log con una sola attività. Il modo in cui l'event log viene suddiviso in sotto-logs dipende dall'operatore, mentre le tracce vuote vengono gestite in maniera differente (in base all'operatore) e comportano l'inserimento di attività silenziose τ.

Le parti che non si riescono a splittare vengono visualizzate come un "flower model", ossia un modello molto complesso e poco sintetizzato nel quale tutte le

tracce possono accadere e non si riesce ad individuare nessun comportamento particolare.

Esistono, tuttavia, varianti dell'algoritmo che sono in grado di gestire il rumore, come l'Infrequent Inductive Miner. L'IMi permette all'utente di definire valore soglia k compreso tra 0 e 1, per separare il comportamento frequente da quello poco frequente.

2.3.2 Proprietà dell'algoritmo Inductive Miner

Il principale punto di forza dell'Inductive Miner è rappresentato dalle garanzie formali che offre. I modelli generati dall'algoritmo, infatti, l'algoritmo IM produce sempre un modello di processo *sound* in grado di riprodurre l'intero event log. A differenza di molti altri algoritmi, la *fitness* è garantita. Poiché i modelli sono strutturati a blocchi e le attività non sono duplicate, i modelli tendono ad essere semplici e generali, perciò il rischio di *overfitting* è minore.

Inoltre, è garantito che il modello è in grado di riprodurre interamente l'event log da cui è stato generato.

Infine, va sottolineata la scalabilità dell'algoritmo. Questo significa che può gestire grandi quantità di dati senza perdere efficacia. In un contesto in cui i dati possono essere sempre più vasti e complessi, questa caratteristica è di fondamentale importanza per garantire che l'analisi dei processi rimanga efficiente e accurata.

La proprietà per cui l'inductive Miner si distingue, però, è la sua capacità di gestire i comportamenti infrequenti, in modo tale da diminuire il rumore.

Complessivamente, l'Inductive Miner offre una combinazione di flessibilità, garanzie formali e scalabilità, rendendolo uno dei principali metodi di analisi dei processi in diverse applicazioni e settori.

2.3.3 Come si valuta un buon modello?

La sfida nella scoperta dei processi è trovare il modello di processo "migliore" dato l'event log considerato. Quale modello di processo sia migliore viene generalmente valutato utilizzando diversi criteri di qualità.

Quattro importanti criteri di qualità sono la "Fitness", la "Generalizzazione", la "Precisione" e la "Semplicità".

Per "Fitness" si intende la capacità di descrivere correttamente le varie tracce. Si ha un'elevata fitness quando lo schema ottenuto si adatta bene alle tracce nel log. Una buona "Generalizzazione" significa che il modello non soffre di *overfitting*. Tuttavia, bisogna evitare un'elevata generalizzazione ma poca precisione, altrimenti si rischia che il modello non descriva in maniera adeguata la realtà, ossia il modello può soffrire di *underfitting*. Quando il modello ha una precisione elevata, si rischia la situazione opposta, ovvero che il modello descrive i dati, ma c'è il

rischio che abbia performance elevate solo con file log specifico, si tratterebbe perciò di overfitting.

Per precisione si intende la capacità del modello di non rappresentare comportamenti completamente diversi da quelli visti nell'event log.

Inoltre, il quarto criterio di "bontà" di un modello è la semplicità, ossia si vuole che l'algoritmo estragga la versione più semplice del modello per favorire la sua interpretazione.

2.4 TECNICHE DI MACHINE LEARNING

2.4.1 Regressione Logistica

L'analisi della regressione è una tecnica statistica utilizzata per comprendere la relazione tra una variabile dipendente continua (solitamente indicata come Y o variabile risposta) e una o più variabili indipendenti (spesso indicate come X_1, X_2, ..., X_k ossia le variabili esplicative o predittori). L'obiettivo principale dell'analisi di regressione è cercare di modellare e comprendere il legame tra queste variabili.

La relazione tra la variabile risposta e le variabili esplicative è espressa attraverso un modello matematico. Nella forma generale, questo modello è rappresentato come $Y = f(X_1, X_2, ..., X_k) + \varepsilon$, dove ε rappresenta l'errore casuale o residuo che cattura la variazione non spiegata dal modello.

Il legame teorico nel modello lineare tra la variabile dipendente Y e i regressori X_1, X_2, ..., X_k è dato dalla media condizionata di Y, dati i valori dei regressori.

Il modello di regressione lineare multipla è espresso dalla seguente formulazione:

$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$.

Dove β_0 è detto termine noto, mentre $\beta_1, ..., \beta_k$ sono detti coefficienti di regressione e, insieme alla varianza dell'errore, sono i parametri del modello da stimare sulla base delle osservazioni campionarie.

L'obiettivo è stimare i coefficienti di regressione ($\beta_0, \beta_1, ..., \beta_k$) che quantificano l'effetto delle variabili indipendenti sulla variabile dipendente. Questi coefficienti indicano quanto cambia la variabile dipendente quando le variabili indipendenti cambiano di una quantità unitaria, mantenendo le altre costanti. Perciò, i coefficienti danno l'idea dell'impatto delle singole variabili sulla variazione della variabile risposta Y: il coefficiente cambia tanto meno quanto le variabili introdotte sono indipendenti una dalle altre.

Il modello di regressione lineare, tuttavia, risulta inadeguato quando la variabile risposta è dicotomica, come vero-falso o 1-0, perché in questo modo si violano le ipotesi sottostanti al modello lineare: perciò si utilizza il modello di Regressione Logistica. Questa tecnica è particolarmente utile quando si desidera studiare o prevedere un risultato binario in base a una serie di variabili indipendenti.

La variabile dicotomica y indica la presenza o l'assenza di una particolare caratteristica, può assumere perciò due valori: 1 con probabilità π e 0 con probabilità $1 - \pi$. La probabilità con la quale Y assume il valore 1 o 0 dipende dalle variabili esplicative x.

Dato che non si può esprimere π come combinazione lineare delle x, perché π varia tra 0 e 1, mentre la combinazione lineare varia tra $-\infty$ e $+\infty$, si sostituisce la variabile dicotomica con *l'odds,* ossia il rapporto tra la probabilità che si verifichi l'evento favorevole e quella che non si verifichi.

Il modello di regressione logistica può essere perciò descritto con la seguente formulazione: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$.

La funzione logaritmica lega π_i a x_i e permette di esprimere questa relazione come una combinazione lineare.

L'effetto dei β sulla variabile risposta y dipende dal coefficiente della variabile per la probabilità del successo su quella dell'insuccesso: se β è positivo significa che la variabile x gioca a favore del successo, al contrario, se β è negativo riduce la probabilità di successo.

In definitiva, la regressione logistica è uno strumento utilizzato per affrontare problemi di classificazione binaria in cui è necessario prevedere o comprendere il comportamento binario di una variabile risposta y.

2.4.2 Support Vector Machine SVM

Support Vector Machine appartiene alla famiglia di modelli denominati “large margin classifier”, i quali hanno l’obiettivo di massimizzare il margine intorno all’iperpiano separatore.

Per comprendere a fondo questa tecnica di machine learning è necessario prima fornire la definizione di miglior bordo decisionale.

Il miglior bordo decisionale è quello che massimizza la distanza tra i punti delle classi: infatti, se ci fosse un bordo vicino ai punti di una delle due classi, sarebbe più alta la probabilità di errore nelle applicazioni reali, perché con una piccola variazione dei dati rispetto al training set si potrebbe missclassificare un elemento. Perciò, un buon bordo è quello che è a maggior distanza dai punti delle due classi.

Notoriamente, le tecniche di machine learning si possono dividere in base all’obiettivo per il quale vengono utilizzate, ossia per regressione o classificazione. Se l’obiettivo dell’analisi è la regressione si utilizza Support Vector Regression, mentre se l’analisi è volta alla classificazione si utilizza Support Vector Machine.

Il Support Vector Machine (SVM) è un modello di classificazione discriminante che apprende i bordi decisionali lineari o non lineari nello spazio degli attributi per separare le classi. Oltre a massimizzare la separabilità delle due classi, l’SVM è in grado di controllare la complessità del modello al fine di garantire buone prestazioni di generalizzazione.

Grazie alla sua capacità di regolarizzare in modo innato il suo apprendimento, SVM è in grado di apprendere modelli altamente espressivi senza soffrire di overfitting. Ha quindi ricevuto una notevole attenzione nella comunità del machine learning ed è comunemente utilizzato in diverse applicazioni pratiche, che vanno dal riconoscimento delle scritte a mano alla categorizzazione del testo.

Un altro aspetto unico dell'SVM è che rappresenta il bordo decisionale utilizzando solo un sottoinsieme degli esempi di training più difficili da classificare, noti come vettori di supporto. Quindi, è un modello discriminativo che è influenzato solo dalle istanze di training vicino al bordo delle due classi, cioè i Support Vectors (SV).

La tecnica Support Vector Machine nasce dall'ottimizzazione matematica ed è utilizzata per la classificazione binaria. Le due classi sono identificate con +1 e -1. Dato un training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, x_i rappresenta l'istanza e y_i è la classe che assume valore +1 e -1.

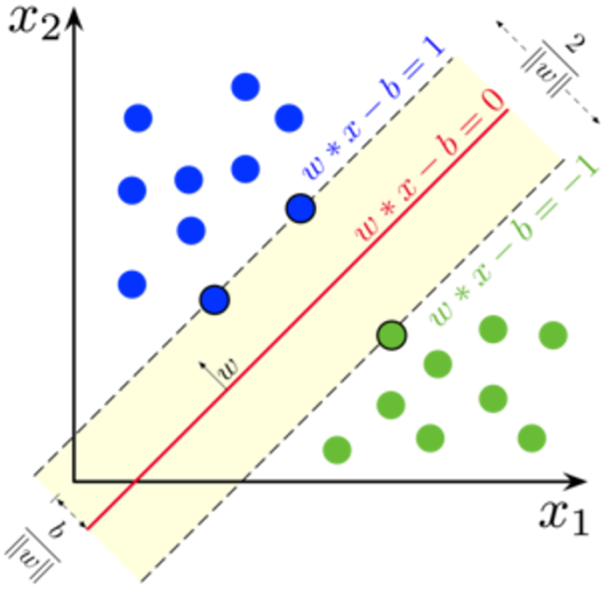

Figura II.6 - Support Vector Machine[13]

I dati bidimensionali linearmente separabili possono essere separati dall'iperpiano, o da una retta, a seconda della dimensionalità che si sta considerando.

L'iperpiano separatore è costruito dalla formula $w^t x + b = 0$, dove x rappresenta gli attributi e w e b rappresentano i parametri dell'iperpiano: w è il vettore perpendicolare al piano che da l'inclinazione, mentre b dà l'asse, ossia quale retta si deve considerare tra tutte le possibili rette che sono perpendicolari al vettore w.

Qualunque punto x_i che appartiene all'iperpiano restituirà un valore zero di $f(x_i)$. Quando i punti si trovano fuori dall'iperpiano si ottengono valori maggiore di zero

[13] Fonte: Wikipedia - https://upload.wikimedia.org/wikipedia/commons/thumb/7/72/SVM_margin.png/600px-SVM_margin.png

(sopra l'iperpiano) o minori di zero (sotto l'iperpiano). Ai fini della classificazione binaria, l'obiettivo è trovare un iperpiano che collochi le istanze di entrambe le classi su lati opposti dell'iperpiano, risultando così in una separazione delle due classi.

I support vectors sono le istanze più vicine al bordo decisionale. L'obiettivo è avere il margine, ρ, il più grande possibile. Quindi tra tutti i possibili iperpiani che separano bisogna scegliere quello che massimizza ρ. Perciò il problema di ottimizzazione matematica consiste nel trovare w e b tali che il margine ρ sia il più grande possibile per ogni elemento del training set (x_i, y_i).

L'ipotesi di partenza dell'SVM è che le classi siano linearmente separabili: la formulazione può essere modificata per apprendere un iperpiano di separazione che tollera un piccolo numero di errori di training utilizzando un metodo noto come *soft margin classification*, il quale consente all'SVM di apprendere iperpiani lineari anche in situazioni in cui le classi non sono linearmente separabili.

Per far ciò, si deve introdurre flessibilità nel vincolo: è possibile accettare come validi alcuni dati classificati erroneamente, ma vicini al bordo, dove "quanto vicino" dipende dalle variabili slack ξ_i. Perciò, in presenza di classi linearmente separabili a meno di alcuni punti, vengono introdotte le variabili di rilassamento (slack variables) $\boldsymbol{\xi}_i$ al vincolo, per ogni istanza del training x_i.

Nel problema di ottimizzazione si nota anche la presenza di un parametro C, il quale è un parametro di regolarizzazione che rappresenta il costo di misclassificazione

delle istanze di training. È perciò un modo per controllare l'overfitting, perché durante l'allenamento del modello non si tiene conto di alcuni punti di rumore, perciò aumenta la generalità del modello.

Nel caso di classi non linearmente separabili, invece, per riuscire a classificare i dati si trasforma lo spazio in modo tale da rendere le classi linearmente separabili. L'idea di base è trasformare i dati dal suo spazio originale degli attributi x in un nuovo spazio $\varphi(x)$ in modo che un iperpiano lineare possa essere utilizzato per separare le istanze nello spazio trasformato, utilizzando l'SVM. L'iperpiano può quindi essere proiettato indietro nello spazio degli attributi originale, risultando in un bordo decisionale non lineare.

In questo contesto, vengono utilizzati i *kernel trick*: consistono nell'applicazione di una trasformazione non lineare ai dati di input per proiettarli in uno spazio delle caratteristiche ad alta dimensionalità, dove la separazione lineare dei dati è più semplice. Questa trasformazione può essere costosa computazionalmente o addirittura impossibile da calcolare direttamente. Tuttavia, il kernel trick consente di calcolare il prodotto scalare tra i vettori trasformati senza effettuare esplicitamente la trasformazione, risparmiando così tempo e risorse computazionali. Le più comuni funzioni di kernel nell'implementazione di SVM sono il Kernel lineare, polinomiale e gaussiano.

Il Kernel lineare è utilizzato per problemi di classificazione lineare, significa non spostarsi in nessun spazio superiore; perciò, le classi sono linearmente separabili e non vanno fatte trasformazioni, si utilizza lo spazio di partenza.

Il Kernel polinomiale è utile quando i dati hanno una struttura non lineare.

Il Kernel gaussiano (RBF - Radial Basis Function) è utilizzato per approssimare bordi decisionali complessi.

In conclusione, è possibile affermare che SVM è un potente algoritmo che fornisce un modo efficace per regolarizzare i parametri del modello massimizzando il margine del bordo decisionale ed è in grado di creare un equilibrio tra la complessità del modello e gli errori di training utilizzando C.

Inoltre, sebbene il tempo di addestramento di un modello SVM possa essere elevato, i parametri possono essere rappresentati utilizzando un piccolo numero di vettori di supporto, rendendo la classificazione delle istanze di test abbastanza veloce.

2.4.3 Metriche per la valutazione delle performance del modello

Per la valutazione delle performance di un modello di classificazione è pratica comune utilizzare l'Accuracy come metrica, la quale misura il numero di dati classificati correttamente sul totale delle osservazioni. Perciò, l'Accuracy (accuratezza) rappresenta la frazione delle previsioni corrette fatte dal modello

rispetto al numero totale di previsioni effettuate. In altre parole, misura quanto il modello sia in grado di classificare correttamente le osservazioni.

L'accuracy è calcolata utilizzando la seguente formula:

$$\text{Accuracy} = \frac{Numero\ di\ previsioni\ corrette}{Numero\ totale\ di\ previsioni}$$

Tuttavia, non in tutti i casi può essere considerata una buona metrica: ipotizzando di avere 100 campioni, con uno sbilanciamento delle classi di 90 e 10, gli errori complessivamente saranno pochi perché il modello classificherà bene la maggior parte degli elementi della classe maggioritaria, ma non riuscirà a prevedere la classe corretta per nessuno, o pochi, degli elementi della classe minoritaria.

In questo caso, l'accuracy sarà molto alta, ma questo nasconde il fatto che il modello non è in grado di classificare gli elementi della classe minoritaria. Perciò si introducono altre metriche: la Precision, la Recall e F1-Score.

La Precision è una metrica che misura quanto il modello sia accurato quando prevede la classe positiva. Perciò, se il modello classifica un'istanza come classe 1, la precisione indica quanto ci si può fidare di questa classificazione.

La precisione è definita come la frazione delle previsioni positive corrette rispetto al numero totale di previsioni positive calcolate dal modello.

La formula per calcolare la precisione è la seguente:

$$\text{Precision} = \frac{Veri\ positivi}{Veri\ positivi + Falsi\ positivi}$$

La Recall è definita come la frazione delle previsioni positive corrette calcolate dal modello rispetto al numero totale di esempi effettivamente positivi nel dataset. Misura quanto bene il modello sia in grado di individuare tutte le istanze positive presenti nel dataset. Indica, perciò, sulle istanze che sono realmente classe 1, quante sono state correttamente classificate. La formula per calcolare la recall è la seguente:

$$\text{Recall} = \frac{Veri\ positivi}{Veri\ positivi + Falsi\ negativi}$$

Per evitare il problema di avere due metriche, si utilizza F1-score. Questa metrica combina la precision e la recall in un unico valore numerico che rappresenta la capacità del modello di classificare correttamente le istanze positive e di evitare falsi positivi. F1-score, perciò, è una media armonica della precision e la recall.

Inoltre, una ulteriore metrica per la valutazione delle performance di un modello di classificazione è l'area sottostante alla curva ROC, Area Under the Curve (AUC). L'AUC misura la capacità del modello di classificare correttamente gli esempi positivi rispetto a quelli negativi: un valore di AUC alto indica che il modello riesce a separare bene le classi.

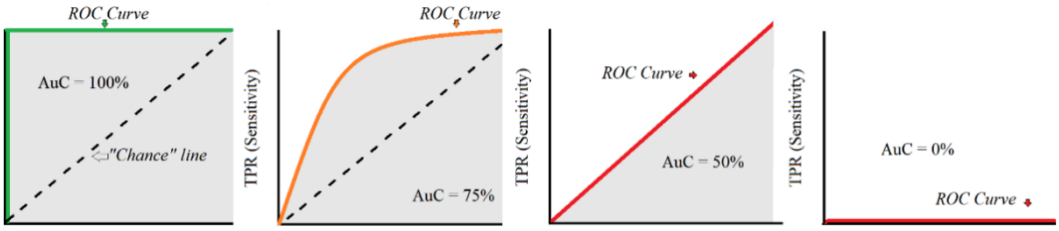

Figura II.7 - Curva ROC[14]

Perciò, il modello "perfetto" avrà un'area al di sotto della curva ROC pari ad 1. Tipicamente, nella visualizzazione si inserisce una retta diagonale, che rappresenta il comportamento di un classificatore random. I punti sopra questa retta rappresentano le prestazioni migliori rispetto al modello standard, mentre i punti al di sotto rappresentano prestazioni peggiori. Perciò, i valori dell'area sottostante la curva superiori a 0.5 indicano che il modello è meglio di un classificatore standard. In definitiva, la valutazione dei modelli di classificazione che seguiranno è stata effettuata guardando alle seguenti metriche: F1 score in primis, ma anche alla precision e recall separatamente per un'analisi delle performance del classificatore più approfondita, ed in seguito si è osservata anche l'area sottostante la curva ROC per maggiore completezza.

[14] Fonte: Data Science Central

III. METODOLOGIA

Sulla base della teoria descritta precedentemente si svilupperanno le analisi future. L'obiettivo di questa tesi è quello di estrarre delle informazioni utili sulle carriere degli studenti, ed in particolare su quegli esami che sono un potenziale collo di bottiglia (ad esempio, hanno un basso tasso di superamento). L'obiettivo è quello di estrarre informazioni sugli sforzi effettivi degli studenti nella preparazione di questi esami. Per perseguire quest'obiettivo, verranno utilizzate le informazioni relative al momento in cui lo studente inizia e termina di studiare per un esame, tenendo conto di quante volte è necessario sostenere nuovamente un esame prima di superarlo. Le domande principali che hanno motivato lo sviluppo di questa tesi sono: Quali corsi sono considerati più impegnativi o difficili dagli studenti? Quali sono i colli di bottiglia?

Per rispondere a queste domande bisogna iniziare dalla disamina di ogni esame nel dettaglio, considerando il tasso di superamento, di bocciati, di assenti e ritirati, oltre che la percentuale di studenti che ha sostenuto l'esame entro la fine del semestre per il quale il corso era stato programmato.

Verranno poi estratti i processi effettuati dagli studenti relativamente ad ogni materia, con un focus particolare sugli esami considerati un ostacolo durante la carriera accademica dello studente.

Inoltre, per approfondire ulteriormente l'analisi, si andranno ad individuare quali sono i principali fattori che influenzano il rendimento scolastico degli studenti, e se quest'ultimo è influenzato dall'aver sostenuto gli esami critici individuati in precedenza. Per fare ciò, verrà addestrato un modello di regressione logistica e SVM per predire se la media del voto finale per ciascuno studente sarà alta o meno, in base a quali esami del primo anno ha sostenuto e quanto tempo ha impiegato per prepararli.

Durante le analisi sviluppate nel corso di questa tesi si può notare un'attenzione maggiore sugli esami del primo anno, perché sono considerati i più critici per la carriera dello studente. Il primo anno universitario, difatti, risulta particolarmente rilevante in quanto getta le basi della formazione dello studente, il cui approccio iniziale è determinante per l'intera durata degli studi.

3.1 PRE-PROCESSING DEI DATI

Come descritto nel primo capitolo, i dati sono stati scaricati dalla piattaforma online ESSE3. Le fasi di pre-processing dei dati sono state effettuate tramite due tool diversi: una prima pulizia è stata effettuata tramite phpMyAdmin, e successivamente i dati sono stati importati e trasformati su Python.

PhpMyAdmin è uno strumento online che permette di gestire database MySQL.

Inizialmente, per ottenere tutte le informazioni del caso, è stato effettuato un *join* con due tabelle. L'operazione di join permette di unire i dati di due o più tabelle diverse.

Tre sono le tabelle di partenza: la prima contiene le informazioni riguardo agli studenti e agli esami dati; la seconda è costituita da tre colonne per identificare in modo univoco ogni studente; e la terza tabella riguarda solamente gli studenti laureati.

La prima tabella contiene le informazioni principali, ossia quelle relative agli esami effettuati, a cui sono state aggiunti i dati per identificare gli studenti, e grazie alla terza tabella è stato possibile creare un sottoinsieme contenente solamente gli studenti già laureati. Infatti, vengono considerati solo i curricula completi, scartando gli studenti che non hanno ancora conseguito la laurea.

Una volta ottenuto il dataset completo, l'analisi e la pulizia dei dati è stata effettuata tramite il linguaggio di programmazione Python.

Il secondo problema di pre-processing che è stato affrontato riguarda le date. All'inizio, il dataset conteneva diverse date, ciascuna riferita a diverse attività dello studente, precisamente:

- "DATA_APP", ossia la data di iscrizione all'appello;
- "DATA_ESA" la data dell'esame, che presenta il valore "NULL" nel caso in cui lo studente non si sia presentato all'appello o sia stato bocciato; perciò, il valore è presente solo in caso di superamento dell'esame;

- “DATA_ESA_CAR” si riferisce alla data in cui il professore ha caricato il voto, ed anche in questo caso il valore è presente solo in caso di superamento dell’esame e NULL in tutti gli altri;
- “DATA_ESA_VERB” costituisce la data di verbalizzazione. L’esame viene verbalizzato anche quando lo stato dell’attività è “chiuso”, ossia l’esito dell’esame non è positivo, mentre è presente un valore nullo nel caso in cui lo studente si sia solo prenotato all’esame ma non si sia presentato. Inoltre, sono verbalizzati anche voti che non sono stati caricati;
- “DATA_INS” è la data dell’inserimento inserimento, precedente al caricamento e alla verbalizzazione;
- “DATA_MOD” è la data della modifica, nel caso in cui il professore vada a modificare ciò che ha inserito nel registro in precedenza;
- “DATA_VAR_STATO” è la data della variazione dello stato dell’attività, ad esempio da “prenotato”, quando lo studente non si è presentato all’esame in un appello, a “Caricato” in un appello successivo.

L’obiettivo è quello di ottenere un dataset contenente solo una data per ciascuna attività dello studente, in modo tale da facilitare la successiva creazione del file log, oltre che per una maggiore chiarezza e semplificazione. Per far ciò, è stata creata una colonna “Date”: quando l’esame risulta caricato la data corrispondente è la data dell’esame, perché è il giorno dell’appello in cui lo studente viene promosso,

mentre in tutti gli altri casi, quando l'esito dell'esame è negativo oppure lo studente è assente, si prende in considerazione la data di prenotazione all'appello.

Lo stato dell'attività sopra menzionato, si riferisce ad una specifica colonna del dataset, ossia "STA_REG_DES", che contiene cinque valori diversi che vanno ad identificare l'attività svolta dallo studente. I suddetti valori sono: caricato, verbalizzato, chiuso, prenotato e annullato.

Il valore "caricato" sta ad indicare che l'esame è stato superato con un voto, mentre "verbalizzato" corrisponde ad un esame non superato. Nel momento in cui il valore corrispondente per un determinato studente ed esame è "chiuso", significa che lo studente in considerazione si è ritirato durante l'esame o era assente, ed infine si ha il valore "prenotato" quando si è solamente prenotato all'esame ma non si è presentato all'appello.

Il valore "annullato" è ambiguo, perché un esame può essere annullato sia prima del caricamento del voto sia successivamente; perciò, questo valore corrisponde sia ad un esame passato che no.

È stata eseguita un'ulteriore pulizia dei dati per la gestione delle incongruenze e del rumore; perciò, i casi in cui l'esame era "annullato" e "verbalizzato" sono stati eliminati, e gli altri valori sono stati organizzati in maniera diversa.

Il valore "caricato" è stato semplicemente sostituito con "promosso" per una maggiore chiarezza. Il valore "chiuso" corrisponde ai casi "assente", "respinto/ritirato" e "bocciato"; perciò, per essere più precisi è stato suddiviso nelle

tre casistiche, in base ai valori corrispondenti delle altre variabili. Infine, il valore "prenotato" è stato sostituito con assente.

Dopo le modifiche appena descritte, le attività degli studenti sono le seguenti: "promosso", "assente", "ritirato", "bocciato".

Per l'estrazione di informazioni più immediata, e per suddividere successivamente i dati, sono state create altre due colonne con l'informazione dell'anno e del semestre in cui sono stati dati gli esami considerati.

Durante la creazione della colonna "Anno" sono stati classificati come esami dati nel primo anno gli esami per i quali dal giorno dell'iscrizione dello studente al giorno dell'esame sono passati meno di 365 giorni. Ugualmente, è stato inserito "2" nella colonna "Anno" se dal giorno dell'iscrizione al giorno dell'esame sono passati più di 365 giorni e meno di 730, mentre per gli esami del terzo anno si è inserito "3" quando il giorno dell'esame superava la data di iscrizione di più di 730 giorni.

Per individuare il semestre, invece, si è utilizzata l'informazione riguardo il mese in cui l'esame è stato svolto: da ottobre a marzo è stato considerato primo semestre e da aprile a settembre secondo semestre.

Durante il corso dell'analisi sono emerse delle incongruenze nei dati, che sono state successivamente risolte con operazioni di pre-processing. Nonostante si stesse considerando solo la porzione di dati relativa agli studenti laureati, per alcuni di essi risultava che per alcuni esami la loro ultima attività è "bocciato", "prenotato" o "assente", perciò risultavano laureati anche se non avevano completato con

successo alcuni esami obbligatori per il loro corso di studi. Si trattava, perciò, di un errore nei dati, perché in alcuni casi lo studente aveva superato l'esame in passato, ma si era prenotato comunque ad un appello successivo, e per questo motivo l'ultima attività risulta "prenotato". In altri casi, lo studente aveva svolto l'esame all'estero durante un programma di Erasmus e perciò il superamento di quel corso non risultava dalla piattaforma.

Per ridurre rumore ed incongruenze, perciò, questi casi sono stati risolti eliminando, per ogni studente e per ogni esame, tutte le righe cronologicamente successive all'attività "promosso".

Infine, sempre in previsione della creazione successiva del file log, sono stati inseriti alcuni riferimenti temporali relativi alle date di fine di ogni semestre e anno. Quest'operazione di pre-processing ha richiesto l'integrazione dei dati sopra-descritti con altre fonti informative, al fine di identificare correttamente il giorno di fine di ogni semestre, che varia di anno in anno.

Al termine, perciò, al dataset sono state aggiunte, per ogni studente, altre attività come: "End first semester", "End first year", "End third semester" e "End second year". Mancano, tuttavia, le informazioni relative al termine del quinto semestre e del terzo anno, perché tutta l'analisi è volta allo studio delle carriere degli studenti, e dato che per il terzo anno il curriculum prevede molti insegnamenti opzionali è difficile generalizzare perché ogni studente effettua un percorso diverso.

Il set di dati finale è una tabella con i seguenti attributi: ID dello studente, nome del corso, data dell'esame, voto dell'esame, tempo necessario per conseguire la laurea definito in giorni, l'anno di iscrizione al corso di laurea, lo stato dell'attività dello studente, l'anno in cui è stato svolto l'esame ed il semestre.

Ad ogni studente è stato associato un identificativo univoco fittizio, per renderli anonimi ed evitare problemi di privacy. La data in cui è stato svolto l'esame e la data di iscrizione dello studente al corso di studi sono caratterizzate dal formato AAAA-MM-GG. Per quanto riguarda la data di immatricolazione, l'informazione iniziale era relativa solamente all'anno, ma per queste analisi è stata fissata arbitrariamente al 1° ottobre dell'anno accademico di iscrizione. Il voto dell'esame è un numero intero compreso tra 18 e 30. Per stato dell'attività dello studente ci si riferisce, come indicato precedentemente, ai valori "promosso", "bocciato", "assente", "ritirato". Infine, il tempo necessario per laurearsi è calcolato come il numero di giorni che intercorrono tra la data dell'esame finale, ossia della discussione della tesi, e la data di iscrizione al corso di studi.

Di seguito un estratto del dataset così formato:

STU_ID	Esame	Date	Voto	Tempo_laurea	AA_ISCR	Attività	Anno	Semestre
F1E6F..A117B	FISICA GENERALE I	2018-03-26	24.0	1171	2017-10-01	Promosso	1	1
F67C6..F1983	ALGEBRA E LOGICA	2020-07-24	26.0	1123	2018-10-01	Promosso	2	2
FC80E..AFC63	ANALISI MATEMATICA 1	2016-01-14	NaN	989	2015-10-01	Bocciato	1	1

Tabella III.1 - Estratto del dataset

3.2 STATISTICHE

All’estrazione dei processi per ogni materia è preceduta un’attenta analisi per ogni corso, relativamente alla percentuale di studenti assenti, bocciati e ritirati per ciascun esame sul totale degli studenti che hanno superato l'esame. Lo scopo di questa analisi è identificare quale esame rappresenta un blocco per gli studenti, al fine di eliminare le difficoltà che si incontrano durante il percorso, le quali spesso sono causa dell’abbandono degli studi.

In primo luogo, è stata creata una tabella raggruppando i dati per ogni esame ed attività, in modo tale da avere per ogni esame il numero esatto di persone promosse, assenti, ritirate e bocciate. Successivamente, è stata calcolata la percentuale sul totale degli studenti che hanno effettuato ogni esame.

L’analisi viene dunque approfondita inserendo la dimensione temporale, ossia individuando in quale semestre gli studenti hanno superato gli esami e la percentuale di coloro che hanno sostenuto gli esami nel periodo giusto, dove per periodo giusto si intende il semestre indicato nel piano di studi di ogni anno accademico.

Perciò, dal set di dati originale è stata estratta solo la porzione di dati che faceva riferimento all’attività “promosso”. Successivamente, si è effettuata una ulteriore divisione in due dataset tra gli studenti iscritti negli anni accademici 2015/2016 - 2016/2017 e quelli iscritti dal 2017/2018 in poi, perché il curriculum previsto dalla Facoltà è leggermente diverso.

In primo luogo, per la definizione della suddetta tabella, sono state create le colonne "End first semester", "End first year", "End third semester", "End second year" e "After second year", e per ogni riga è stato inserito "1" nella colonna giusta in base a quando l'esame preso in considerazione era stato fatto e "0" nelle altre. In altre parole, se nella riga il valore "Anno" era uguale ad "1", così come nella colonna "Semestre", allora l'esame è considerato effettuato prima della fine del primo semestre, e così via.

In seguito, si è raggruppato per esame in modo tale da sommare i valori per ciascuno studente ed avere il numero esatto di alunni che hanno effettuato l'esame prima della fine di ogni periodo, tenendo in considerazione solo gli esami obbligatori del primo e secondo anno.

Il passo successivo di quest'analisi è stata creare la colonna "Right period" ed inserire nella stessa le informazioni relative a quando ciascun esame era previsto nel piano di studi, e successivamente creare una ulteriore colonna con la percentuale di studenti che effettivamente ha passato l'esame nel periodo previsto.

Ciascuna di queste statistiche è stata calcolata sia per gli studenti nella loro totalità, che per ciascuna delle categorie in cui sono stati suddivisi, ossia gli studenti che si laureano in tempo, coloro che completano gli studi con un anno di ritardo e chi con oltre un anno.

3.3 PROCESSI

Come affermato nel Capitolo 2, la creazione dell'event log è propedeutica all'estrazione dei processi.

Perciò, l'obiettivo da perseguire è avere un event log così formato: Case ID, Activity, Timestamp.

Per ottenere la colonna delle attività, per ogni riga sono state concatenate le stringhe relative al nome dell'esame e dello stato dell'attività: perciò, se per l'esame "FISICA GENERALE II" lo stato dell'attività dello studente è "Bocciato", l'attività nell'event log sarà "FISICA GENERALE II-Bocciato".

L'event log risultante, quindi, sarà il seguente: ogni studente rappresenta un caso, e le attività per ogni studente sono gli eventi con i relativi riferimenti temporali nella colonna "Timestamp".

I dati, perciò, sono stati pre-processati su Python, mentre per la creazione dell'event log si è utilizzato un altro *tool* caratteristico del process mining, ossia Disco. Disco è uno strumento utilizzato durante le analisi di process mining per gestire event log, la loro conversione, la creazione di modelli ed il filtraggio.

Una volta ottenuto il file log in formato *xes*[15], sono stati estratti i processi relativamente ad ogni materia, sia per la totalità degli studenti che per ogni

[15] eXtensible Event Stream (XES)

sottocategoria, perciò per gli studenti laureati in corso, un anno fuori corso e oltre un anno fuori corso.

In base ai risultati ottenuti dalle statistiche precedentemente descritte, in questa tesi verranno approfondite le analisi dei processi di quei corsi che sono stati individuati come “critici”.

I processi sono stati estratti utilizzando l’algoritmo Infrequent Inductive Miner, tramite la libreria di Python dedicata alle tecniche di Process Mining, pm4py.

È stato scelto l’algoritmo Infrequent Inductive Miner perché permette una migliore gestione del rumore. Difatti, per l’individuazione del threshold k ottimale da inserire, si è creato un loop nel quale si esegue il modello con tutti i valori k, da 0 a 1, e per ognuno vengono calcolate le quattro metriche per la valutazione dei modelli. Per la scelta del threshold adeguato si è guardato specialmente ad un miglior compromesso tra la fitness e la precision.

In particolare, il codice utilizzato è il seguente:

```
result = pd.DataFrame(columns=['Threshold','Fitness','Precision', 'Generalization', 'Semplicity'])
for k in np.arange(0,1.1,0.1):
   mod = pm4py.discovery.discover_process_tree_inductive(log, k)
   net, initial_marking, final_marking = pm4py.convert_to_petri_net(mod)
   fitness = pm4py.fitness_token_based_replay(log,, net, initial_marking, final_marking)
```

```
prec = pm4py.precision_token_based_replay(log, net, initial_marking, final_marking)
gen = generalization_evaluator.apply(log, net, initial_marking, final_marking)
simp = simplicity_evaluator.apply(net)
row = pd.DataFrame({'Threshold':i,'Fitness':fitness['average_trace_fitness'], 'Precision':prec, 'Generalization':gen, 'Semplicity':simp},index=[0])
result = pd.concat([result, row])
```

Il risultato sarà una tabella nella quale ogni riga conterrà un threshold con le relative informazioni riguardo alla fitness, precision, generalizzazione e semplicità del modello allenato con il valore k considerato. Una volta ottenuto questo valore, il modello è stato estratto utilizzando la seguente funzione:

tree = pm4py.discovery.discover_process_tree_inductive(*event log*, t*hreshold*)

Per la visualizzazione dei processi, invece, sono state utilizzate le Reti di Petri, utilizzando anche le informazioni sulle frequenze. Di seguito il codice scritto.

```
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 = event log)
pt_visualizer.view(gviz)
```

3.4 PREDIZIONE

Dopo aver analizzato l'iter con il quale gli alunni studiano ogni esame, commentato adeguatamente i corsi considerati un blocco per gli studenti ed estratto le relative conclusioni, l'analisi è volta a rispondere alle seguenti domande: Quali sono i principali fattori che influenzano il rendimento scolastico degli studenti? È possibile prevedere il voto medio degli studenti alla fine del loro percorso di laurea? Per rispondere alle suddette domande sono state utilizzate tecniche di machine learning quali la Regressione Logistica e Support Vector Machine. Le analisi sono state svolte considerando sia i dati relativi solamente al primo semestre che relativi al primo anno. Si tengono in considerazione solamente i primi periodi perché l'obiettivo è comprendere quanto l'inizio della carriera dello studente incide sul loro percorso totale, ossia in che misura la progressione della loro carriera nel primo anno influisce sul rendimento alla laurea.

Il dataset utilizzato come input è stato creato nel seguente modo.

Partendo dal set di dati prima descritto, si sono andate ad estrarre le attività degli studenti effettuate rispettivamente nel primo semestre e nel primo anno, filtrando solo per gli esami passati.

Raggruppando i dati per ogni studente, si sono andati a contare il numero di esami svolti, nel primo semestre e nel primo anno, mentre la media dei voti è stata calcolata relativamente all'intera carriera dello studente.

La tabella, dunque, è stata costruita nel seguente modo: “STU_ID”, “Numero di esami”, “Voto medio”, “Tempo laurea”.

Perciò, prendendo questo dataset in input, è stato addestrato un modello di regressione logistica, che ha come predittori le variabili scalate riguardanti il numero di esami ed il tempo laurea, e come variabile dipendente y una variabile binaria che rappresenta se lo studente si è laureato con una media elevata o meno.

La variabile “High performance” è stata creata assegnando “1” quando uno studente si è laureato con una media superiore al 27, e assegnando “0” per coloro con una media inferiore.

L’analisi è volta a scoprire le possibili relazioni tra il numero di esami superati durante il primo anno, il tempo impiegato dagli studenti per terminare il percorso di studi ed il voto medio ottenuto.

Il voto è una variabile interessante da predire, un voto alto può significare che lo studente è bravo, ma può anche significare più tempo a disposizione per studiare al meglio. L’analisi è volta a scoprire, perciò, se il numero di esami ed il tempo di laurea incidono positivamente o negativamente sul voto.

Inoltre, sono stati addestrati altri due modelli, utilizzando le tecniche della regressione logistica e Support Vector Machine.

L’obiettivo è capire se si può predire il voto finale, perciò le performance degli studenti, partendo dalle informazioni degli esami dati al primo anno e dal tempo impiegato per studiarli.

Per la creazione del dataset da sottoporre a queste due tecniche, sono state eliminate le attività che erano state aggiunte in precedenza in virtù dell'estrazione dei processi, quali "End first semester", "End first year", "End third semester", "End second year" e "After second year", e creato un sottoinsieme considerando solo gli esami svolti nel primo anno. Si prendono in considerazione solo questi ultimi perché si vuole estrarre la relazione tra le performance degli studenti, e quindi il loro voto di laurea, e quanti e quali esami vengono svolti nel primo anno.

Si è pensato di inserire un'altra informazione rilevante, ossia il tempo impiegato da ciascuno studente per studiare ogni esame. Questo tempo è stato calcolato come la differenza tra la prima attività dello studente registrata relativa a quel determinato esame e l'ultima, ossia "Superato": perciò il tempo, calcolato in giorni, che uno studente impiega per passare ogni esame.

Il set di dati utilizzato come input agli algoritmi sopra citati è composto da 410 righe, perciò una riga per ogni studente, e per ogni esame il tempo impiegato: se lo studente non ha effettuato l'esame entro il tempo considerato è stato inserito il valore "1000", che indica perciò un numero di giorni superiore all'Anno Accademico.

L'ultimo step prima di applicare le tecniche di machine learning è quello di dividere il dataset, che andrà in input all'algoritmo in training set e test set, in modo tale da allenare il modello e testarlo su due differenti set di dati, per avere una migliore

consapevolezza dell'errore che il modello può commettere in altre applicazioni reali.

Il dataset è stato diviso in due sotto-sezioni, una per le variabili esplicative x e l'altra per la variabile da predire y. In seguito, tutte le variabili sono state scalate utilizzando la funzione fornita dalla libreria sklearn, MinMaxScaler(), la quale effettua la standardizzazione per ogni variabile sottraendo il valore minimo della stessa e poi dividendo per la differenza tra il valore massimo e minimo.

Infine, si è proceduto a creare il training e test set, rispettivamente contenenti l'80% dei dati ed il 20%, con la funzione train_test_split(x, y, test_size=0.2).

Per l'esecuzione del modello di Regressione logistica, si è utilizzata la funzione della libreria sklearn: linear_model._base.LinearRegression().fit(X_train, y_train).

Per quanto riguarda il modello di Support Vector Machine, invece, prima della sua esecuzione è stata necessaria la ricerca dei migliori parametri da utilizzare.

La ricerca è stata effettuata nel seguente modo: per ogni kernel, lineare, polinomiale e rbf[16], è stato creato un loop all'interno del quale venivano eseguiti tutti i possibili modelli cambiando i parametri, iterando per il parametro C da 1 a 1000, e per quanto riguarda il parametro gamma nel kernel rbf è stato iterato da 0 a 0.1, con uno step di 0.01, mentre per trovare il miglior grado del polinomio si è iterato per un valore da 1 a 5.

[16] Radial Basis Function (RBF)

Il codice utilizzato per l'iterazione dei modelli con diversi kernel e parametri è il seguente:

```
parametri = [
   {"kernel": ["rbf"], "gamma": np.arange(0,0.1,0.01), "C": np.arange(1,1000,100)},
   {"kernel": ["linear"], "C": np.arange(1,1000,100)},
   {"kernel": ["poly"], "degree":np.arange(1,5,1),"gamma": np.arange(0,0.1,0.01), "C": np.arange(1,1000,100)}]

svc = SVC(random_state=0)
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=10, random_state=0)
search = GridSearchCV(estimator=svc, param_grid=parametri, scoring="roc_auc", cv=cv)
search.fit(X_train, y_train)

best_params = search.best_params_
print("Migliori parametri:", best_params)
best_model = search.best_estimator_
y_pred = best_model.predict(X_test)
f1 = f1_score(y_test, y_pred, average='weighted')
print(f"F1-score del miglior modello: {f1}")
```

Il modello finale è stato scelto in base al modello che garantiva un F1 score maggiore.

La funzione utilizzata per l'esecuzione del modello scelto è integrata nella libreria sklearn, ed è la seguente: svc = SVC(*kernel*, parametri).

IV. STATISTICHE

In questo capitolo, di seguito saranno elencate diverse statistiche volte ad estrarre informazioni utili sia per gli studenti che per i professori ed infine per l'amministrazione dell'Ateneo, ossia si andrà ad identificare quale esame rappresenta un blocco per la carriera universitaria degli studenti.

4.1 **ANALISI DEI TASSI DI SUPERAMENTO DEGLI ESAMI**

Le tre tabelle seguenti rappresentano la percentuale di studenti assenti, bocciati e ritirati per ciascun esame sul totale degli studenti che hanno superato l'esame, rispettivamente per gli esami del primo anno, del secondo e, per ultimi, gli esami a scelta del secondo anno. La percentuale è stata calcolata dividendo il numero di assenti, bocciati e ritirati sul numero totale di persone che hanno superato l'esame, che per i corsi obbligatori sarà uguale a 410, ossia la totalità degli studenti presenti nel dataset. Si noti che i valori possono superare il 100% perché uno studente potrebbe essere assente, ritirato o bocciato all'esame anche più volte prima di superarlo.

Per quanto riguarda gli esami a scelta, sarà aggiunta una colonna che andrà ad indicare la percentuale di studenti che hanno scelto di svolgere quel determinato esame sul totale, per dare un'idea di quante persone effettivamente hanno dato quell'esame.

Esame	Assenti/Promossi	Bocciati/Promossi	Ritirati/Promossi
Fisica Generale 1	68.89%	166.17%	4.20%
Algebra Lineare e Geometria	41.38%	0%	12.81%
Analisi Matematica 1	75.99%	133.66%	11.39%
Fisica Generale 2	74.5%	80.75%	2.25%
Analisi Matematica 2	93.62%	21.57%	17.89%
Fondamenti Di Informatica	9.56%	0%	14.46%
Economia Dell'impresa	74.15%	0%	0%

Tabella IV.1- I ANNO: Percentuali Assenti, Bocciati, Promossi

Come già affermato in precedenza, il primo anno rappresenta il momento più critico per la carriera dello studente; perciò, osservare questi dati è utile per individuare preventivamente i possibili blocchi in modo tale da poter consigliare gli studenti all'inizio dei loro studi.

Si nota subito come la percentuale più alta di studenti bocciati rispetto ai promossi sia relativa all'esame "Fisica generale 1". Difatti, avere il 166.17% di bocciati rispetto ai promossi significa che, in media, una persona è stata bocciata almeno una volta e molti sono stati bocciati più di una volta.

Similmente, anche per l'esame "Analisi matematica 1" si nota una percentuale di bocciati superiore al 100%. Per questi due esami, inoltre notiamo anche un'alta percentuale di assenti: perciò ci sono molti studenti che cominciano a studiare il materiale per quell'esame, ma scelgono di non presentarsi il giorno dell'appello perché non si sentono pronti per superare la prova.

Da questa tabella, si evince che anche altri due corsi richiedono un'attenzione particolare, ossia "Fisica generale 2" e "Analisi matematica 2". Per quanto riguarda il primo, vi sono delle percentuali preoccupanti sia per quanto riguarda i bocciati che gli assenti. Il corso di Analisi Matematica 2, invece, risulta avere un basso tasso di studenti che non superano l'esame, ma in compenso un elevato numero di assenti.

Di seguito verranno analizzati gli esami del secondo anno.

Esame	Assenti/Promossi	Bocciati/Promossi	Ritirati/Promossi
Elementi di Elettronica	20.78%	34.96%%	9.05%
Fondamenti di Automatica	6.22%	0%	37.06%
Elettrotecnica	4.39%	231.22%	0.24%
Elettromagnetismo per la Trasmissione dell'Informazione	47.75%	12.5%	5.75%
Controlli Automatici	27.32%	147.07%	10.49%
Algoritmi e Strutture Dati	16.22%	13.24%	11.59%

Tabella IV.2– II ANNO: Percentuali Assenti, Bocciati, Promossi

Analizzando il tasso di superamento degli esami del secondo anno si osserva una situazione critica per l'esame di Elettrotecnica, con una percentuale di bocciati superiore al 200%, ossia in media uno studente è stato bocciato almeno due volte prima di superare l'esame.

Un altro esame che può essere ritenuto critico è Controlli Automatici, con un tasso di studenti bocciati del 147,07%. In entrambi i casi sottolineati, non si registrano particolari criticità per quanto riguarda gli studenti assenti e ritirati.

Di seguito saranno analizzati, invece, gli esami a scelta. Nel piano di studi sono previsti, infatti, dei crediti formativi che lo studente può colmare scegliendo tra

diversi esami. Durante il secondo anno, lo studente può scegliere di svolgere un esame tra: Calcolo delle Probabilità e Statistica Matematica, Meccanica Razionale, Algebra e Logica, Analisi Numerica.

Esame	Assenti/Promossi	Bocciati/Promossi	Ritirati/Promossi	Iscritti/Totale
Calcolo delle Probabilità e Statistica Matematica	14.20%	0%	48.15%	40%
Meccanica Razionale	7.70%	0%	11.54%	6%
Algebra e Logica	58.03%	0%	3.11%	47%
Analisi Numerica	29.77%	12.36%	6.74%	43%

Tabella IV.3– II ANNO Esami a scelta: Percentuali Assenti, Bocciati, Promossi

Per quanto riguarda gli insegnamenti a scelta non si notano particolari criticità. Algebra e logica, tuttavia, presenta un tasso di studenti assenti superiore rispetto agli altri, ma nel complesso non risulta un esame critico, anche perché ha un tasso di bocciati uguale a zero: potremmo affermare, perciò, che molti studenti preferiscono non presentarsi il giorno dell'appello per poi svolgere l'esame quando si sentono più preparati e sicuri di superarlo.

La percentuale di iscritti rispetto al totale degli studenti (410), è stata calcolata prendendo in considerazione il numero di studenti promossi, ossia la totalità di coloro i quali hanno sostenuto l'esame, e quindi di coloro che hanno scelto quel corso. Possiamo dedurre, perciò, che Algebra e Logica, nonostante sia un corso con una percentuale di assenti abbastanza elevata, è l'esame maggiormente scelto dagli studenti.

In sintesi, da questa prima analisi risulta che gli insegnamenti di Fisica Generale 1 e 2, Analisi Matematica 1 e 2, Elettrotecnica e Controlli Automatici sono ritenuti dei possibili colli di bottiglia nella carriera degli studenti.

Per i suddetti esami, perciò, si analizzano le differenze tra gli studenti laureati in corso e coloro i quali si sono laureati con più di un anno di ritardo.

Esame	Assenti/Promossi	Bocciati/Promossi	Ritirati/Promossi
Fisica Generale 1	48.13%	121.99%	0.83%
Fisica Generale 2	51.67%	82.08%	1.25%
Analisi Matematica 1	58.26%	99.59%	6.61%
Analisi Matematica 2	48.15%	15.64%	18.52%
Elettrotecnica	3.7%	218.11%	0.41%
Controlli Automatici	9.47%	118.93%	1.65%

Tabella IV.4 - Studenti in corso: esami critici

Esame	Assenti/Promossi	Bocciati/Promossi	Ritirati/Promossi
Fisica Generale 1	146.0%	286.0%	14.0%
Fisica Generale 2	145.83%	106.25%	4.17%
Analisi Matematica 1	142.0%	216.0%	30.0%
Analisi Matematica 2	202.0%	52.0%	20.0%
Elettrotecnica	7.69%	259.62%	0.0%
Controlli Automatici	67.31%	188.46%	21.15%

Tabella IV.5 - Studenti fuori corso: esami critici

In generale, si nota che le criticità sottolineate in precedenza vengono rispecchiate anche nei dati relativi agli studenti in corso, tranne per quanto riguarda l'esame Matematica 2 che non rappresenta una criticità per questi ultimi.

Notiamo, inoltre, che Elettrotecnica rimane il corso con il più alto tasso di bocciati, che anche per gli studenti laureati in corso è superiore al 200%.

Come differenze tra le due tipologie di studenti possiamo notare che, per l'esame Fisica Generale 1, la percentuale di bocciati è più che raddoppiata per gli studenti fuori corso, mentre per quanto riguarda gli assenti aumenta di quasi il 100%. Un comportamento simile si osserva per l'esame Analisi Matematica 1.

Inoltre, se l'esame Analisi Matematica 2 per gli studenti laureati in corso sembrava non essere un problema, per quanto riguarda gli studenti fuori corso il tasso di assenti è superiore al 200%: questo significa che, in media, uno studente si prenota agli appelli per almeno due volte prima di presentarsi all'esame effettivamente.

Per quanto riguarda l'esame di Fisica Generale 2, il tasso di non superamento dell'esame è leggermente aumentato, ma ciò che differisce di molto rispetto alle statistiche riguardanti gli studenti in corso è la percentuale degli assenti.

Possiamo affermare, infatti, che è presente la tendenza negli studenti fuori corso a prenotarsi agli appelli ma non andarci effettivamente, magari perché non si sentono abbastanza pronti per sostenere l'esame o perché ne stanno preparando anche altri nel mentre.

4.2 **ANALISI TEMPORALE**

Dopo aver analizzato i tassi di superamento di ogni esame, è interessante inserire nell'analisi anche dei riferimenti temporali.

La tabella seguente mostra il numero di persone che hanno superato l'esame in ogni periodo considerato. L'informazione più interessante da estrapolare da quest'analisi è quella inserita nella colonna "*On time*", ossia la percentuale di studenti che superano l'esame considerato entro il periodo prestabilito dal piano di studi.

L'obiettivo di quest'analisi è osservare quanti studenti effettivamente riescono a superare gli esami entro il periodo stabilito nel piano di studi, e soprattutto notare le differenze tra gli studenti laureati in corso e fuori corso, per verificare se i primi seguono maggiormente il manifesto degli studi o meno.

Quest'analisi è il punto di partenza per individuare gli esami che costituiscono un blocco per gli studenti. Infatti, possiamo dedurre che se un corso ha un basso tasso di superamento entro il periodo definito significa che gli studenti non riescono a superare l'esame la prima volta che lo provano, oppure decidono di rimandare la sua preparazione in un semestre successivo, perché il corso è considerato difficile da sostenere negli studi o perché è considerato tale in base a voci diffuse negli ambienti studenteschi, inducendo comunque lo studente a rimandare l'esame in altri momenti.

Le statistiche seguenti verranno calcolate per gli studenti iscritti negli anni accademici dal 2015 al 2016, e dal 2017 al 2019, perché il piano di studi cambia leggermente per quanto riguarda le materie del secondo anno.

Esame	1° Semestre	1° Anno	3° Semestre	2° Anno	Oltre il 2° Anno	Periodo	On time
Fisica Generale 1	41	34	31	4	76	1° Semestre	22.04%
Algebra Lineare e Geometria	96	76	4	4	8	1° Semestre	51.06%
Analisi Matematica 1	78	50	15	10	32	1° Semestre	42.16%
Fisica Generale 2	0	21	24	17	121	1° Anno	11.48%
Analisi Matematica 2	0	60	51	29	49	1° Anno	31.75%
Fondamenti Di Informatica	0	151	25	9	6	1° Anno	79.06%
Economia Dell'impresa	0	146	36	1	8	1° Anno	76.44%

Tabella IV.6 – AA 2015/2016, 2016/2017 I ANNO: Percentuale di superamento entro il periodo stabilito

Da quest'analisi condotta sulle performance degli studenti iscritti negli anni accademici 2015/2016 e 2016/2017, si nota che l'esame di Fisica Generale 2 ha la percentuale di superamento entro il periodo prestabilito più bassa; infatti, la maggior parte degli studenti supera l'esame dopo il secondo anno, mentre il corso è previsto per il secondo semestre del primo anno. È importante, però, fare una considerazione: nel corso di studi non esistono propedeuticità, ossia non è obbligatorio sostenere prima un esame per farne un altro, quindi, ad esempio, si potrebbe sostenere Analisi Matematica 2 senza aver superato Analisi Matematica 1, e similmente per Fisica Generale. Tuttavia, seppur non sono propedeutici esiste comunque una relazione tra i due esami, ed è lecito sostenere che è meglio conoscere gli argomenti di Fisica Generale 1 prima di iniziare a studiare Fisica Generale 2, così come per Analisi Matematica.

Perciò il basso tasso di superamento entro la fine del primo anno di Fisica Generale 2 potrebbe essere influenzato dall'aver fatto o meno Fisica Generale 1: anche per quest'ultimo, infatti, si nota una percentuale critica, ossia del 22%, e ci sono molti più studenti che danno l'esame dopo la fine del secondo anno che entro il primo semestre del primo anno, quando il corso è previsto.

Inoltre, meritano un'attenzione particolare gli esami Analisi Matematica 1 e 2, il cui tasso di superamento entro il periodo stabilito è inferiore al 50%, tuttavia non sembra essere una situazione critica perché la maggior parte degli studenti, comunque, entro il semestre successivo supera l'esame.

Infine, si notano degli elevati tassi di successo per gli esami Algebra lineare e Geometria, Fondamenti di Informatica ed Economia dell'Impresa, rispettivamente del 51.06%, 79.06% e 76.44%. Solo pochi studenti, infatti, decidono di rimandare questi esami nei periodi successivi.

Esame	1° Semestre	1° Anno	3° Semestre	2° Anno	Oltre il 2° Anno	Periodo	On time
Elementi di Elettronica	0	0	107	47	37	3° Semestre	56.02%
Fondamenti di Automatica	0	0	65	31	91	3° Semestre	34.76%
Elettrotecnica	0	0	42	121	28	3° Semestre	21.99%
Elettromagnetismo per la Trasmissione dell'Informazione	0	0	0	47	141	2° Anno	25.0%
Controlli Automatici	0	0	0	109	82	2° Anno	57.07%
Algoritmi e Strutture Dati	0	0	0	91	100	2° Anno	47.64%

Tabella IV.7 – AA 2015/2016, 2016/2017 II ANNO: Percentuale di superamento entro il periodo stabilito

Dall’analisi temporale effettuata considerando le materie del secondo anno, la percentuale minore rispetto al superamento “in tempo” dell’esame si può notare relativamente ad Elettrotecnica: infatti, solo 42 studenti su 191 iscritti agli anni accademici 2015/2016 e 2016/2017 superano l’esame entro il primo semestre del secondo anno, mentre 121 lo rimandano al semestre successivo.

Similmente per i corsi di Fondamenti di Automatica, Elettromagnetismo per la Trasmissione dell'Informazione, Algoritmi e Strutture Dati, si può notare che la maggioranza degli studenti rimanda questi esami dopo la fine del secondo anno, ed infatti hanno una percentuale di superamento entro le giuste tempistiche inferiore al 50%.

Per quanto riguarda Controlli Automatici, invece, seppur dall’analisi precedente risultasse essere un esame problematico, perché il tasso di bocciatura era elevato, gli studenti comunque iniziano a prepararlo in tempo e preferiscono non rimandare questo esame: ha infatti una percentuale del 57.07%.

In definitiva, relativamente agli studenti iscritti agli anni accademici 2015/2016 e 2016/2017, gli esami che risultano essere un blocco per gli studenti sono Fisica 1, Fisica 2 ed Elettrotecnica.

Di seguito le analisi riguardanti gli studenti iscritti agli anni accademici 2017/2018, 2018/2019 e 2019/2020.

Esame	1° Semestre	1° Anno	3° Semestre	2° Anno	Oltre il 2° Anno	Periodo	On time
Fisica Generale 1	63	76	22	5	53	1° Semestre	28.77%
Algebra Lineare e Geometria	122	77	14	3	2	1° Semestre	55.96%
Analisi Matematica 1	103	84	9	10	13	1° Semestre	47.03%
Fisica Generale 2	0	61	53	28	75	1° Anno	28.11%
Analisi Matematica 2	0	138	52	9	20	1° Anno	63.01%
Fondamenti Di Informatica	0	174	24	14	5	1° Anno	80.18%
Economia Dell'impresa	0	181	31	1	6	1° Anno	82.65%

Tabella IV.8 – AA 2017/2018, 2018/2019, 2019/2020 I ANNO: Percentuale di superamento entro il periodo stabilito

Gli studenti iscritti agli anni accademici 2017/2018, 2018/2019 e 2019/2020 sono in totale 219.

Si può immediatamente notare un miglioramento di tutte le percentuali: ciò significa che, rispetto agli anni precedenti, gli studenti cercano di rimandare il meno possibile gli esami e tentano di superarli entro il periodo stabilito dal piano di studi.

Tuttavia, gli esami che risultavano critici negli anni precedenti, ossia Fisica Generale 1 e 2, sono caratterizzati comunque da una percentuale più bassa di superamento entro il periodo stabilito.

Esame	1° Semestre	1° Anno	3° Semestre	2° Anno	Oltre il 2° Anno	Periodo	On time
Elementi di Elettronica	0	0	111	65	42	3° Semestre	50.92%
Fondamenti di Automatica	0	0	83	75	57	3° Semestre	38.6%
Elettrotecnica	0	0	96	104	19	3° Semestre	43.84%
Elettromagnetismo per la Trasmissione dell'Informazione	0	0	0	117	**95**	2° Anno	55.19%
Controlli Automatici	0	0	0	145	74	2° Anno	66.21%
Programmazione ad Oggetti	0	0	0	46	12	2° Anno	79.31%

Tabella IV.9 – AA 2017/2018, 2018/2019, 2019/2020 II ANNO: Percentuale di superamento entro il periodo stabilito

Lo stesso comportamento si può notare per gli esami del secondo anno. Inoltre, l'esame di Elettrotecnica risultava critico per gli studenti iscritti agli anni accademici 2015/2016 e 2016/2017, negli anni successivi si nota un netto

miglioramento, anche se comunque la maggior parte degli studenti rimanda l'esame al semestre successivo.

Andremo ora ad analizzare le differenze tra gli studenti che si laureano in corso e coloro che si laureano fuori corso, tenendo in considerazione solo gli esami ritenuti critici, analizzando i dati relativi alla totalità degli iscritti. A causa delle difficoltà incontrate nello studio e nel superamento di questi esami, gli studenti potrebbero non riuscire a seguire il manifesto e di conseguenza non riuscire a laurearsi in tempo: questi esami, perciò, potrebbero rappresentare dei colli di bottiglia.

Esame	On time In corso	On time Fuori corso
Fisica Generale 1	37.76%	4.0%
Fisica Generale 2	33.75%	0.0%
Analisi Matematica 1	59.09%	6.0%
Analisi Matematica 2	66.67%	6.0%
Elettrotecnica	40.33%	17.31%

Tabella IV.10- Differenze di comportamento tra i laureati in corso e fuori corso per gli esami critici

La tabella mostra che, per alcuni esami, la situazione è critica anche per gli studenti laureati in corso, poiché meno del 50% di loro riesce a sostenere questi esami in tempo. Tuttavia, le statistiche più preoccupanti le osserviamo per gli studenti fuori corso, dove nemmeno il 10% di loro riesce a sostenere i suddetti esami nei tempi previsti.

In particolare, nessuno degli studenti laureati fuori corso sostiene Fisica Generale 2 entro il primo anno, mentre meno del 5% sostiene Fisica Generale 1, e meno del 10% matematica Generale 1 e 2. Elettrotecnica, invece, è la materia che ha la percentuale più alta anche negli studenti fuori corso, ma presenta comunque una criticità poiché meno del 20% di loro riesce a sostenere l'esame in tempo.

Infine, possiamo affermare che esiste una notevole differenza tra gli studenti laureati in corso e fuori corso per quanto riguarda le tempistiche in cui danno gli esami, questo a sostegno dell'ipotesi che seguendo il piano di studi più fedelmente è possibile migliorare le proprie performance accademiche.

Di seguito un *boxplot* per l'analisi della distribuzione del tempo di superamento di ogni esame per gli studenti in corso e fuori corso, per sottolinearne le differenze. L'analisi dei boxplot verrà fatta sia per quanto riguarda gli studenti iscritti negli anni accademici 2015/2016-2016/2017 e per gli anni accademici dal 2017 in poi.

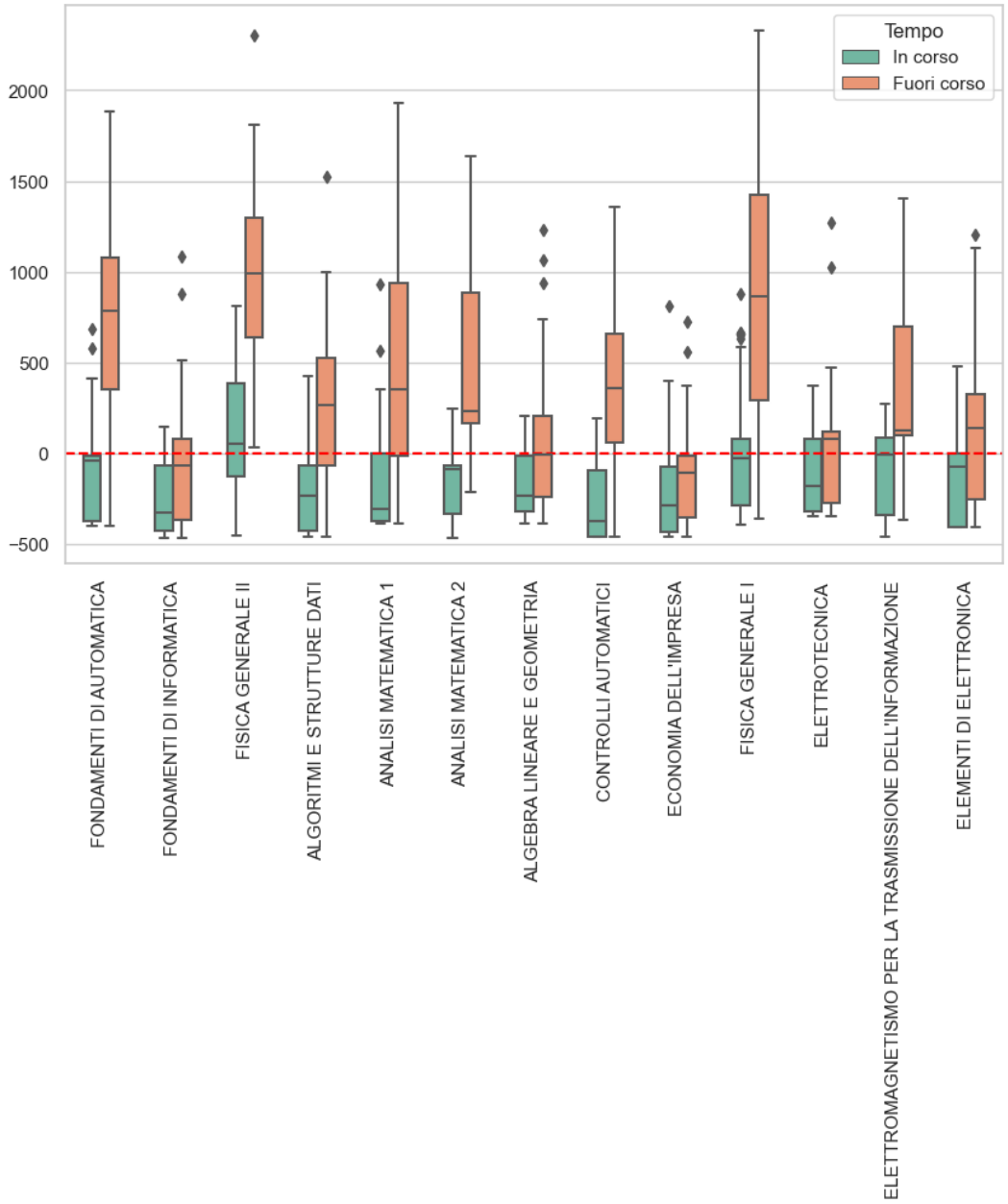

Figura IV.1 - AA 2015/2016-2016/2017: Boxplot distribuzione del tempo di preparazione di studenti in corso e fuori corso

Il boxplot è un grafico utile per sintetizzare al meglio le proprietà di una distribuzione: vengono rappresentati, infatti, il valore minimo della distribuzione (Q_0), il primo quartile (Q_1), la mediana (Q_2), il terzo quartile (Q_3) ed il valore massimo (Q_4) di una variabile.

Il rettangolo contiene il 50% centrale delle osservazioni, comprese tra il primo ed il terzo quartile, mentre la linea centrale rappresenta la mediana e i due segmenti verticali indicano la dispersione dei valori inferiori al primo quartile e superiori al terzo quartile non classificati come outliers. Questi ultimi, invece, sono rappresentati come dei punti al di sopra o al di sotto della distribuzione.

Osservando i boxplot, perciò, si riesce a comprendere il comportamento del 50% dei valori osservati, quanto sono dispersi i dati e se ci sono molti outliers.

In questo caso, il grafico rappresenta la differenza, calcolata in giorni, tra la fine del semestre entro il quale l'esame doveva essere dato ed il giorno in cui effettivamente lo studente lo ha superato. È presente, inoltre, una linea tratteggiata rossa orizzontale in corrispondenza dello zero, il quale valore indica che la differenza tra le due date è zero; perciò, significa che lo studente ha superato l'esame il giorno della fine del semestre indicato dal piano di studi relativo a quell'esame.

Se il boxplot si trova al di sotto della linea tratteggiata, perciò, significa che il core della distribuzione, ossia il 50% centrale, sono studenti che hanno superato l'esame prima della fine del semestre per il quale era previsto. Se il boxplot si

trova al di sopra della linea orizzontale fissata sullo zero, invece, gli studenti superano l'esame in una data successiva a quella della fine del semestre per il quale era previsto l'esame.

Si nota che per alcuni esami le differenze tra gli studenti in corso e fuori corso non sono particolarmente accentuate, come Fondamenti di Automatica, Algebra Lineare e Geometria, Economia dell'impresa, Elettrotecnica ed Elementi di elettronica. Altri insegnamenti, invece, denotano una notevole differenza tra le due categorie di studenti citate, ad esempio Fisica Generale 1 e 2, le quali distribuzioni degli studenti fuori corso si trovano nettamente al di sopra della linea tratteggiata. In particolare, per quanto riguarda Fisica Generale 2, non vi è nessun valore che si trova al di sotto o in corrispondenza dello zero; perciò, significa che nessuno studente supera l'esame entro il semestre stabilito.

Si può notare lo stesso comportamento nella distribuzione dei dati relativi agli studenti iscritti agli anni accademici 2017/2018, 2018/2019 e 2019/2020. Tuttavia, vediamo una maggiore sparsità della distribuzione relativa all'esame Fisica generale 1.

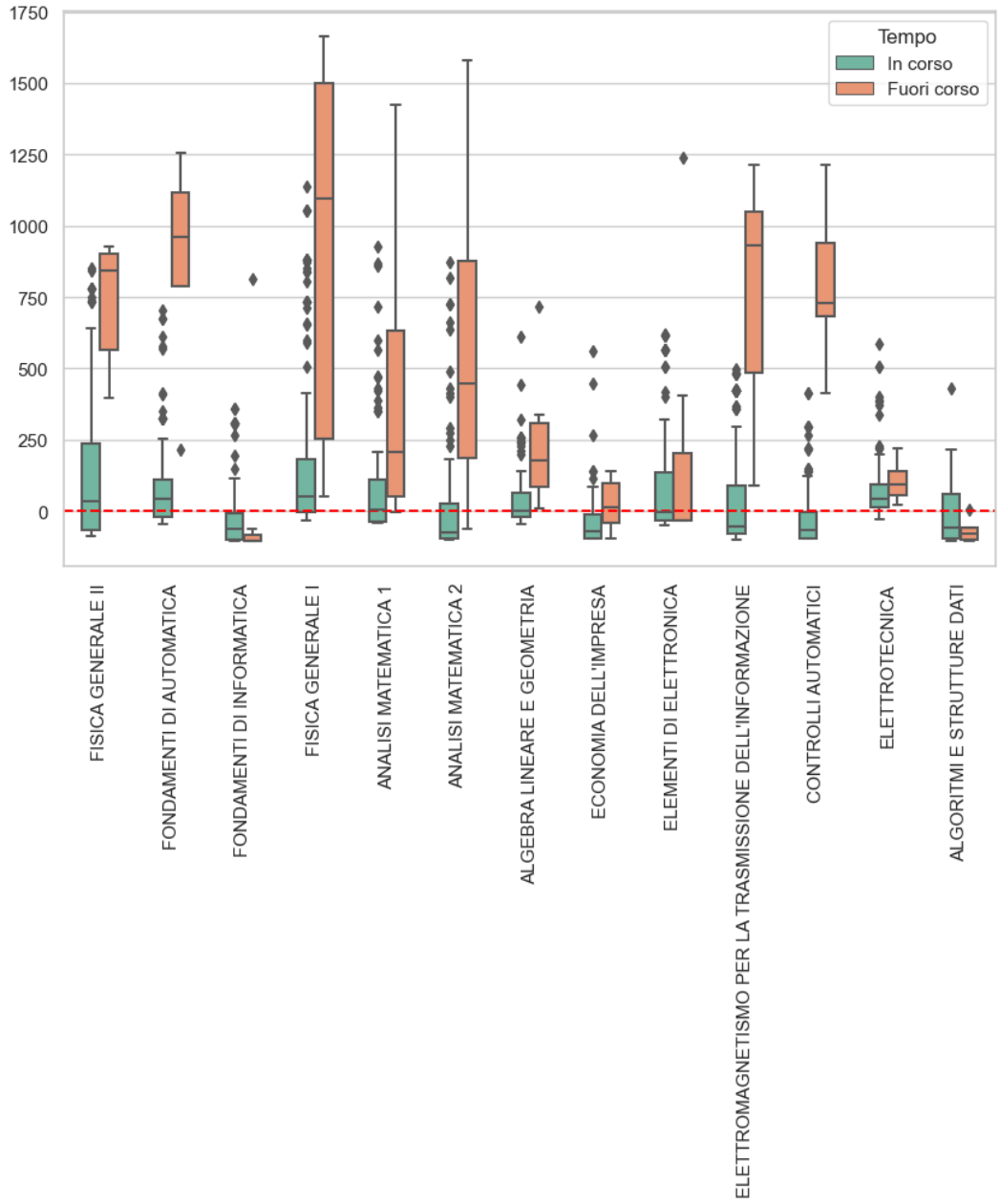

Figura IV.2 - AA 2017/2018-2018/2019-2019/2020: Boxplot distribuzione del tempo di preparazione di studenti in corso e fuori corso

4.2.1 Analisi del tempo di preparazione degli esami

Per approfondire l'analisi temporale, si va ora spostare il focus sul tempo di preparazione degli studenti per ciascun esame. Il tempo di preparazione degli esami è stato calcolato dalla differenza in giorni dalla data dell'ultima attività registrata per ciascuno studente riferito ad un esame, ossia "Superato", alla prima.

Perciò, attraverso un grafico a barre si andrà a visualizzare i corsi per i quali gli studenti impiegano più tempo per superarli con successo.

Si nota, quindi, che l'esame che richiede più tempo è Fisica Generale 1, e a seguire Fisica Generale 2, Analisi Matematica 1, Elettrotecnica e Analisi Matematica 2.

Gli ultimi risultati, infatti, confermano quanto detto in precedenza, ossia che i suddetti esami rappresentano dei colli di bottiglia per la carriera accademica degli studenti. Di seguito, perciò, verranno estratti i processi relativi ai suddetti esami per esaminarli nel dettaglio.

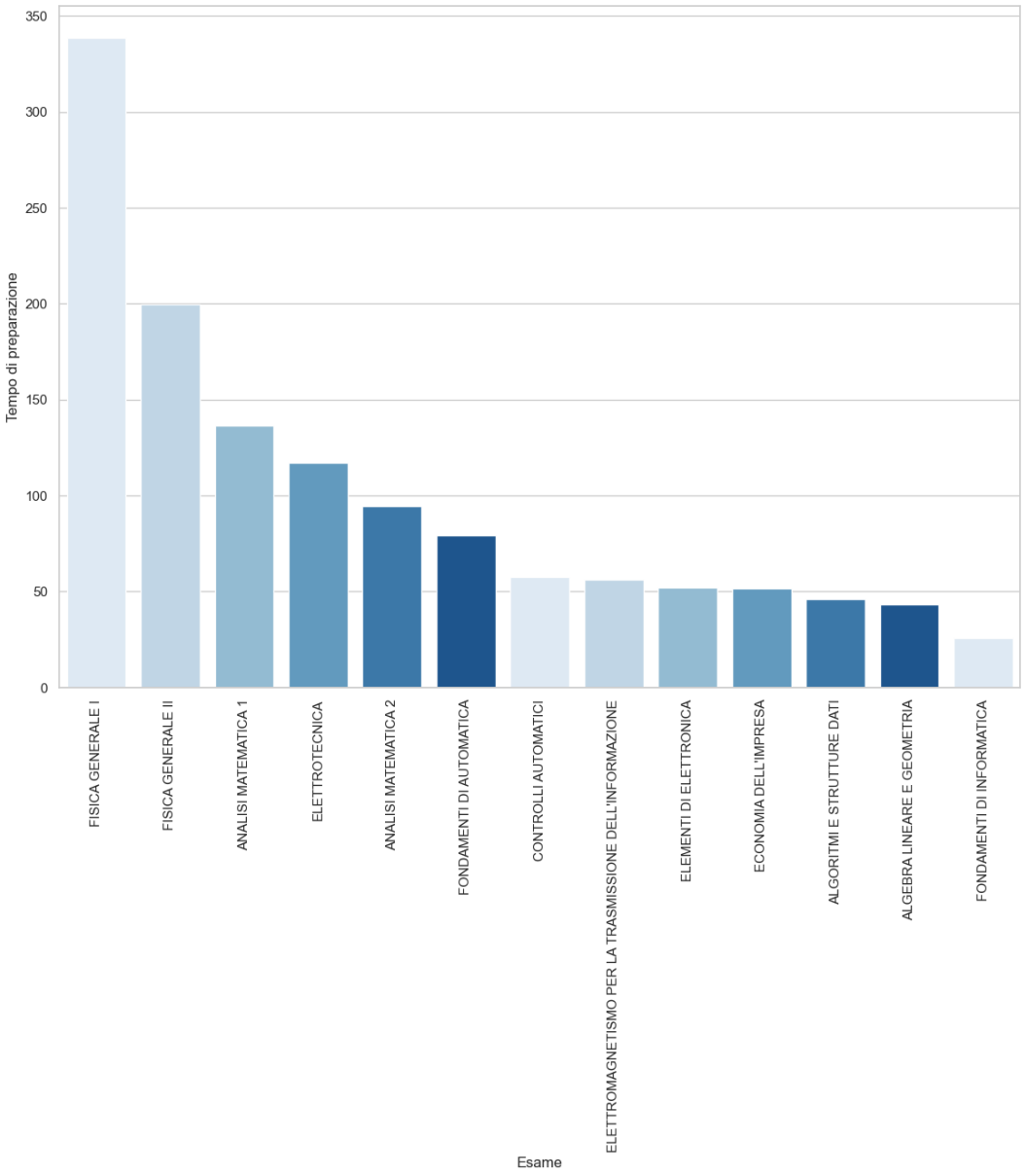

Figura IV.3 - Tempo di preparazione per ciascun esame

V. PROCESSI

Per approfondire l'analisi, si è proseguito estraendo i modelli di processo corrispondenti ai comportamenti degli studenti per la preparazione degli esami critici rilevati nel capitolo precedente, per poi sottolineare le differenze tra i modelli ottenuti per gli studenti in corso, un anno fuori corso e fuori corso di oltre un anno. Lo scopo, perciò, è quello di estrarre informazioni riguardo il momento in cui lo studente inizia e termina il processo di studio per un esame, identificando dove si trovano i blocchi.

Di seguito verranno analizzati i processi in ordine di tempo impiegato dagli studenti per la preparazione dell'esame, ossia: Fisica Generale 1, Fisica Generale 2, Analisi Matematica 1, Elettrotecnica e Analisi Matematica 2.

I seguenti processi sono stati estratti utilizzando l'algoritmo Infrequent Inductive Miner.

5.1 FISICA GENERALE 1

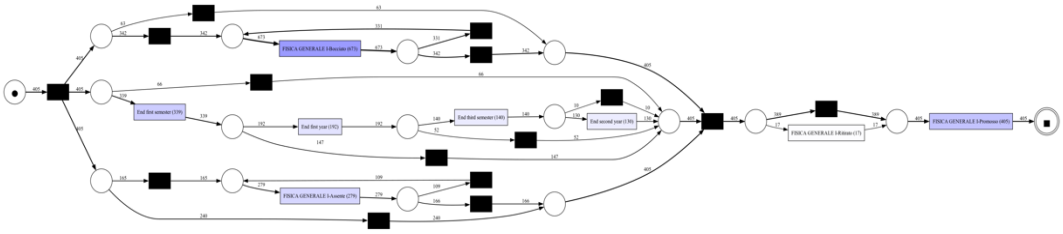

Figura V.1 - Processo Fisica Generale 1

Il processo relativo a tutti gli studenti che effettuano l'esame di Fisica Generale 1 è stato estratto inserendo come valore ottimale per la gestione del rumore 0.1, perciò i comportamenti non generali sono stati filtrati in minima parte.

Il processo ottenuto riproduce il log molto fedelmente, infatti ha una fitness del 0.975 e una precisione del 0.876. Inserendo un filtro per la gestione del rumore, anche la generalizzazione è abbastanza elevata, ossia del 0.849, mentre la semplicità del modello è pari a 0.667.

Inizialmente si possono notare tre gruppi di attività in parallelo.

Nel primo gruppo è presente un loop di studenti che vengono bocciati all'esame; perciò, si può notare immediatamente un blocco. Tuttavia, questo loop può essere saltato per arrivare direttamente alla fine del processo, ossia al superamento dell'esame.

Il secondo gruppo di attività aggiunge una prospettiva temporale all'analisi: gli studenti che saltano questa attività superano l'esame prima della fine del primo

semestre, entro quando dovrebbe essere sostenuto. Tuttavia, così come è anche emerso dalle analisi precedenti, se si osservano le frequenze si può notare che il numero di studenti che riesce a superare l'esame durante il primo semestre è poco cospicuo rispetto a coloro i quali rimandano l'esame al semestre successivo, o al secondo anno.

Il terzo gruppo di attività mostra un loop di studenti assenti, ossia che hanno studiato per l'esame ma non pensavano di essere sufficientemente preparati per superarlo e quindi non si sono presentati. Questo loop però può anche essere saltato: ci sono infatti studenti che riescono a superare l'esame la prima volta che lo sostengono.

Prima della fine del processo, con l'ultima attività "Promosso", vi è anche l'attività "Ritirato", ma con una frequenza molto bassa: difatti, sono pochi gli studenti che decidono di ritirarsi durante lo svolgimento dell'esame.

Di seguito si procede con l'analisi dei processi per ogni categoria di studenti considerati fino ad ora: studenti in corso, un anno fuori corso e fuori corso di oltre un anno.

5.1.1 Fisica Generale 1 – studenti in corso

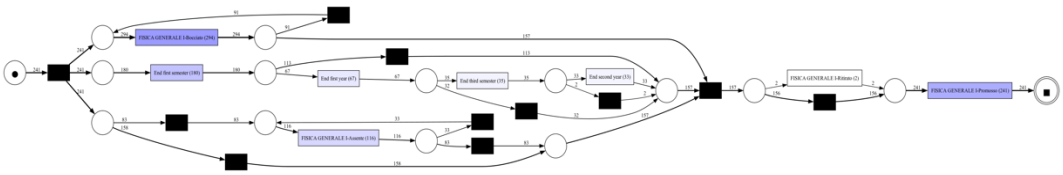

Figura V.2 – Studenti in corso: Processo Fisica Generale 1

Il threshold utilizzato per la gestione dei comportamenti infrequenti è di 0.3, mentre la fitness del processo è pari a 0.901, la precisione è 0.893, la generalizzazione 0.829, e la semplicità 0.680.

Il processo relativo agli studenti in corso è molto simile al processo generale; infatti, anche in questo caso possiamo notare gli stessi tre gruppi di attività parallele. La differenza con il processo descritto precedentemente descritto sta nelle frequenze: infatti, come previsto, vi sono un maggior numero di studenti che saltano le attività "End first year", "End third semester" e "End second year": ciò sta a significare che molti studenti che riescono a laurearsi in corso superano l'esame di Fisica Generale 1 entro la fine del primo anno.

Il risultato ottenuto è conforme alle considerazioni effettuate sulle statistiche precedenti: infatti le percentuali di studenti che riuscivano a superare l'esame entro il semestre previsto, riguardanti fisica generale 1, sono molto più alte per gli studenti che si laureano in corso, anche se comunque molti di questi studenti rimandano l'esame al semestre successivo.

Tuttavia, anche questi studenti affrontano dei problemi nella preparazione di Fisica Generale 1, come si può notare dal loop dell'attività "Bocciato" con una frequenza molto alta, e anche dagli studenti assenti.

5.1.2 Fisica Generale 1 – studenti un anno fuori corso

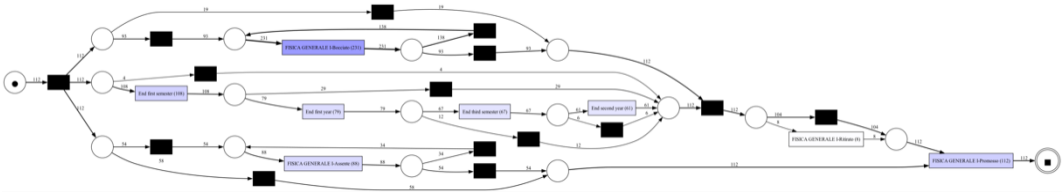

Figura V.3 – Studenti un anno fuori corso: Processo Fisica Generale 1

Il processo è stato estratto utilizzando un valore per la gestione del rumore di 0.1, ottenendo perciò la massima adesione al log, con una fitness di 1, ma una precisione di 0.637, una generalizzazione di 0.833, e semplicità del modello 0.667.

Anche nel caso del processo degli studenti laureati un anno fuori corso è presente lo stesso parallelismo con tre gruppi di attività, ed è ancora visibile una maggiore frequenza dell'attività "End first semester" rispetto alle altre attività temporali: quindi significa che anche se la maggior parte degli studenti che si laureano con solo un anno di ritardo non riesce a superare l'esame di Fisica 1 entro il primo semestre, supera l'esame comunque entro la fine del primo anno; mentre pochissimi studenti vengono promossi nel primo semestre.

Il processo dei laureati in corso e dei laureati con un anno di ritardo, per quanto riguarda la preparazione dell'esame di Fisica Generale 1, è perciò molto simile.

5.1.3 Fisica Generale 1 – studenti fuori corso di oltre un anno

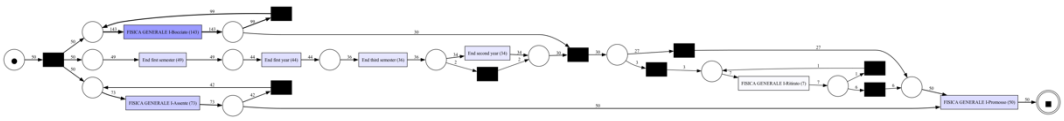

Figura V.4 – Studenti oltre un anno fuori corso: Processo Fisica Generale 1

Il threshold per la gestione del rumore scelto dopo una valutazione del miglior compromesso tra fitness e precisione è di 0.3. Il processo ottenuto ha una fitness di 0.93, precisione 0.893, una generalizzazione di 0.829 e semplicità del modello di 0.68.

Nel gruppo degli studenti laureati fuori corso, si può notare che vengono effettuate le attività "End first semester" e "End first year", perciò gli studenti superano l'esame dopo il primo anno, e tuttavia la maggior parte di loro viene promosso dopo la fine del secondo anno.

Inoltre, nonostante vi siano meno studenti in questo gruppo, la frequenza delle bocciature è comunque molto elevata.

5.2 FISICA GENERALE 2

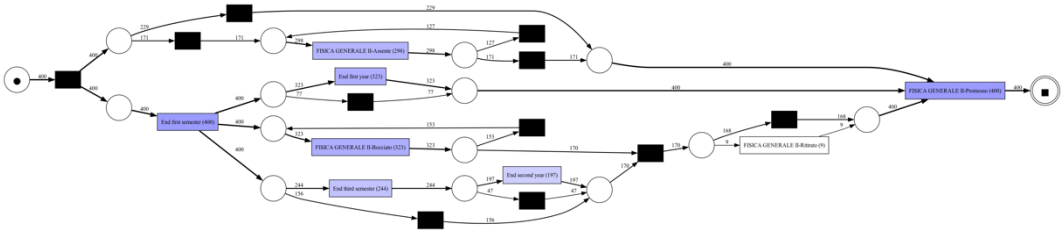

Figura V.5 - Processo Fisica Generale 2

Il corso di Fisica Generale 1 non è propedeutico per lo svolgimento di Fisica Generale 2, tuttavia, è ragionevole pensare che il superamento di quest'ultimo possa essere influenzato dal primo.

Per l'estrazione di questo processo è stato utilizzato un threshold di 0.1, ottenendo una fitness di 0.925, precisione di 0.85, una capacità di generalizzazione dei comportamenti pari a 0.912 e semplicità del modello di 0.66.

Il processo relativo al comportamento degli studenti che preparano l'esame Fisica Generale 2 è composto da due gruppi di attività eseguite in parallelo.

Il primo ramo contiene un loop dell'attività "Assente", infatti questo comportamento era già emerso nelle statistiche precedenti, con una percentuale di assenti del 74.5%. Tuttavia, il suddetto loop può anche essere saltato, per arrivare direttamente alla fine del processo superando l'esame.

Il secondo gruppo di attività comincia con “End first semester” dato che Fisica Generale 2 è un corso che si tiene al secondo semestre del primo anno; perciò, nessuna attività riguardo quest’esame è svolta nel primo semestre.

Dopo la fine del primo semestre si diramano altre tre attività in parallelo.

Alcuni studenti superano l’esame senza ostacoli, mentre la maggior parte di loro lo supera dopo la fine del primo anno, perciò effettua l’attività “End first year”.

Dopo il primo semestre vi è un loop relativo all’attività “Bocciato: si nota che quest’esame rappresenta un blocco per gli studenti per l’alta frequenza delle bocciature.

L’ultimo gruppo di attività rappresenta la sfera temporale, il quale fa emergere che molti studenti effettuano l’esame dopo il terzo semestre o addirittura dopo la fine del secondo anno.

Infine, prima di terminare il processo con la promozione all’esame, vi è un piccolo gruppo di studenti che si ritirano durante il corso dell’appello.

5.2.1 Fisica Generale 2 – studenti in corso

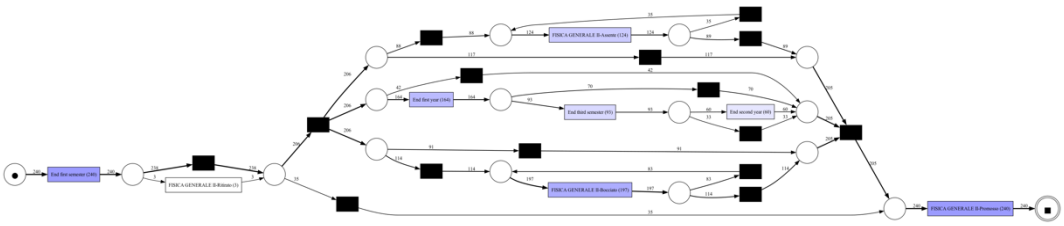

Figura V.6 – Studenti in corso: Processo Fisica Generale 2

Il valore scelto come threshold per non considerare anche i comportamenti infrequenti è di 0.1. Il processo così estratto ha un'elevata aderenza al log, con una fitness di 0.998 e una precisione di 0.904, ottenendo un modello con una generalizzazione di 0.875 e semplicità di 0.667.

Per quanto riguarda gli studenti laureati in corso, il processo inizia con la fine del primo semestre, essendo, come già detto, Fisica Generale 2 un corso situato al secondo semestre nel piano di studi.

Come prima attività nel processo vi è "Ritirato", ma come si può facilmente notare osservando le frequenze, viene saltata da quasi tutti gli studenti; infatti, è molto inusuale che uno studente si ritiri durante l'esame.

Una parte di studenti viene promossa immediatamente e non incontra, perciò, nessun ostacolo, mentre alcuni effettuano altre attività prima di superare l'esame.

Si nota, perciò, un parallelismo simile a quello analizzato in precedenza per il processo relativo a tutti gli studenti. Tuttavia, a differenza del processo generale, seppur poche persone riescono a passare l'esame entro il periodo previsto, ossia il

secondo semestre del primo anno, pochi studenti rimandano l'esame dopo il secondo anno. Infine, anche in questo processo, si nota un'alta frequenza dell'attività "Bocciato".

5.2.2 Fisica Generale 2 – studenti un anno fuori corso

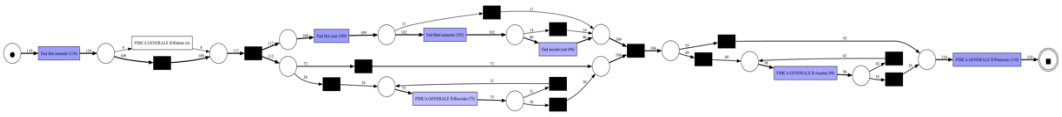

Figura V.7 – Studenti un anno fuori corso: Processo Fisica Generale 2

Il processo relativo agli studenti laureati un anno fuori corso è stato estratto scegliendo un threshold per la gestione del rumore di 0.2, ottenendo una fitness tra il modello ed il log di 0.982, una precisione pari a 0.939, la generalizzazione del modello ottenuto invece è 0.845 e semplicità uguale a 0.725.

Anche in questo caso il processo inizia con l'attività "End first semester" ed il loop, seppur poco popolato, dell'attività "Ritirato". Successivamente si notano due gruppi di attività in parallelo.

Il primo gruppo relativo alle attività temporali, dalle quali emerge una tendenza degli studenti a superare l'esame tra la fine del secondo anno e i periodi successivi.

Dal secondo gruppo di attività si diramano due scelte mutuamente esclusive: il superamento dell'esame dopo la fine del primo semestre, oppure superarlo dopo essere stati bocciati almeno una volta.

5.2.3 Fisica Generale 2 – studenti fuori corso

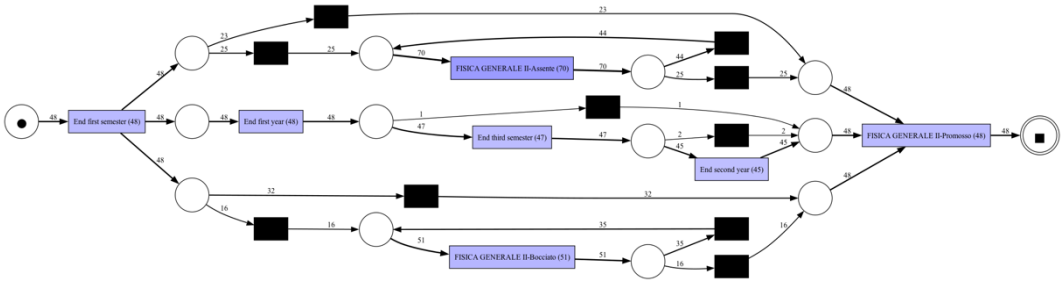

Figura V.8 - Studenti un anno fuori corso: Processo Fisica Generale 2

Il modello di comportamento degli studenti fuori corso nel preparare l'esame di Fisica Generale 2 è stato estratto utilizzando un valore del threshold di 0.1, ottenendo così un processo con un'elevata fitness, 0.999, precisione di 0.717, e generalizzazione della realtà rappresentata dal log di 0.747, con una risultante semplicità del modello pari a 0.689.

Nuovamente, dopo la fine del primo semestre, si hanno tre gruppi di attività in parallelo.

Dal primo gruppo si dipartono due attività esclusive: il primo arco porta alla fine del processo, e perciò al superamento dell'esame, e l'altro porta ad entrare nel ciclo dell'attività "Assente". Dal secondo gruppo di attività emerge ciò che distingue il processo di questo gruppo di studenti da quelli in corso, ossia che la maggior parte dei primi supera quest'esame dopo la fine del secondo anno. Il terzo gruppo è caratterizzato dal loop dell'attività "Bocciato", il quale, tuttavia, può essere saltato. Inoltre, si nota un'elevata frequenza dell'attività "Assente".

5.3 ANALISI MATEMATICA 1

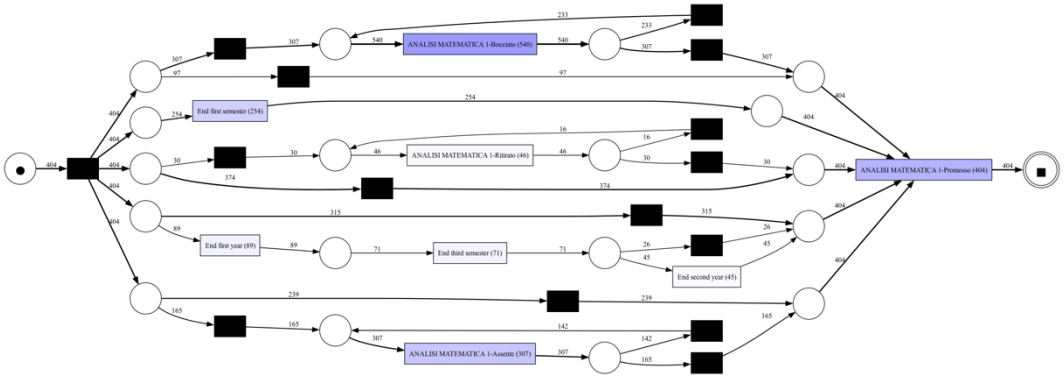

Figura V.9 - Processo Analisi Matematica 1

Dall'analisi delle statistiche presentate nel capitolo precedente, è emerso che tra gli esami che rappresentano un blocco per gli studenti vi sono Analisi Matematica 1 ed Analisi Matematica 2. In particolare, Analisi Matematica 1 è il terzo corso per il quale gli studenti impiegano maggior tempo a superarlo.

Il processo sopra presentato rappresenta il percorso che gli studenti seguono prima di essere promossi all'esame. È stato estratto utilizzando un threshold di 0.1 per non includere nel modello anche alcuni comportamenti a bassissima frequenza, ed il risultato ha una fitness del 0.97, una precisione di 0.711, un'elevata generalizzazione di 0.9 e semplicità 0.661.

Il processo generale è caratterizzato da un elevato parallelismo, si notano infatti cinque gruppi di attività parallele.

Il primo gruppo si dirama in due percorsi, il primo è rappresentato da un loop attorno all'attività "Bocciato", dove si nota una frequenza molto alta, mentre il

secondo percorso permette di saltare il loop ed arrivare alla fine del processo e quindi superare l'esame.

Dal secondo gruppo di attività emerge che poco più della metà degli studenti riesce a superare l'esame entro la fine del primo anno, ma dopo il primo semestre.

Nel terzo gruppo di attività è presente un ciclo incentrato sull'attività "Ritirato", leggermente più frequente rispetto agli esami di Fisica Generale analizzati in precedenza, ma comunque rimane un'attività poco popolare tra gli studenti: infatti, si nota che l'arco che permette di saltare questo loop è maggiormente marcato, il che significa che è un percorso più utilizzato.

La maggior parte degli studenti, come era emerso anche dall'analisi temporale condotta precedentemente, riesce a superare l'esame entro il primo semestre o, al massimo, entro il primo anno, anche se c'è comunque una parte degli studenti che effettuano le attività "End third semester" e "End second year".

Analisi matematica 1, tuttavia, è un esame caratterizzato da un'alta percentuale di assenti, come si nota anche dall'elevata frequenza rilevata nel ciclo dell'attività "Assente".

5.3.1 Analisi Matematica 1 – Studenti in corso

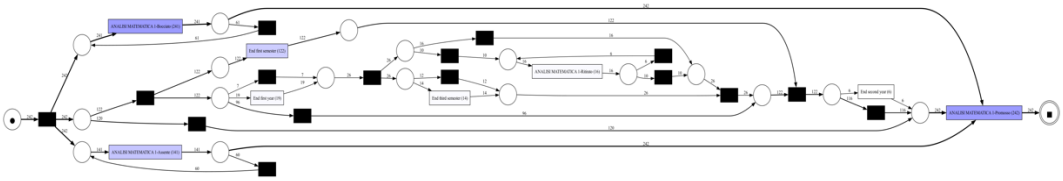

Figura V.10 – Studenti in corso: Processo Analisi Matematica 1

Il processo estratto con un valore di 0.2 per la gestione del rumore, ha un livello di corrispondenza con il file log di 0.897, una precisione di 0.569, generalizzazione 0.808 e semplicità 0.647.

Si nota la presenza di due cicli: uno attorno all'attività "Bocciato", effettuata da molti studenti, che successivamente porta alla fine del processo; e l'altro attorno all'attività "Assente", anch'essa molto frequente.

Il gruppo di attività centrali aggiunge informazioni relative a quando gli studenti effettuano le attività: perciò si nota che circa la metà degli studenti supera l'esame rispettando le tempistiche previste, gran parte invece lo supera al secondo semestre, e solo pochi studenti lo rimandano ulteriormente. Infatti, la percentuale di studenti "precoci" che dà l'esame di Analisi Matematica 1 è molto elevata, circa il 60%, con un elevato distacco dagli studenti fuori corso, la quale percentuale, invece, è molto bassa.

Inoltre, come si notava anche nel processo generale, anche molti degli studenti che si laureano in corso si assentano spesso agli appelli.

5.3.2 Analisi Matematica 1 – Studenti un anno fuori corso

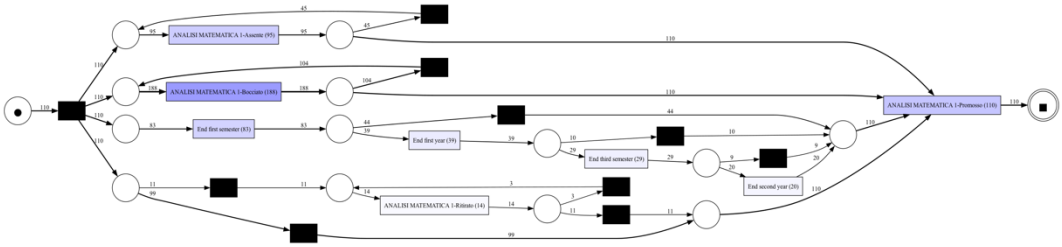

Figura V.11 - Studenti un anno fuori corso: Processo Analisi Matematica 1

Il processo relativo agli studenti laureati in ritardo di un anno è stato estratto imponendo come threshold un valore di 0.2, e si è ottenuto così un modello con una fitness di 0.9, una precisione di 0.764, con capacità di generalizzare la realtà del 0.80 e semplicità pari a 0.647.

Il modello ottenuto è molto simile al modello relativo agli studenti laureati in corso; tuttavia, si nota che la maggior parte degli studenti supera l'esame dopo il primo semestre. Inoltre, si hanno quattro gruppi di attività in parallelo: uno per il ciclo relativo all'attività "Assente", per l'attività "Bocciato" e "Ritirato", e l'altro gruppo contiene le attività che indicano i semestri.

5.3.3 Analisi Matematica 1 – Studenti fuori corso

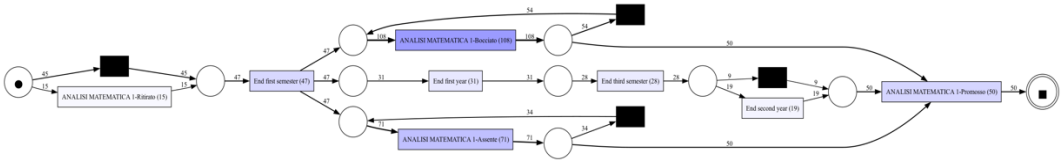

Figura V.12 - Studenti fuori corso: Processo Analisi Matematica 1

Il processo riguardante Analisi Matematica 1 per gli studenti fuori corso è stato estratto con un threshold di 0.4, ottenendo una fitness e precisione rispettivamente di 0.84 e 0.83, mentre generalizzazione e semplicità sono pari a 0.82 e 0.67.

Il processo inizia con l'attività "Ritirato", la quale può essere saltata, per poi arrivare ad "End first semester": ciò significa che nessuno di questi studenti supera l'esame entro il periodo stabilito dal piano di studi. La maggior parte di questi studenti, tuttavia, supera l'esame anche dopo il secondo anno.

Inoltre, si nota un arco molto più spesso che attraversa l'attività "Bocciato" ed "Assente", a sottolineare l'alta frequenza che caratterizza queste attività.

Come descritto in precedenza, infatti, la percentuale di superamento entro il primo semestre dell'esame Analisi Matematica 1, per gli studenti fuori corso, è molto bassa, pari al 6%, coerentemente a quanto emerge dai processi appena visualizzati.

5.4 ELETTROTECNICA

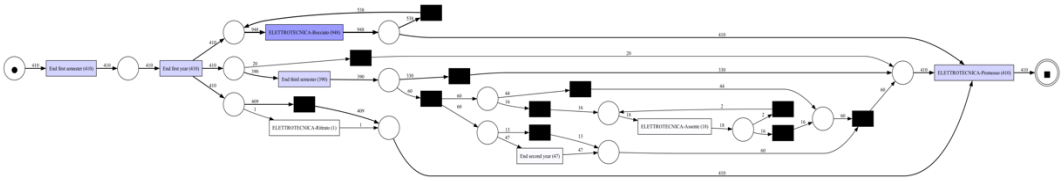

Figura V.13 - Processo Elettrotecnica

Elettrotecnica è un esame che, dalle statistiche precedenti, è risultato un potenziale blocco per gli studenti a causa dell'alto tasso di bocciature, superiore al 200%. Inoltre, dall'analisi temporale, risulta una bassa percentuale di superamento entro il periodo in cui è stato stabilito dal piano di studi, ossia il primo semestre del secondo anno: tuttavia si nota un miglioramento dagli studenti iscritti nell'anno accademico 2015/2016 e 2016/2017 agli anni dal 2017/2018 in poi.

Si è scelto, perciò, di approfondire l'analisi estraendo prima il processo relativo a tutti gli studenti che preparano questo esame, per poi esaminare le differenze tra i processi delle diverse categorie di studenti.

Il valore soglia scelto è di 0.1, risultando in un processo altamente conforme al log, con una fitness di 1, una precisione di 0.98, una generalizzazione pari a 0.85 e semplicità di 0.73.

Il processo inizia con le attività temporali "End first semester" e "End first year", dato che Elettrotecnica è una materia del primo semestre del secondo anno. Successivamente si notano tre gruppi di attività in parallelo.

L’arco relativo all’attività “Bocciato” è molto spesso, a causa dell’alta frequenza di quest’ultima: infatti, come sopradetto, l’esame di Elettrotecnica ha un elevato tasso di bocciature, ed è per questo che si ritiene che sia un potenziale blocco per gli studenti.

Si nota, poi, che alcuni studenti sono in grado di superare l’esame prima del secondo semestre del secondo anno. Altri studenti, invece, superano Elettrotecnica prima della fine del secondo anno, ed alcuni di questi si prenotano all’appello senza presentarsi, prima di arrivare alla fine del processo con la promozione.

Come si evince dal processo, vi sono pochi studenti, rispetto al totale, che superano l’esame dopo la fine del secondo anno, e solo uno di questi si ritira durante l’appello.

5.4.1 Elettrotecnica - Studenti in corso

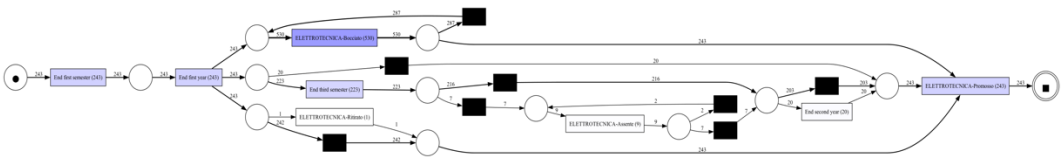

Figura V.14 - Studenti in corso: Processo Elettrotecnica

Il processo è stato estratto con un valore soglia per la gestione dei comportamenti infrequenti pari a 0.1, infatti l’aderenza al log è massima, con fitness pari a 1, precisione di 0.984, generalizzazione di 0.83 e semplicità 0.76.

Il processo relativo agli studenti in corso è molto simile al processo riguardante la totalità degli alunni: l'unica differenza riguarda le frequenze, più basse nel loop dei bocciati, trattandosi anche di un numero inferiore di studenti.

5.4.2 Elettrotecnica - Studenti un anno fuori corso

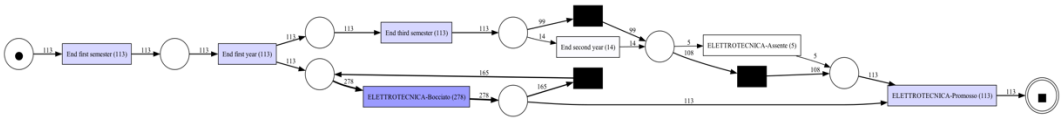

Figura V.15 Studenti un anno fuori corso: Processo Elettrotecnica

Il processo relativo agli studenti laureati un anno fuori corso è stato estratto senza eliminare i comportamenti infrequenti, dato che il modello risultante aveva già di per sé una semplicità abbastanza elevata, di 0.674, una fitness massima, precisione e generalizzazione rispettivamente di 0.747 e 0.762.

Come già individuato in precedenza, il loop dell'attività "Bocciato" presenta un arco più marcato a causa dell'elevata frequenza delle bocciature; perciò, significa che molti studenti vengono bocciati, e perciò attraversano quel percorso, prima di essere promossi.

Inoltre, si nota che molti studenti effettuano l'esame il semestre successivo a quello previsto.

5.4.3 Elettrotecnica - Studenti fuori corso

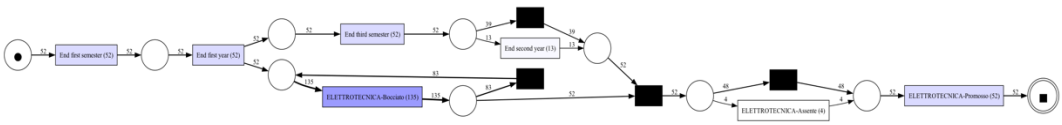

Figura V.16 Studenti fuori corso: Processo Elettrotecnica

Anche per quest'ultimo processo, si è scelto di non utilizzare alcun valore soglia per filtrare i comportamenti infrequenti, e come risultato si è ottenuto un modello avente: fitness pari a 1, precisione di 0.868, generalizzazione 0.821 e semplicità 0.778.

Il processo relativo agli studenti laureati fuori corso è molto simile al modello appena descritto. Si nota anche qui, dalle elevate frequenze del ciclo intorno all'attività "Bocciato", il motivo principale per il quale questo esame è stato considerato un blocco per gli studenti, ossia l'elevato tasso di bocciature.

5.5 ANALISI MATEMATICA 2

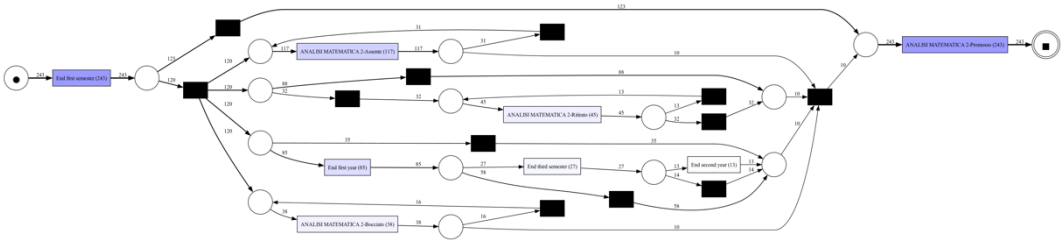

Figura V.17 - Processo Analisi Matematica 2

Infine, si procede con l'analisi del modello di comportamento degli studenti relativo alla preparazione del corso Analisi Matematica 2.

Questo esame, per gli studenti iscritti all'anno accademico 2015/2016 e 2016/2017, aveva un tasso di superamento entro il secondo semestre del primo anno del 31.75%, mentre questa percentuale è raddoppiata per gli studenti iscritti negli anni accademici successivi.

Inoltre, Analisi Matematica 2 è stato identificato come un esame problematico a causa della tendenza degli studenti di rimandare l'esame ai semestri successivi e, soprattutto, dell'elevato tasso di assenti.

Il processo è stato estratto utilizzando come valore soglia per filtrare i comportamenti meno frequenti di 0.2, ottenendo un modello con una fitness di 0.852, precisione 0.836, generalizzazione di 0.87 e semplicità 0.633.

La prima attività del processo è "End first semester", dato che l'esame è previsto nel secondo semestre del primo anno, dalla quale si dipartono due attività

mutuamente esclusive: da una di queste si diramano quattro possibilità in parallelo, e l'altro percorso viene seguito dagli studenti che riescono a superare immediatamente l'esame.

I quattro gruppi di attività paralleli si riferiscono ad un loop per l'attività "Ritirato", per "Assente" e per "Bocciato", mentre il terzo ramo di attività permette di inserire riferimenti temporali. Da quest'ultimo si evince, infatti, che la maggior parte degli studenti supera l'esame dopo il primo semestre o comunque entro la fine del primo anno; perciò, pochi studenti rimandano l'esame oltre la fine del secondo anno.

Tuttavia, come era emerso anche dalle statistiche precedenti, il ciclo relativo all'attività "Assente" ha una frequenza molto elevata.

Da questa analisi emerge, perciò, che l'esame rappresenta un collo di bottiglia nella carriera accademica degli studenti perché, questi ultimi, impiegano più tempo nella preparazione dello stesso dato che è prassi comune non presentarsi all'appello finché non si ritiene di essere abbastanza preparati: perciò il tasso di bocciature è basso per il suddetto corso.

5.5.1 Analisi Matematica 2 – Studenti in corso

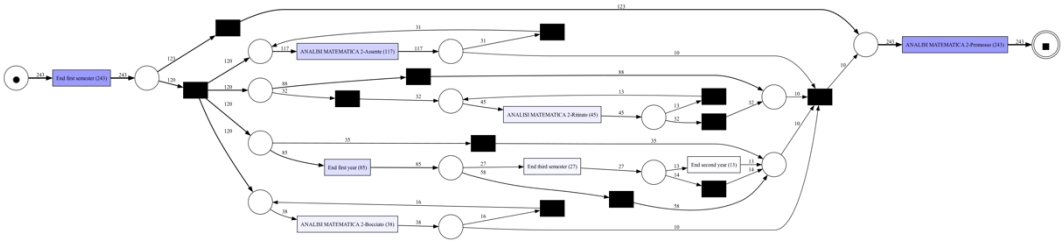

Figura V.18 - Studenti laureati in corso: Processo Analisi Matematica 2

Il threshold scelto per filtrare il rumore durante l'estrazione del processo è di 0.2, con un'aderenza al log ottenuta di 0.896, una precisione di 0.842, capacità di generalizzazione di 0.833 e semplicità di interpretazione del modello di 0.643.

Il modello estratto dai dati relativi agli studenti laureati in corso è molto simile al modello generale appena descritto, ed anche per questi studenti si nota un'elevata frequenza dell'attività "Assente". La metà di questi studenti non incontra nessun intoppo nel proprio percorso e supera l'esame entro il periodo stabilito, riuscendo a superarlo subito. Una piccola parte di questi studenti, invece, rimanda l'esame ai semestri successivi. Si nota, tuttavia, una bassa frequenza dell'attività "Bocciato".

5.5.2 Analisi Matematica 2 – Studenti un anno fuori corso

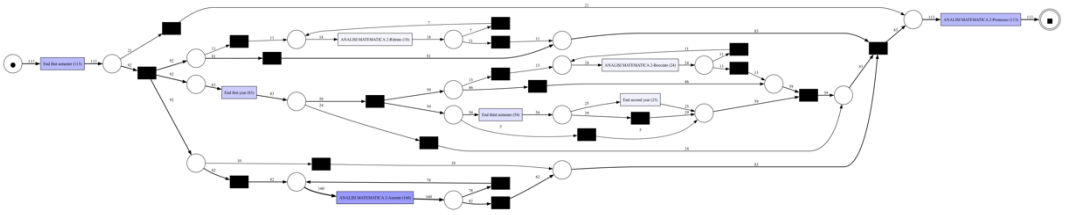

Figura V.19 - Studenti laureati un anno fuori corso: Processo Analisi Matematica 2

Il modello ottenuto riguardante il comportamento degli studenti fuori corso di un anno nel preparare Analisi Matematica 2, ha una fitness di 0.92, precisione di 0.738, una generalizzazione di 0.798 e semplicità 0.667. Il valore soglia utilizzato per la gestione del rumore è di 0.2.

Nel modello, seppur molto simile a quello descritto precedentemente, si può notare un maggiore ritardo da parte degli studenti nel superare l'esame: infatti le attività "End first year" e "End third semester" hanno una frequenza maggiore rispetto a quelle nel processo relativo agli studenti laureati in corso, proporzionalmente al numero di studenti delle due categorie.

5.5.3 Analisi Matematica 2 – Studenti fuori corso

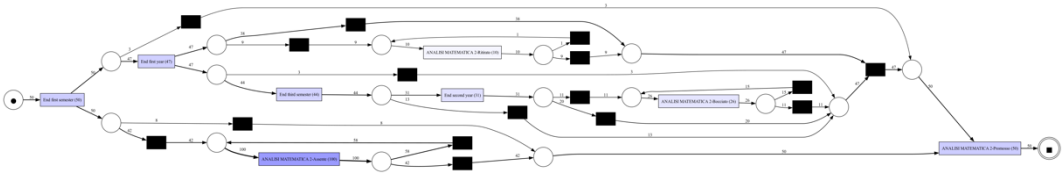

Figura V.20- Studenti laureati fuori corso: Processo Analisi Matematica 2

Il processo estratto, con un valore di filtro del rumore di 0.1, è caratterizzato da: fitness massima, precisione di 0.751, generalizzazione pari a 0.723 e semplicità 0.677.

Gli studenti che si laureano fuori corso, come ci si poteva anche aspettare dal tasso di superamento dell'esame entro il periodo prestabilito pari al 6%, sono gli studenti che rimandano di più quest'esame: infatti quasi la totalità di loro lo supera dopo il primo anno, molti dopo il secondo semestre del secondo anno ed anche dopo la fine di quest'ultimo.

Inoltre, come è caratteristico di quest'esame in ogni categoria di studenti considerata, l'attività "Assente" ha un arco più spesso, a delineare l'alta frequenza dell'attività.

5.6 COME LAUREARSI IN TEMPO?

In conclusione, dalle differenze emerse tra i processi relativi agli studenti che riescono a laurearsi in corso e quelli fuori corso, si può affermare che il "segreto" per avere una carriera accademica di successo è effettuare il prima possibile questi esami "critici", perché si ritiene che rimandare la loro preparazione allunghi i tempi di laurea.

Infatti, ciò che caratterizza gli studenti "precoci" è l'aver superato gli esami entro il tempo stabilito dal piano di studi.

VI. PREDIZIONE

Una volta approfondita l'analisi relativa agli esami considerati un "intoppo" nel percorso universitario, focalizzandosi sui processi relativi ai comportamenti degli studenti nella preparazione dei suddetti esami, ora si andranno a considerare i principali fattori che influenzano il rendimento scolastico degli studenti e se è possibile o meno prevedere il voto medio degli studenti alla fine del loro percorso di laurea.

6.1 PREDIZIONE DELLE PERFORMANCE IN BASE AL NUMERO DEGLI ESAMI

6.1.1 Primo semestre

Di seguito un estratto del set di dati utilizzato per addestrare il modello di Regressione Logistica per la predizione della performance degli studenti, utilizzando come informazioni preliminari il numero di esami svolti durante il primo semestre ed il tempo, calcolato in giorni, che gli stessi impiegano per arrivare al termine del percorso di studi.

Per "High performance" si intende una media dei voti superiore a 27.

STU_ID	N esami	Tempo laurea	High performance
000B9…AC952	2	1123	1
0049D…72EEF	2	1018	1
00521…FAA91	2	1133	0
00E61…A816D	1	1216	0

Tabella VI.1 Estratto del dataset per il modello predittivo basato sul numero degli esami

Per l'addestramento del modello, si è eliminata la colonna relativa all'identificatore univoco degli studenti e si è applicata una standardizzazione dei dati. La colonna "STU_ID" è stata eliminata in quanto contiene un valore personalizzato per ogni studente; perciò, non contribuisce all'estrazione di informazioni generali per l'intera popolazione statistica.

Di seguito una tabella riassuntiva delle performance del modello.

Categoria	Precision	Recall	F1-Score	Supporto
0	0.82	0.93	0.87	84
1	0.79	0.56	0.66	39
Media pesata	0.81	0.81	0.80	123

Tabella VI.2 - Performance del modello di Regressione Logistica basato sul numero di esami effettuato nel primo semestre

Si nota una capacità maggiore del modello di classificare la categoria "0", ossia gli studenti che hanno una media inferiore a 27, perché all'interno del set di dati sono presenti più studenti di questa categoria. Tuttavia, guardando la metrica per la

valutazione delle performance F1-Score, si nota una buona capacità generale del modello di predire se gli studenti avranno una media elevata o meno.

Analizzando i coefficienti del modello, invece, si nota che la variabile "Numero di esami" ha un coefficiente di 2.605, mentre "Tempo laurea" di -2.45: perciò l'aumentare del numero degli esami ha un'influenza positiva sulla probabilità di essere uno studente con una "prestazione elevata"; mentre l'aumentare del tempo impiegato per il completamento degli studi influisce negativamente sulla suddetta probabilità.

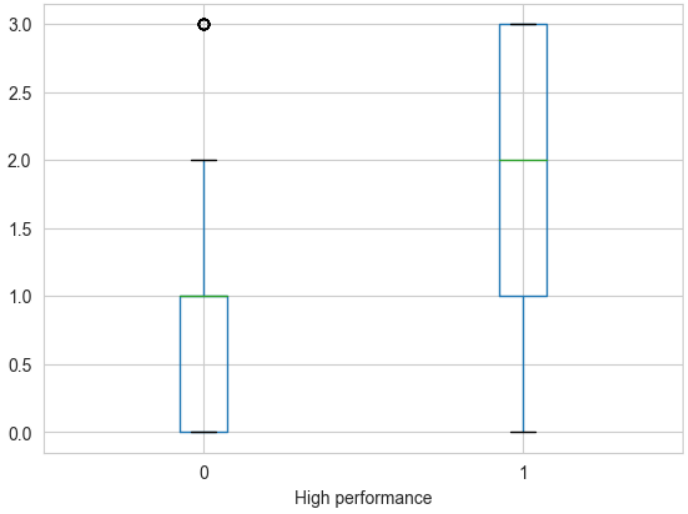

Figura VI.1 - Boxplot della distribuzione del numero di esami dati nel primo semestre

Quando si esamina la distribuzione del numero di esami superati dagli studenti, si nota una differenza significativa tra due gruppi: quelli con una media elevata e quelli con una media inferiore a 27. Nel gruppo degli studenti con una media

elevata, la distribuzione è caratterizzata da una tendenza verso l'alto, con una mediana pari a 2. Ciò significa che la maggior parte di questi studenti ha superato almeno due esami. D'altra parte, nel gruppo degli studenti con una media inferiore a 27, la distribuzione ha una mediana di 0, il che indica che molti di loro non hanno superato alcun esame nel primo semestre. In sintesi, gli studenti con una media più alta sembrano avere una tendenza a superare un numero di esami maggiore, mentre gli studenti con una media inferiore tendono a superarne meno.

6.1.2 Primo anno

Dopo aver allenato il modello alla predizione delle performance degli studenti in base al numero di esami svolti nel primo semestre, lo stesso modello è stato applicato anche ai dati relativi agli esami del primo anno.

Di seguito le performance ottenute:

Categoria	Precision	Recall	F1-Score	Supporto
0	0.77	0.90	0.83	80
1	0.72	0.49	0.58	43
Media pesata	0.75	0.76	0.74	123

Tabella VI.3 - Performance del modello di Regressione Logistica basato sul numero di esami effettuato nel primo anno

Anche nel caso dei dati relativi all'intero primo anno, gli studenti con una media inferiore al 27 sono all'incirca il doppio rispetto agli studenti con una media superiore.

Analizzando i coefficienti relativi alle variabili “Numero di esami” e “Tempo laurea”, rispettivamente di 3.769 e -2.153, si nota che il numero di esami superati entro il primo anno ha un’influenza ancora maggiore rispetto a quelli del primo semestre.

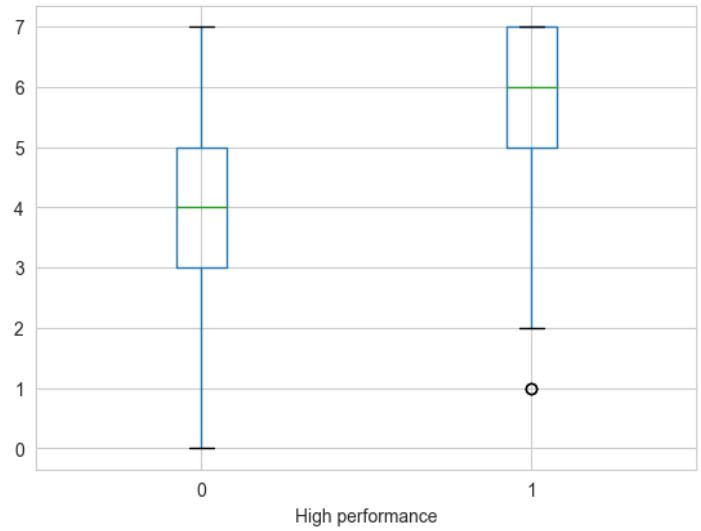

Figura VI.2 - Boxplot della distribuzione del numero di esami dati nel primo anno

Inoltre, analizzando la distribuzione del numero di esami per entrambe le categorie di studenti, si nota che la distribuzione che riguarda gli studenti con una media elevata è più spostata verso l’alto, con mediana pari a 6; mentre la prima distribuzione ha una mediana pari a 4, perciò gli studenti con una media inferiore a 27 solitamente, entro il primo anno, superano meno esami rispetto a coloro che hanno una media più alta.

È un risultato che, da un lato ci si potrebbe aspettare dato che gli studenti più studiosi, e che quindi effettuano più esami, hanno una media più elevata; ma d'altro canto dare meno esami potrebbe anche significare avere più tempo a disposizione per ogni corso ed essere più preparati, e quindi prendere voti più alti.

Si può affermare, perciò, che quanti esami si riescono a superare nel primo semestre, e a maggior ragione nel primo anno, ha una notevole influenza sulle performance accademiche degli studenti.

6.2 PREDIZIONE DELLE PERFORMANCE IN BASE AL TEMPO IMPIEGATO PER LA PREPARAZIONE DEGLI ESAMI

Si vuole indagare ora se ha rilevanza o meno quali esami si effettuano nel primo anno, ossia si vuole rispondere alla domanda: superare o meno durante il primo anno gli esami "critici" ha effetto sul voto di laurea? Data l'informazione su quali esami uno studente ha superato al primo anno, è possibile predire se si laureerà con un voto alto o meno?

Per rispondere a queste domande si è proceduto alla costruzione di due modelli di classificazione, mediante Regressione Logistica e Support Vector Machine, partendo da un set di dati così formato:

STU ID	Fisica Generale 2	Fondamenti di Informatica	Fisica Generale 1	Analisi Matematica 1	Analisi Matematica 2	Algebra Lineare e Geometria	Economia Dell'impresa	High Performance
0..952	41	0	21	48	0	0	49	1
0..EEF	83	0	76	6	0	0	0	1
0..A91	1000	0	7	10	103	125	0	0
0..160	0	0	109	51	0	0	0	0

Tabella VI.4 - Estratto del dataset per il modello predittivo basato sul tempo impiegato per la preparazione degli esami

Anche in questo caso, la variabile "STU_ID" è stata eliminata per una migliore generalizzazione del modello. La variabile da predire, invece, è "High Performance", ossia se lo studente si laureerà con una media superiore al 27 o meno.

6.2.1 Regressione Logistica

Di seguito le performance del modello di Regressione Logistica:

	Precision	Recall	F1-Score	Support
0	0.85	0.82	0.84	56
1	0.64	0.69	0.67	26
Media pesata	0.79	0.78	0.78	82

Tabella VI.5 - Performance del modello di Regressione Logistica basato sul tempo impiegato per la preparazione degli esami effettuati nel primo anno

Il modello presenta buone performance, con un F1-score medio pesato di 0.78. Tuttavia, nella predizione della categoria "0", quindi degli studenti con una media inferiore a 27, performa meglio, a causa di un leggero sbilanciamento delle classi. La capacità del modello di individuare la categoria esatta in base alle informazioni date in input, è misurata anche dalla curva sottostante alla curva ROC, che è di 0.76.

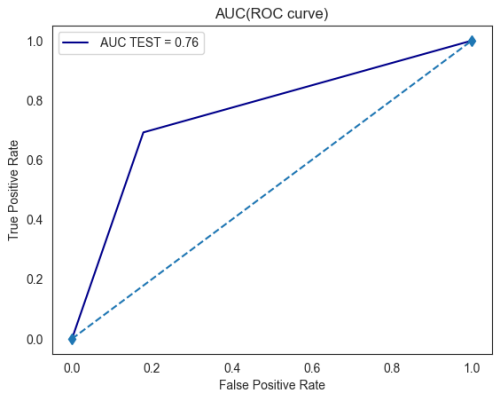

Figura VI.3 - Curva ROC del modello di Regressione Logistica per la predizione delle performance degli studenti

Di seguito, i coefficienti del modello.

	Coefficienti	P-Value
Fisica Generale 2	-1.271	0.00905
Fondamenti di Informatica	-0.204	0.56051
Fisica Generale 1	-1.411	0.00006
Analisi Matematica 1	-0.738	0.01196
Analisi Matematica 2	-0.945	0.04798
Algebra Lineare e Geometria	0.120	0.61992
Economia Dell'impresa	-0.699	0.00185

Tabella VI.6 - Coefficienti del modello di Regressione Logistica

Analizzando i coefficienti, si nota che le materie Fisica Generale 2, Fisica Generale 1 e Analisi Matematica 2 hanno una maggiore influenza negativa sulla variabile "voto": ciò significa che all'aumentare del tempo di preparazione dei suddetti esami diminuisce la probabilità che lo studente rientri tra coloro i quali si laureano con "alte prestazioni".

Infatti, un tempo di preparazione molto elevato, di mille giorni, per costruzione significa che l'esame non è stato sostenuto.

È coerente affermare, perciò, che non superare i suddetti esami, identificati come critici nelle analisi precedenti, porta ad un conseguente decremento della probabilità di riuscire a laurearsi con un voto elevato.

I coefficienti nella regressione logistica indicano, ad esempio, se il tempo di preparazione per l'esame di Fisica Generale 2 aumenta di 1, la probabilità di laurearsi con una media superiore a 27 aumenta di $e^{-1.271}$, ossia la probabilità viene moltiplicata per 0.28, perciò diminuisce.

Si nota, inoltre, che la significatività statistica di questi coefficienti è elevata perché i p-value corrispettivi sono bassi, mentre non è possibile commentare il significato dei coefficienti delle materie Fondamenti di Informatica e Algebra Lineare e Geometria.

6.2.2 Support Vector Machine

Partendo dagli stessi dati, perciò utilizzando le informazioni riguardo il tempo di preparazione degli esami del primo anno, si procede alla predizione delle performance accademiche degli studenti, classificando in "performance elevate" e non, utilizzando anche l'algoritmo di classificazione SVM.

Per la scelta dei migliori parametri del modello è stato creato un ciclo for che ha eseguito tutti i possibili modelli, per scegliere quale fosse il kernel da usare e con

quali valori. Il miglior modello risultante è stato selezionato, usando come kernel "rbf", e come parametri: 'C' = 901 e 'gamma': 0.05.

Le performance del modello sono le seguenti:

	Precision	Recall	F1-Score	Support
0	0.87	0.82	0.84	56
1	0.66	0.73	0.69	26
Media pesata	0.80	0.79	0.80	82

Tabella VI.7 - Performance del modello SVM basato sul tempo impiegato per la preparazione degli esami effettuati nel primo anno

In questo modello di predizione si nota una performance leggermente migliore rispetto a quello dove si è utilizzato la Regressione Logistica, con una Precision e Recall più alta per la categoria "0", che indica una capacità maggiore di predire la classe di studenti con una media inferiore a 27. Difatti, la metrica F1-score è abbastanza elevata, ossia di 0.80.

Per una ulteriore valutazione delle performance del modello si è calcolata l'area al di sotto della curva ROC, che misura la capacità discriminativa di un modello di classificazione, che in questo caso è elevata perché vicino ad 1, ossia 0.78.

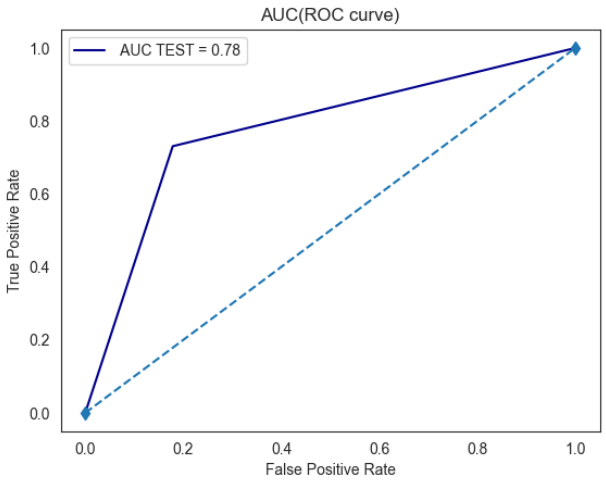

Figura VI.4 - Curva ROC del modello SVM per la predizione delle performance degli studenti

Successivamente, si va ad applicare il modello a dati al di fuori del campione.

Supponendo di avere uno studente che supera in un tempo minimo tutti gli esami, e sottoponendo questo dato al modello, si ottiene come predizione "1"; mentre se si inserisce il valore "1000" in corrispondenza degli esami critici, lo studente verrà etichettato nella categoria "0".

In definitiva, si può concludere che riuscire a superare durante il primo anno gli esami individuati come critici è cruciale per la carriera accademica.

Inoltre, avere un modello predittivo delle performance degli studenti, può essere utile per individuare preventivamente gli studenti potenzialmente "a rischio" e dar loro suggerimenti a riguardo. L'obiettivo principale della comunità accademica globale, infatti, è la diminuzione del livello di dispersione scolastica, una situazione

che comunemente si verifica nel corso del primo anno, poiché questo rappresenta una fase particolarmente delicata.

Si potrebbe anche pensare ad un'applicazione reale del suddetto modello e metterlo a disposizione degli studenti, per far in modo che, in base agli esami da loro superati, gli si mostri la probabilità di rientrare nella categoria di studenti ad "elevate performance": in questo modo, se gli studenti fossero indecisi su quale esame preparare successivamente, potrebbero scegliere quello che permette loro di aumentare questa probabilità.

VII. CONCLUSIONI

I dati raccolti sui processi educativi offrono nuove opportunità per migliorare l'esperienza di apprendimento degli studenti.

Le analisi finora condotte hanno illustrato i modelli di comportamento tipici degli studenti che si approcciano allo studio degli esami presenti nel loro curriculum di laurea. Le tecniche di process mining permettono di tracciare i percorsi degli studenti per individuare i comportamenti da loro adottati nella preparazione degli esami.

In particolare, è interessante individuare le differenze di comportamento che esistono tra gli studenti che si laureano entro la durata normale del corso di studi e coloro che si laureano in ritardo: quest'analisi è finalizzata a scoprire i comportamenti "vincenti" in modo da poter consigliare gli studenti che si approcciano ora agli studi.

L'esplorazione dei comportamenti degli studenti "precoci" e "ritardatari" ha rivelato quali esami potrebbero rappresentare dei colli di bottiglia nel processo di studi; per il corso di laurea Ingegneria Informatica e dell'Automazione, sono stati individuati come tali: Fisica Generale 1, Fisica Generale 2, Analisi Matematica 1, Analisi Matematica 2 ed Elettrotecnica.

Dalle analisi è risultato che questi esami influiscono sui tempi di conseguimento del diploma di laurea degli studenti, rendendoli meritevoli di maggiore attenzione.

Una volta individuati, si è proseguito approfondendo il processo di preparazione di tali esami, scoprendo modelli di comportamento differenti nel modo in cui le categorie di studenti si approcciano allo studio degli stessi.

Infine, utilizzando la regressione logistica ed SVM, si è quantificato l'impatto del tempo impiegato per la preparazione di ogni esame del primo anno, sulle performance degli studenti, dove per performance si intende il voto di laurea superiore alla soglia del 27 o inferiore.

In questa analisi si è diviso il campione analizzato in "studenti ad elevate prestazioni" e non, in modo tale da individuare i fattori che influiscono sul rendimento accademico. Come risultato, si è ottenuto che il numero di esami effettuati durante il primo semestre, e ancora di più nel primo anno, influisce notevolmente sulle performance accademiche.

In definitiva, tutte le analisi che hanno caratterizzato questa tesi hanno evidenziato la necessità di dare priorità agli esami cruciali, soprattutto entro il primo anno.

In conclusione, questa ricerca invita a svolgere maggiori analisi in futuro per migliorare la qualità dell'istruzione, prevedendo il rendimento accademico degli studenti, affinché si possano individuare gli studenti "a rischio" ed elaborare strategie efficaci al loro supporto.

Si ritiene che questo lavoro sia di fondamentale importanza per prevenire l'abbandono degli studi e rendere l'esperienza universitaria meno stressante e, al

contempo, ugualmente, se non maggiormente, utile al bagaglio culturale dello studente. Ci si augura, perciò, che possa porre le basi per studi futuri in tale ottica.

Uno studio futuro potrebbe implicare la scoperta di un modello di predizione migliore che tenga conto di ulteriori dati, inserendo soprattutto informazioni demografiche quali l'età dello studente, il sesso e la scuola superiore di provenienza.

Un aspetto positivo del modello predittivo estratto riguarda la sua duplice utilità: non è solo un'importante risorsa per i professori, che lo potrebbero utilizzare per individuare gli studenti "a rischio" e fornire loro maggiori attenzioni, ma può anche essere un prezioso strumento per gli studenti stessi nel prendere decisioni ponderate sul loro percorso accademico.

L'obiettivo ultimo, seppur ambizioso, che ha ispirato questo studio è il miglioramento del sistema accademico e ciò richiede sforzi continui per tradurre le intuizioni emerse dalla ricerca in raccomandazioni concrete e pratiche. Raccomandazioni personalizzate che mireranno a fornire un valido supporto agli studenti, aiutandoli nel processo decisionale durante le varie fasi della loro carriera accademica, tenendo in considerazione la loro attuale progressione e dando feedback continui sulla loro efficacia.

I risultati ottenuti, quindi, forniscono spunti per le università che cercano di migliorare i corsi da loro offerti per abbattere le difficoltà che gli studenti incontrano durante il loro percorso e smentire gli stereotipi che rendono una facoltà

“più difficile” o meno, per assicurare che la scelta del corso di studi ideale non sia influenzata da paure di questo tipo.

VIII. BIBLIOGRAFIA

1. WIL VAN DER AALST, *Process Mining Data Science in Action*, Springer, Eindhoven, The Netherlands, 2016.

2. YUPEI ZHANG, YUE YUN, RUI AN, JIAQI CUI, HUAN DAI1 AND XUEQUN SHANG, *Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis***.**

3. HAMEED ALQAHERI AND MRUTYUNJAYA PANDA, *An Education Process Mining Framework: Unveiling Meaningful Information for Understanding Students' Learning Behavior and Improving Teaching Quality,* 2022.

4. VAN DER AALST, *Process Mining - Discovery, Conformance and Enhancement of Business Processes*, Springer, 2011.

5. AHER, S. B., & LOBO, L, *Applicability of data mining algorithms for recommendation system in e-learning. In Proceedings of International Conference on Advances in Computing, Communications and Informatics*, 2012.

6. ALEJANDRO PEÑA-AYALA, *Educational data mining: A survey and a data mining-based analysis of recent works.*

7. MUSTAFA YAĞCI, *Educational data mining: prediction of students' academic performance using machine learning algorithms.*

8. RYAN S. BAKER, LIEF ESBENSHADE, JONATHAN VITALE, SHAMYA KARUMBAIAH, *Using Demographic Data as Predictor Variables: A Questionable Choice,* 2023.

9. Documentazione pm4py*: https://pm4py.fit.fraunhofer.de/docs*

10. Documentazione scikit-learn*: https://scikit-learn.org/stable/*

11. Documentazione statsmodel: *https://www.statsmodels.org/stable/index.html*

12. PANG-NING TAN, MICHAEL STEINBACH, ANUJ KARPATNE, VIPIN KUMAR, *Introduction to Data Mining*, Pearson, Seconda edizione, 2019.

13. PAOLO GIUDICI, SILVIA FIGINI*, Applied Data Mining for business and industry,* Wiley, seconda edizione, 2009.

14. VITO RICCI, *Principali tecniche di regressione con R*, 2006.

15. J.C.A.M. BUIJS, B.F. VAN DONGEN, W.M.P. VAN DER AALST, Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity. International Journal of Cooperative Information Systems, 2014.

16. SANDER J.J. LEEMANS, DIRK FAHLAND, WIL M.P. VAN DER AALST, *Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour,* Eindhoven University of Technology, the Netherlands.

17. S.J.J. LEEMANS, D. FAHLAND, W.M.P. VAN DER AALST, *Discovering Block-Structured Process Models From Event Logs - A Constructive Approach,* Eindhoven University of Technology, The Netherlands.

18. WIL M.P. VAN DER AALST, ALESSANDRO BERTI, *Discovering Object-Centric Petri Nets,* 2020.

19. CHRISTIAN W. GU¨NTHER AND WIL M.P. VAN DER AALST, *Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics,* Eindhoven University of Technology

E so che ogni cosa la devo

A mio padre e al suo sudore

Al sorriso di mia madre

Al viso di ogni nonno che proietta amore

A mio fratello piccolo ora più alto di me

E nonna mi protegge sulla stella più bella che c'è

E ai miei amici esauriti

Alle notti felici

- *Articolo 31, Gente che spera*

Ai miei genitori, a cui devo tutto e che mi hanno permesso di arrivare fin qui.

A Simone, non vedo l'ora di vedere quanto in alto arriverai.

A Donato, per essere la mia persona, per essermi sempre stato vicino e per avermi fatto capire che tutto è più bello in due.

Alla bro, per il reciproco miglioramento quotidiano e per avermi regalato due anni spettacolari (non è finita).

A Velia e Francesca, siamo cresciute insieme (cresciute ma non maturate, per fortuna) ed anche se lontane grazie per non aver mai cambiato il nostro legame.

A Valeria, Alessandra, Giulia e Alice, per aver reso Via Indipendenza 18 un posto magico.

A Simone, Damiano, Marco e Daniele, i bros di dark data science, con i quali dalla preparazione degli esami alla preparazione della zizzona di Battipaglia era un attimo.

Ai professori D. Potena e L. Genga, per la loro disponibilità e per averci presentato questo progetto e dato quest'opportunità.