



UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

REPORT PROGETTO

Statistica e Analisi dei Dati

DOCENTI

Dott. Stefano Cirillo

Dott. Luigi Di Biasi

Università degli Studi di Salerno

STUDENTI

Francesco Alessandro Pinto

Matricola: 0522501981

Stefano Guida

Matricola: 0522502054

Anno Accademico 2024-2025

Indice

Elenco delle Figure	v
Elenco delle Tabelle	ix
1 Introduzione	1
1.1 Introduzione al Dominio del Problema	1
1.2 Introduzione al Dataset Selezionato	2
1.3 Obiettivi dell'Analisi	3
1.4 Struttura del Report	4
2 Analisi Univariata	5
2.1 Controlli Preliminari	6
2.2 Indici di Sintesi	7
2.3 Fattori Demografici	9
2.4 Fattori Socio-Economici	12
2.5 Dati sull'Iscrizione e Accesso ai Corsi	17
2.6 Informazioni Primo Anno Accademico	24
2.7 Fattori Macro-Economici	29
2.8 Insight Analisi Univariata	31
3 Analisi Bivariata	33

3.1	Correlazione tra Feature Numeriche	34
3.2	Aggregazione Feature Categoricali	37
3.3	Feature Construction	46
3.4	Fattori Demografici	48
3.5	Fattori Socio-Economici	51
3.6	Dati sull'Iscrizione e Accesso ai Corsi	55
3.7	Informazioni Primo Anno Accademico	63
3.8	Fattori Macro-Economici	70
3.9	Insight Analisi Bivariata	73
4	Test Statistici	74
4.1	Test di Indipendenza del Chi-Quadro	75
4.1.1	V di Cramér	76
4.2	Test di Kruskal-Wallis	77
4.2.1	Epsilon-Quadrato (ϵ^2)	78
4.3	Esecuzione e Risultati dei Test	79
4.3.1	Analisi dei Risultati: Variabili Categoricali	79
4.3.2	Analisi dei Risultati: Variabili Numeriche	81
5	Analisi dei Cluster	84
5.1	Feature Selection	84
5.2	Preprocessing	85
5.2.1	Gestione dei Valori Mancanti	86
5.2.2	Gestione degli Outlier	87
5.2.3	Riduzione della Dimensionalità	89
5.3	Scelta dell'Algoritmo di Clustering	91
5.4	K-means: Scelta del Numero Ottimale di Cluster	92
5.5	Analisi dei Risultati	94
5.5.1	Cluster 2 (Verde)	98
5.5.2	Cluster 1 (Rosso)	98
5.5.3	Cluster 4 (Viola)	99
5.5.4	Cluster 3 (Azzurro)	100
5.5.5	Considerazioni Finali	100

6	Stima delle Medie dei Voti e Differenze di Genere	103
6.1	Stima Intervallare della Media dei Voti	104
6.2	Stima Intervallare della Differenza tra le Medie	105
6.3	1st Year Grade	106
6.3.1	Stima Intervallare della Media	106
6.3.2	Confronto Maschi e Femmine	107
6.4	Admission Grade	108
6.4.1	Stima Intervallare della Media	108
6.4.2	Confronto Maschi e Femmine	109
6.5	Previous Qualification Grade	110
6.5.1	Stima Intervallare della Media	110
6.5.2	Confronto Maschi e Femmine	111
7	Analisi Dataset Sintetico	113
7.1	Strategia di Prompting	113
7.1.1	Struttura del Prompt	114
7.1.2	Struttura del File JSON	115
7.2	Distribuzione delle Feature Categoriche	117
7.3	Distribuzione delle Feature Numeriche	119
7.4	Associazioni con la Variabile Target	123
7.5	Correlazioni tra Feature Numeriche	126
7.6	Test del Chi-Quadro Bilaterale	128
7.6.1	Regola di Scott	128
7.6.2	Metodologia del Test	128
7.6.3	Risultati del Test	129
7.7	Considerazioni Finali	130
8	Conclusioni	132
8.1	Research Question 1	132
8.2	Research Question 2	133
8.3	Research Question 3	134
8.4	Research Question 4	135

Bibliografia

137

Elenco delle figure

2.1	Distribuzione Variabile Target	7
2.2	Analisi Univariata: Gender	9
2.3	Analisi Univariata: Marital Status	10
2.4	Analisi Univariata: Nationality	10
2.5	Analisi Univariata: Age At Enrollment	11
2.6	Analisi Univariata: Debtor	12
2.7	Analisi Univariata: Scholarship Holder	13
2.8	Analisi Univariata: Tuition Fees Up To Date	13
2.9	Analisi Univariata: Mother's / Father's Qualification	14
2.10	Analisi Univariata: Mother's / Father's Occupation	15
2.11	Analisi Univariata: Displaced	17
2.12	Analisi Univariata: Educational Special Needs	18
2.13	Analisi Univariata: Daytime/Evening Attendance	18
2.14	Analisi Univariata: International	19
2.15	Analisi Univariata: Application Mode	19
2.16	Analisi Univariata: Application Order	20
2.17	Analisi Univariata: Course	21
2.18	Analisi Univariata: Previous Qualification	21
2.19	Analisi Univariata: Previous Qualification Grade	22

2.20	Analisi Univariata: Admission Grade	23
2.21	Analisi Univariata: Curricular Units Credited	25
2.22	Analisi Univariata: Curricular Units Enrolled	25
2.23	Analisi Univariata: Curricular Units Evaluations	26
2.24	Analisi Univariata: Curricular Units Approved	27
2.25	Analisi Univariata: Curricular Units Grade	28
2.26	Analisi Univariata: Curricular Units Without Evaluations	28
2.27	Analisi Univariata: GDP	30
2.28	Analisi Univariata: Inflation Rate	30
2.29	Analisi Univariata: Unemployment Rate	31
3.1	Matrice di Correlazione tra features quantitative	35
3.2	Scatterplot: Previous Qualification VS Admission Grade	36
3.3	Scatterplot: CU Grade 1st sem. VS 2nd sem.	36
3.4	Effetto Aggregazione "Marital Status"	38
3.5	Effetto Aggregazione "Nationality"	39
3.6	Effetto Aggregazione "Mother's / Father's Qualification"	41
3.7	Effetto Aggregazione "Mother's / Father's Occupation"	41
3.8	Effetto Aggregazione "Application Mode"	43
3.9	Effetto Aggregazione "Previous Qualification"	44
3.10	Effetto Aggregazione "Course"	45
3.11	Correzione Anomalia di Application Order	46
3.12	Analisi Univariata: Completed Exams Ratio	47
3.13	Analisi Univariata: Passed Exams Ratio	48
3.14	Analisi Bivariata: Gender VS Target	49
3.15	Analisi Bivariata: Marital Status VS Target	50
3.16	Analisi Bivariata: Nationality VS Target	50
3.17	Analisi Bivariata: Age At Enrollment VS Target	51
3.18	Analisi Bivariata: Debtor VS Target	52
3.19	Analisi Bivariata: Tuition Fees Up to Date VS Target	53
3.20	Analisi Bivariata: Scholarship Holder VS Target	53
3.21	Analisi Bivariata: Parent's Qualification VS Target	54

3.22	Analisi Bivariata: Parent's Occupation VS Target	55
3.23	Analisi Bivariata: Displaced VS Target	56
3.24	Analisi Bivariata: BES VS Target	57
3.25	Analisi Bivariata: Attendance VS Target	57
3.26	Analisi Bivariata: International VS Target	58
3.27	Analisi Bivariata: Application Mode VS Target	59
3.28	Analisi Bivariata: Application Order VS Target	60
3.29	Analisi Bivariata: Previous Qualification VS Target	60
3.30	Analisi Bivariata: Course VS Target	61
3.31	Analisi Bivariata: Previous Qualification Grade VS Target	62
3.32	Analisi Bivariata: Admission Grade VS Target	63
3.33	Analisi Bivariata: CU Credited VS Target	64
3.34	Analisi Bivariata: CU Enrolled VS Target	65
3.35	Analisi Bivariata: CU Evaluations VS Target	66
3.36	Analisi Bivariata: CU Approved VS Target	66
3.37	Analisi Bivariata: CU Grade VS Target	67
3.38	Analisi Bivariata: CU Without Evaluations VS Target	68
3.39	Analisi Bivariata: Completed Exams Ratio VS Target	69
3.40	Analisi Bivariata: Passed Exams Ratio VS Target	70
3.41	Analisi Bivariata: GDP VS Target	71
3.42	Analisi Bivariata: Inflation Rate VS Target	72
3.43	Analisi Bivariata: Unemployment Rate VS Target	72
5.1	Regressione Lineare: 1st Sem. Grade / 2nd Sem. Grade	88
5.2	Screeplot PCA	91
5.3	Elbow Method: WCSS per $K = 1, \dots, 10$	93
5.4	Silhouette Score medio per $K = 1, \dots, 10$	93
5.5	Visualizzazione 2D dei Cluster	95
5.6	Clustering: Distribuzione delle features nei Cluster (Parte 1)	96
5.7	Clustering: Distribuzione delle features nei Cluster (Parte 2)	97
5.8	Clustering: Distribuzione delle Variabile Target nei Cluster	102
7.1	Dataset Sintetico: analisi "Debtor"	118

7.2	Dataset Sintetico: analisi "Mother's Qualification"	118
7.3	Dataset Sintetico: analisi "Age At Enrollment"	120
7.4	Dataset Sintetico: analisi "Curricular Units Enrolled"	120
7.5	Dataset Sintetico: analisi "Admission Grade"	121
7.6	Dataset Sintetico: analisi "Inflation Rate"	122
7.7	Dataset Sintetico: Mother's Qualification VS Target	123
7.8	Dataset Sintetico: Age At Enrollment VS Target	124
7.9	Dataset Sintetico: Admission Grade VS Target	125
7.10	Dataset Sintetico: CU Grade 1st Sem. VS 2nd Sem.	126
7.11	Dataset Sintetico: Previous Qualification Grade Vs Admission Grade	127

Elenco delle tabelle

2.1	Indici di Sintesi per le Features Quantitative	8
2.2	Fattori Demografici: Descrizione Features	11
2.3	Organizzazione del sistema scolastico portoghese	15
2.4	Fattori Socio-Economici: Descrizione Features	16
2.5	Dati sull'Iscrizione e Accesso ai Corsi: Descrizione Features	23
2.6	Informazioni Primo Anno Accademico: Descrizione Features	29
2.7	Fattori Macro-Economici: Descrizione Features	31
3.1	Aggregazione: "Marital Status"	38
3.2	Aggregazione: "Nationality"	39
3.3	Aggregazione "Mother's / Father's Qualification"	40
3.4	Aggregazione "Mother's / Father's Occupation"	42
3.5	Aggregazione "Application Mode"	43
3.6	Aggregazione "Previous Qualification"	44
3.7	Aggregazione "Course"	45
4.1	Risultati del Test di Indipendenza del Chi-Quadro	80
4.2	Risultati del Test di Kruskal-Wallis	82
7.1	Risultati del test del Chi-Quadro bilaterale	130

CAPITOLO 1

Introduzione

1.1 Introduzione al Dominio del Problema

L'identificazione e la prevenzione dell'abbandono studentesco rappresentano un obiettivo cruciale per le istituzioni di istruzione superiore, in quanto il successo accademico degli studenti influisce sia sul loro percorso personale sia sulla reputazione e l'efficienza dell'ente formativo. Interventi precoci rivolti agli studenti a rischio si sono dimostrati efficaci nel ridurre i tassi di abbandono e favorire il conseguimento degli obiettivi formativi [1, 2].

Negli ultimi anni, l'ampia disponibilità di dati demografici, socioeconomici e accademici, raccolti durante l'immatricolazione e nel corso del percorso di studi, ha favorito il ricorso a tecniche di analisi avanzate, in particolare di machine learning. Queste metodologie consentono di sviluppare modelli predittivi in grado di individuare con tempestività gli studenti che potrebbero incorrere in difficoltà o abbandonare gli studi. In tal modo, le istituzioni possono attuare strategie di supporto mirate, migliorando contemporaneamente la qualità dell'esperienza formativa e l'impiego efficiente delle risorse.

Tuttavia, l'implementazione di tali strumenti predittivi presenta ancora diverse sfide. In primo luogo, la questione dello sbilanciamento delle classi incide negativamente

sulla capacità dei modelli di riconoscere in modo accurato i casi di rischio [2]. Inoltre, la disponibilità di dati completi e affidabili non è sempre garantita, e la qualità del dataset gioca un ruolo determinante sulla robustezza delle previsioni. Infine, resta aperto il problema della generalizzabilità: modelli altamente specializzati in un determinato contesto potrebbero necessitare di adattamenti significativi per essere trasferiti con successo in altre istituzioni o in differenti realtà nazionali [1].

In questo scenario, la letteratura ha proposto diverse soluzioni, fra cui l'impiego di algoritmi di classificazione, l'uso di tecniche di resampling (*e.g.*, SMOTE) per gestire le classi minoritarie e l'adozione di *ensemble methods* o metodi di *boosting* al fine di migliorare ulteriormente le prestazioni predittive. Nonostante i progressi, è evidente che la ricerca in questo ambito debba proseguire, con l'obiettivo di perfezionare i modelli attuali e ampliare la loro applicabilità in contesti sempre più eterogenei e complessi.

1.2 Introduzione al Dataset Selezionato

Il dataset impiegato per il presente progetto, denominato “Predict Students’ Dropout and Academic Success”, è stato reso disponibile dall’UCI Machine Learning Repository ed è stato originariamente costruito a partire da diversi database disgiunti del Polytechnic Institute of Portalegre (IPP) in Portogallo [1, 2]. Esso raccoglie i dati di studenti iscritti tra gli anni accademici 2008/09 e 2018/19 a corsi di laurea eterogenei, quali agronomia, design, infermieristica, giornalismo, management e ingegneria. Inizialmente, il dataset era focalizzato soltanto sulle informazioni disponibili al momento dell’immatricolazione, vale a dire:

- **Fattori demografici:** età, genere, stato civile, nazionalità, ecc.
- **Fattori socio-economici:** titolo di studio e occupazione dei genitori, stato di borsista, regolarità del pagamento delle tasse universitarie, ecc.
- **Dati sull’Iscrizione e Accesso ai Corsi:** voto di ammissione o criteri d’ingresso, modalità e ordine di iscrizione al corso, qualifica o diploma precedente, ecc.
- **Fattori macro-economici:** dati economici nazionali (Portogallo).

Successivamente, il dataset è stato arricchito con ulteriori indicatori di performance accademica riguardanti il primo anno di corso, con particolare riferimento all'andamento nel primo e nel secondo semestre. Gli autori hanno formulato il problema di ricerca come un problema classificazione multiclasse, suddividendo gli studenti in tre categorie principali [1]:

- **Graduate:** studenti che completano il percorso di studi entro la durata prevista;
- **Enrolled:** studenti ancora attivamente iscritti al termine del periodo canonico;
- **Dropout:** studenti che abbandonano il percorso prima del completamento.

Gli autori hanno dichiarato di aver già effettuato un processo di pulizia e preparazione dei dati (*preprocessing*), con lo scopo di gestire anomalie, valori mancanti e outlier inspiegabili [2], tuttavia non sono note nel dettaglio tutte le operazioni effettuate.

1.3 Obiettivi dell'Analisi

Alla luce delle informazioni fornite dal dataset selezionato e delle problematiche evidenziate riguardo al rischio di abbandono e alla performance accademica, sono state formulate le seguenti Research Questions:

Q RQ₁. *Quali fattori sono maggiormente associati all'esito accademico (Graduate, Enrolled, Dropout) degli studenti del Polytechnic Institute of Portalegre (IPP)?*

Q RQ₂. *È possibile identificare gruppi distinti di studenti sulla base dei loro comportamenti e delle loro performance accademiche nel primo anno di studi, così da attuare adeguate strategie di supporto?*

Q RQ₃. *Qual è l'intervallo di confidenza al 99% per la media dei voti ottenuti dagli studenti? Inoltre, qual è l'intervallo di confidenza al 99% per la differenza tra la media dei voti degli studenti maschi e quella delle studentesse femmine?*

Q RQ₄. *È possibile, attraverso un Large Language Model (LLM), generare un dataset sintetico con proprietà e distribuzioni statistiche simili a quelle del dataset originale?*

1.4 Struttura del Report

Il presente report è suddiviso nei seguenti capitoli principali:

- **Capitolo 1 - Introduzione:** viene introdotto il dominio del problema e descritto il dataset utilizzato per l'analisi. Inoltre, vengono presentati gli obiettivi dello studio, formulati sotto forma di Research Questions (RQ);
- **Capitolo 2 - Analisi Univariata:** vengono descritte le analisi preliminari condotte per esplorare il dataset, con particolare attenzione alla distribuzione delle singole variabili, all'identificazione di anomalie, outlier e pattern rilevanti;
- **Capitolo 3 - Analisi Bivariata:** vengono illustrate le operazioni effettuate per valutare le possibili associazioni tra le diverse variabili e la variabile target, fornendo una prima evidenza delle relazioni esistenti;
- **Capitolo 4 - Test Statistici:** vengono descritti i test statistici eseguiti per confermare, con un adeguato livello di significatività, le associazioni rilevate nel capitolo precedente, al fine di rispondere in modo rigoroso alla **RQ1**;
- **Capitolo 5 - Analisi dei Cluster:** viene presentato il processo di clustering, descrivendo le tecniche impiegate, le operazioni eseguite e i risultati ottenuti, con l'obiettivo di rispondere alla **RQ2**;
- **Capitolo 6 - Stima delle Medie dei Voti e Differenze di Genere:** vengono illustrate le metodologie di stima intervallare e di confronto tra popolazioni applicate per rispondere alla **RQ3**;
- **Capitolo 7 - Analisi del Dataset Sintetico:** viene descritta la generazione del dataset sintetico e analizzate brevemente le principali differenze rispetto al dataset reale, evidenziando eventuali criticità e difficoltà riscontrate, con lo scopo di rispondere alla **RQ4**;
- **Capitolo 8 - Conclusioni:** vengono sintetizzate le risposte fornite alle Research Questions e discusse le principali implicazioni dei risultati ottenuti.

CAPITOLO 2

Analisi Univariata

In questo capitolo viene illustrato il processo di analisi univariata delle features presenti nel dataset, con l'obiettivo di esplorare la loro distribuzione e identificare tendenze, dispersione, anomalie e altre caratteristiche rilevanti. Per facilitare l'interpretazione dei risultati, le variabili sono state suddivise in diverse categorie, ciascuna analizzata in una sezione dedicata. Inoltre, in base alla natura delle feature stesse, sono stati utilizzati strumenti specifici di sintesi e visualizzazione.

Le **Variabili Categoriche (qualitative)** rappresentano informazioni suddivise in classi o categorie (es. genere, stato civile, corso di laurea). Per queste variabili sono state utilizzate:

- **Tabelle di frequenza:** per calcolare la distribuzione delle categorie (modalità).
- **PieChart:** per rappresentare graficamente la distribuzione delle categorie di variabili binarie o con poche modalità.
- **Barplot:** per rappresentare graficamente la distribuzione delle categorie di variabili con svariate modalità.

Le **Variabili Quantitative** comprendono variabili numeriche che possono assumere valori discreti o continui (es. età, media dei voti, tasso di inflazione). L'analisi è stata condotta attraverso:

- **Istogrammi:** per analizzare la distribuzione della variabile.
- **Kernel Density Plot:** per stimare la densità e visualizzare la forma della distribuzione delle variabili continue. Tipicamente per questa tipologia di variabili Istogramma e Kernel density plot sono stati combinati.
- **Boxplot:** per individuare valori anomali e analizzare la dispersione dei dati.
- **Indici di Sintesi:** per descrivere le principali caratteristiche della distribuzione (tendenza centrale, dispersione, forma).

È importante notare che nei grafici, quando indicate esplicitamente sopra ciascuna modalità, le frequenze relative sono state approssimate per questioni di praticità. Pertanto, le frequenze indicate come 0% si riferiscono a valori molto bassi, prossimi allo 0% ma comunque non nulli

2.1 Controlli Preliminari

In primo luogo, sono stati eseguiti controlli preliminari per verificare la coerenza e la qualità dei dati rispetto alla documentazione fornita. In particolare, è stato riscontrato che:

- Il dataset è composto da 4424 osservazioni (studenti) e 37 variabili (feature).
- Tra le features, 36 sembrano essere variabili numeriche e 1 è categorica (variabile Target). Tuttavia, dalla documentazione emerge che molte delle variabili numeriche rappresentano in realtà variabili categoriche codificate. La loro codifica sarà illustrata nel dettaglio durante l'analisi univariata.
- Il dataset non sembra presentare valori nulli e non contiene duplicati. Tuttavia, durante l'analisi univariata saranno condotti controlli aggiuntivi per individuare eventuali dati mancanti o anomalie.

La variabile target, di particolare interesse per questa analisi, rappresenta l'esito accademico dello studente al termine della durata prevista del corso di studi. Come mostrato in Figura 2.1, la distribuzione delle categorie risulta sbilanciata:

- **Graduate:** Gli studenti che si laureano nei tempi previsti dal corso di studi rappresentano la categoria più numerosa (49,9% – 2210 studenti).
- **Enrolled:** Gli studenti che al termine del periodo standard risultano ancora iscritti al corso rappresentano la categoria meno numerosa (17,9% – 794 studenti).
- **Dropout:** Gli studenti che abbandonano gli studi prima del termine previsto hanno invece una frequenza relativa intermedia tra gli altri due gruppi (32,1% – 1421 studenti).

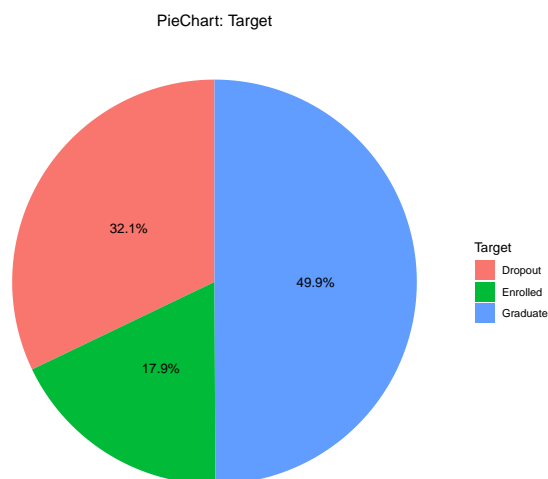


Figura 2.1: Distribuzione Variabile Target

2.2 Indici di Sintesi

Questa sezione presenta gli indici di sintesi calcolati per le variabili quantitative presenti nel dataset. Gli indici statistici descrittivi consentono di riassumere le principali caratteristiche delle distribuzioni, fornendo informazioni sulla tendenza centrale, la dispersione e la forma dei dati. Nella Tabella 2.1 sono riportati gli indici di sintesi calcolati per ciascuna variabile analizzata, i quali possono essere suddivisi in:

- **Indici di Tendenza Centrale:** Forniscono una misura del valore attorno al quale i dati tendono a concentrarsi. In questa analisi sono stati calcolati la **Moda** (valore più frequente), la **Media** (valore medio dei dati) e la **Mediana** (valore centrale della distribuzione).
- **Indici di Dispersione:** Descrivono il grado di variabilità dei dati attorno alla tendenza centrale. In questa analisi sono stati calcolati la **Varianza** (misura la dispersione rispetto alla media), la **Deviazione Standard** (radice quadrata della varianza), il **Coefficiente di Variazione** (misura la dispersione relativa della variabile) e l'**Intervallo Interquartile (IQR)** (differenza tra il terzo e il primo quartile: $Q3 - Q1$).
- **Indici di Forma:** Descrivono la forma della distribuzione dei dati rispetto alla distribuzione normale. In particolare, sono stati calcolati la **Skewness** (misura il grado di simmetria della distribuzione) e la **Curtosi** (misura il grado di appiattimento della distribuzione).

Feature	Moda	Media	Mediana (Q2)	Q1	Q3	Varianza	Dev. Std.	CV (%)	IQR	Skewness	Curtosi
Age At Enrollment	18	23.27	20	19	25	57.57	7.59	32.61	6	2.05	7.12
Previous Qualification Grade	//	132.61	133.1	125	140	173.93	13.19	9.94	15	0.31	3.97
Admission Grade	//	126.98	126.1	117.7	134.6	209.73	14.48	11.41	16.9	0.53	3.66
CU Credited (1st Sem.)	0	0.71	0	0	0	5.57	2.36	332.47	0	4.17	22.18
CU Credited (2nd Sem.)	0	0.54	0	0	0	3.68	1.92	354.09	0	4.63	27.4
CU Enrolled (1st Sem.)	6	6.27	6	5	7	6.15	2.48	39.55	2	1.62	11.93
CU Enrolled (2nd Sem.)	6	6.23	6	5	7	4.82	2.2	35.24	2	0.79	10.13
CU Evaluations (1st Sem.)	8	8.3	8	6	10	17.46	4.18	50.36	4	0.98	8.46
CU Evaluations (2nd Sem.)	8	8.06	8	6	10	15.59	3.95	48.96	4	0.34	5.06
CU Approved (1st Sem.)	6	4.71	5	3	6	9.57	3.09	65.74	3	0.77	6.09
CU Approved (2nd Sem.)	6	4.44	5	3	7	9.09	3.01	67.96	4	0.31	3.84
CU Grade (1st Sem.)	//	10.64	12.29	10.0	12.4	23.46	4.84	45.52	2.4	-1.57	3.91
CU Grade (2nd Sem.)	//	10.23	12.2	9.9	12.5	27.15	5.21	50.94	2.58	-1.31	3.07
CU Without Evaluation (1st Sem.)	0	0.14	0	0	0	0.48	0.69	501.88	0	8.2	92.76
CU Without Evaluation (2nd Sem.)	0	0.15	0	0	0	0.57	0.75	501.46	0	7.27	69.73
GDP	0.32	0	0.32	-1.7	1.79	5.15	2.27	115295	3.49	-0.39	2
Inflation Rate	1.4	1.23	1.4	0.3	2.6	1.91	1.38	112.6	2.3	0.25	1.96
Unemployment Rate	7.6	11.57	11.1	9.4	13.9	7.1	2.66	23.03	4.5	0.21	2

Tabella 2.1: Indici di Sintesi per le Features Quantitative

Gli indici di sintesti calcolati verranno analizzati più nel dettaglio nelle sezioni successive, insieme alle variabili a cui essi fanno riferimento.

2.3 Fattori Demografici

In questa sezione vengono analizzate le caratteristiche demografiche degli studenti, considerando variabili quali genere, età, stato civile e nazionalità. Nella Tabella 2.2 è possibile trovare una panoramica dettagliata delle variabili analizzate. Dall'analisi di questi fattori emergono le seguenti osservazioni:

- **Gender** (Figura 2.2): la distribuzione mostra un marcato squilibrio, con una maggiore presenza di studenti di sesso **0-Female** (64.8% - 2868 studenti) rispetto a quelli di sesso **1-Male** (35.2% - 1556 studenti). Tale disparità potrebbe riflettere una maggiore partecipazione femminile ai corsi offerti o, più in generale, una tendenza legata alla scelta universitaria in determinati ambiti disciplinari.

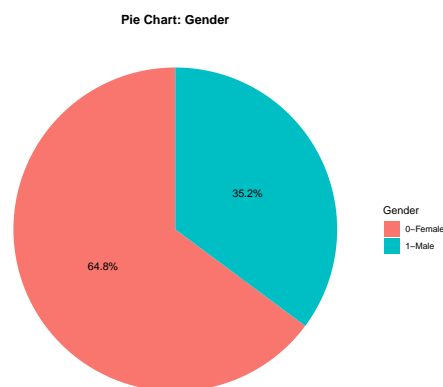


Figura 2.2: Analisi Univariata: Gender

- **Marital Status** (Figura 2.3): possiamo notare che la maggior parte degli studenti è **1-Single** (88.6% - 3920 studenti), mentre una quota nettamente inferiore risulta **2-Married** (8.6% - 380 studenti). Questo dato suggerisce che la popolazione studentesca sia prevalentemente composta da persone tipicamente non coniugate. Il grafico a barre evidenzia uno squilibrio marcato tra le categorie, con alcune modalità caratterizzate da una frequenza prossima allo 0%, rendendole statisticamente poco rilevanti.

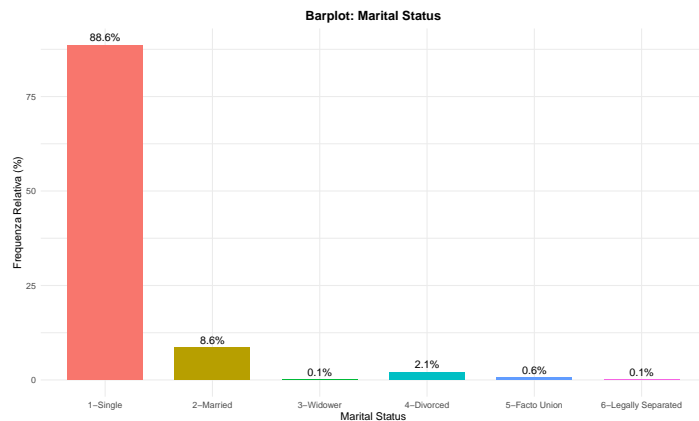


Figura 2.3: Analisi Univariata: Marital Status

- **Nationality** (Figura 2.4): si può osservare una forte predominanza di studenti di nazionalità **1-Portoghese** (97.5% - 4313 studenti), suggerendo che l'istituzione analizzata accolga prevalentemente studenti locali. Le altre nazionalità sono presenti con percentuali significativamente inferiori e prossime allo 0%. Tra gli studenti internazionali, il gruppo più numeroso è rappresentato dai **41-Brazilian** (0.9% - 40 studenti).

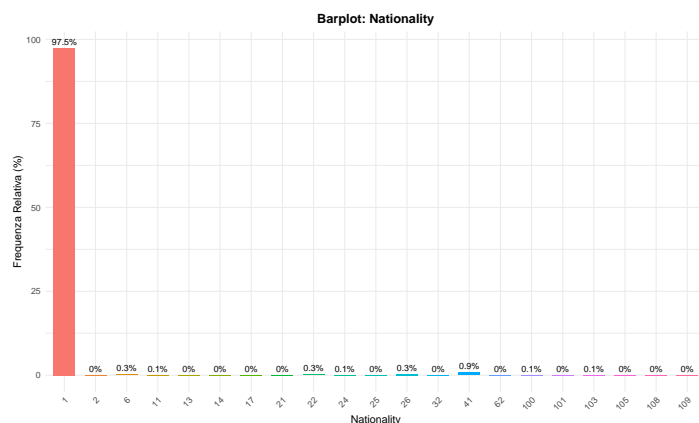


Figura 2.4: Analisi Univariata: Nationality

- **Age At Enrollment** (Figura 2.5): l'età media degli studenti al momento dell'iscrizione è di **23.27 anni**, con una mediana di **20 anni**. La moda, pari a **18 anni**, indica che l'età di immatricolazione più frequente corrisponde a quella immediatamente successiva al completamento della scuola secondaria. L'analisi della distribuzione evidenzia una marcata **asimmetria positiva** (skewness = 2.05), con una concentrazione elevata di studenti nelle fasce d'età più giovani e

una coda lunga verso destra. La dispersione dei dati è significativa, con una deviazione standard di **7.59 anni** e un coefficiente di variazione del **32.61%**, indicando una notevole eterogeneità nelle età di iscrizione. L'intervallo interquartile mostra che il 50% degli studenti si immatricola tra i **19 e i 25 anni**. Inoltre, il boxplot fa notare la presenza di numerosi **outlier**, rappresentati da studenti che si immatricolano in età superiore ai **34 anni**.

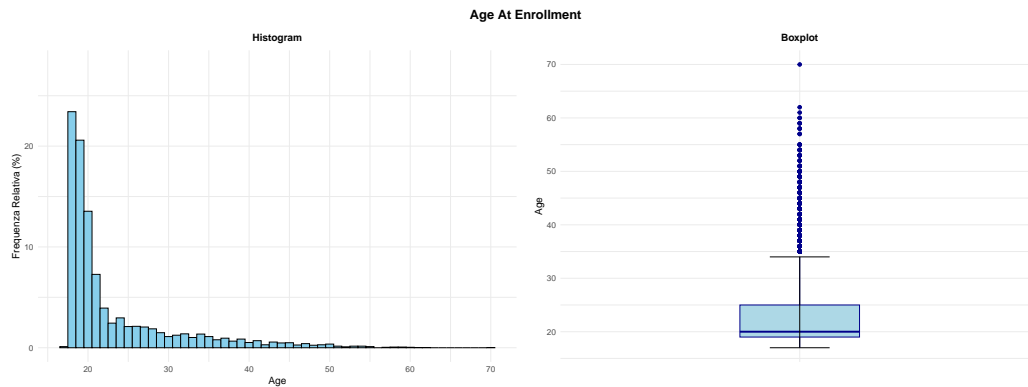


Figura 2.5: Analisi Univariata: Age At Enrollment

Nome	Tipologia	Descrizione	Valori
Gender	Categorica	Sesso dello studente	0-Female, 1-Male
Marital Status	Categorica	Stato civile dello studente	1-Single, 2-Married, 3-Widower, 4-Divorced, 5-Facto Union, 6-Legally Separated
Nationality	Categorica	Nazionalità dello studente	1-Portuguese, 2-German, 6-Spanish, 11-Italian, 13-Dutch, 14-English, 17-Lithuanian, 21-Angolan, 22-Cape Verdean, 24-Guinean, 25-Mozambican, 26-Santomean, 32-Turkish, 41-Brazilian, 62-Romanian, 100-Moldova (Republic of), 101-Mexican, 103-Ukrainian, 105-Russian, 108-Cuban, 109-Colombian
Age At Enrollment	Quantitativa Discreta	Età dello studente al momento dell'immatricolazione	[17 – 70]

Tabella 2.2: Fattori Demografici: Descrizione Features

2.4 Fattori Socio-Economici

In questa sezione vengono analizzate le condizioni socio-economiche degli studenti, considerando variabili come titolo di studio e occupazione dei genitori, possesso di borse di studio e situazione debitoria. La Tabella 2.4 fornisce una panoramica dettagliata delle variabili analizzate. Inoltre, prima di procedere con la lettura delle osservazioni, è consigliabile consultare la Tabella 2.3 per avere maggiori informazioni sull'ordinamento del sistema scolastico Portoghese, in modo da comprendere al meglio le categorie relative alle qualifiche. Dall'analisi fattori socio-economici emergono le seguenti osservazioni:

- **Debtor** (Figura 2.6): La distribuzione risulta fortemente sbilanciata, con una netta predominanza di studenti che non presentano situazioni debitorie (88.6% - 3920 studenti), mentre solamente l'11.4% (504 studenti) degli studenti presenta debiti economici.

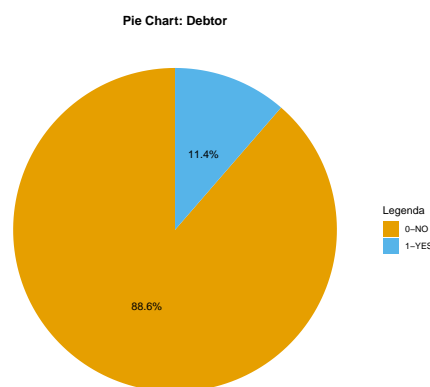


Figura 2.6: Analisi Univariata: Debtor

- **Scholarship Holder** (Figura 2.7): La maggior parte degli studenti non beneficia di una borsa di studio (75.2% - 3327 studenti), mentre il 24.8% (1097 studenti) ne è titolare. Anche in questo caso c'è uno squilibrio evidente tra le due categorie, tuttavia la percentuale di studenti con borsa di studio non è trascurabile, indicando una presenza significativa di sostegno economico all'interno dell'istituto.

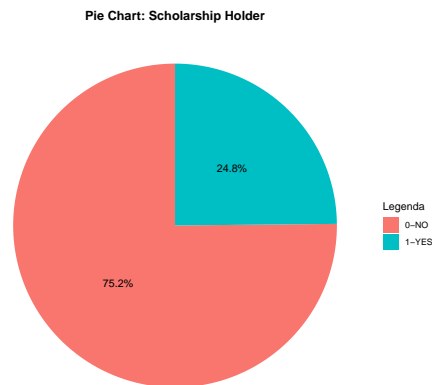


Figura 2.7: Analisi Univariata: Scholarship Holder

- **Tuition Fees Up To Date** (Figura 2.8): la maggior parte degli studenti (88.1% - 3898 studenti) risulta in regola con il pagamento delle tasse universitarie, mentre l'11.9% (526 studenti) presenta arretrati. La distribuzione è fortemente sbilanciata e la bassa percentuale di studenti non in regola suggerisce che la maggioranza riesce a rispettare le scadenze amministrative previste.

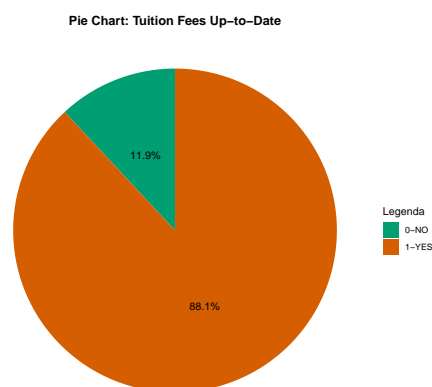


Figura 2.8: Analisi Univariata: Tuition Fees Up To Date

- **Mother's / Father's Qualification** (Figura 2.9): l'analisi del livello di istruzione dei genitori evidenzia una maggiore concentrazione nei livelli di istruzione medio-bassi. La categoria più rappresentata per le madri è **Secondary Education - 12th Year of Schooling or Equivalent (1)** con il 24.2% (1070), seguita da **Basic Education 1st Cycle (4th/5th Year) or Equivalent (37)** con il 22.8% (1009) e **Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv (19)** con il 21.5% (951). Anche per i padri, il pattern è simile, con una predominanza della categoria

Basic Education 1st Cycle (4th/5th Year) or Equivalent (37) con il 27.3% (1208), seguita da **Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv (19)** con il 21.9% (969) e da **Secondary Education - 12th Year of Schooling or Equivalent (1)** con il 21.5% (951). In entrambi i casi possiamo notare una presenza moderatamente bassa di titoli universitari. Confrontando i due grafici possiamo notare come i padri tendano ad avere un livello di istruzione leggermente inferiore rispetto alle madri. Inoltre, è importante notare la presenza di molte categorie che si presentano con una frequenza prossima allo 0%, le quali rappresentano principalmente situazioni particolari di interruzione del percorso accademico prima del conseguimento del titolo associato.

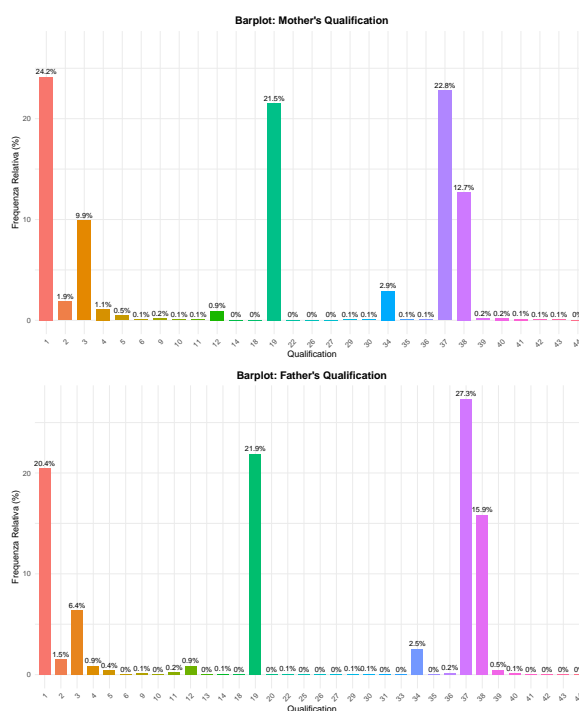


Figura 2.9: Analisi Univariata: Mother's / Father's Qualification

- **Mother's / Father's Occupation** (Figura 2.10): l'analisi della situazione lavorativa dei genitori mostra una predominanza di occupazioni a bassa qualifica. Per quanto riguarda le madri, la categoria più rappresentata è **9-Unskilled Worker** con il 35.6% (1575), seguita da **4-Administrative Staff** con il 18.5% (818) e da **5-Personal Services, Security and Safety Workers and Sellers** con il 12% (531). Anche per i padri si osserva una distribuzione simile, con una maggiore concentrazione nella categoria **9-Unskilled Worker** con il 22.8% (1009), seguita

da **7-Skilled Workers in Industry, Construction and Craftsmen** con il 15.1% (668) e da **5-Personal Services, Security and Safety Workers and Sellers** con l'11.7% (518). È evidente la scarsa presenza di professioni ad alta specializzazione, con frequenze vicine allo 0%. Inoltre, possiamo notare come i padri tendono a svolgere una gamma di lavori più diversificata rispetto alle madri. È interessante notare anche la presenza, seppur con una frequenza bassa, di genitori appartenenti alla categoria **0-Student**.

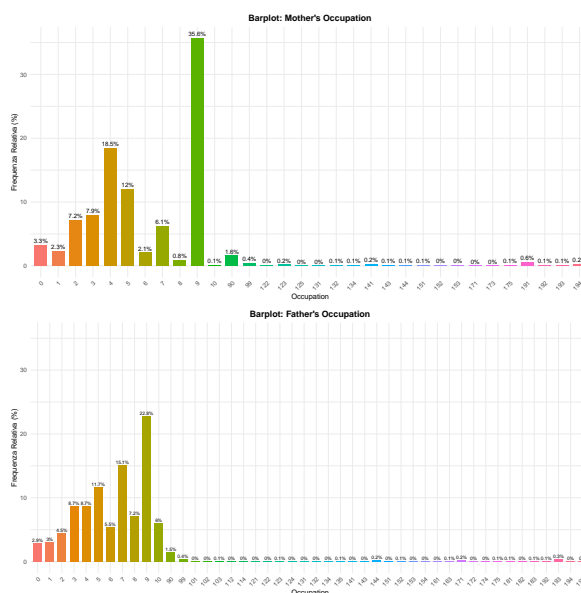


Figura 2.10: Analisi Univariata: Mother's / Father's Occupation

Livello di Istruzione	Età Tipica	Ciclo	Descrizione
Pre-School Education	3-5 anni	-	Facoltativo, include asili nido e scuole materne per la preparazione alla scuola primaria.
Basic Education	6-15 anni	1° Ciclo (6-9 anni)	Dura 4 anni, un solo insegnante per tutte le materie.
		2° Ciclo (10-11 anni)	Dura 2 anni, insegnanti diversi per le varie materie.
		3° Ciclo (12-15 anni)	Dura 3 anni, sistema più strutturato con materie specifiche.
Secondary Education	15-18 anni	-	Dura 3 anni, può essere generale (per l'accesso all'università) o professionale (per l'ingresso nel mondo del lavoro).
Higher Education	18+ anni	-	Include università e politecnici, con lauree di primo livello (Licenciatura), secondo livello (Mestrado) e dottorati (Doutoramento).

Tabella 2.3: Organizzazione del sistema scolastico portoghese

Nome	Tipologia	Descrizione	Valori
Debtor	Categorica	Studiante con debiti economici	0-No, 1-Yes
Scholarship Holder	Categorica	Studiante beneficiario di una borsa di studio	0-No, 1-Yes
Tuition Fees Up To Date	Categorica	Studiante con pagamenti delle tasse universitarie in regola	0-No, 1-Yes
Mother's / Father's Qualification	Categorica	Titolo di studio dei genitori dello studente	1-Secondary Education - 12th Year of Schooling or Eq., 2-Higher Education - Bachelor's Degree, 3-Higher Education - Degree, 4-Higher Education - Master's, 5-Higher Education - Doctorate, 6-Frequency of Higher Education, 9-12th Year of Schooling - Not Completed, 10-11th Year of Schooling - Not Completed, 11-7th Year (Old), 12-Other - 11th Year of Schooling, 13-2nd Year Complementary High School Course, 14-10th Year of Schooling, 18-General Commerce Course, 19-Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv., 20-Complementary High School Course, 22-Technical-Professional Course, 25-Complementary High School Course - Not Concluded, 26-7th Year of Schooling, 27-2nd Cycle of the General High School Course, 29-9th Year of Schooling - Not Completed, 30-8th Year of Schooling, 31-General Course of Administration and Commerce, 33-Supplementary Accounting and Administration, 34-Unknown, 35-Can't Read or Write, 36-Can Read Without Having a 4th Year of Schooling, 37-Basic Education 1st Cycle (4th/5th Year) or Equiv., 38-Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv., 39-Technological Specialization Course, 40-Higher Education - Degree (1st Cycle), 41-Specialized Higher Studies Course, 42-Professional Higher Technical Course, 43-Higher Education - Master (2nd Cycle), 44-Higher Education - Doctorate (3rd Cycle)
Mother's / Father's Occupation	Categorica	Occupazione Lavorativa dei genitori dello studente	0-Student, 1-Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers, 2-Specialists in Intellectual and Scientific Activities, 3-Intermediate Level Technicians and Professions, 4-Administrative staff, 5-Personal Services, Security and Safety Workers and Sellers, 6-Farmers and Skilled Workers in Agriculture, Fisheries and Forestry, 7-Skilled Workers in Industry, Construction and Craftsmen, 8-Installation and Machine Operators and Assembly Workers, 9-Unskilled Workers, 10-Armed Forces Professions, 90-Other Situation, 99-(blank), 101-Armed Forces Officers, 102-Armed Forces Sergeants, 103-Other Armed Forces personnel, 112-Directors of administrative and commercial services, 114-Hotel, catering, trade and other services directors, 121-Specialists in the physical sciences, mathematics, engineering and related techniques, 122-Health professionals, 123-Teachers, 124-Specialists in finance, accounting, administrative organization, public and commercial relations, 131-Intermediate level science and engineering technicians and professions, 132-Technicians and professionals, of intermediate level of health, 134-Intermediate level technicians from legal, social, sports, cultural and similar services, 135-Information and communication technology technicians, 141-Office workers, secretaries in general and data processing operators, 143-Data, accounting, statistical, financial services and registry-related operators, 144-Other administrative support staff, 151-Personal service workers, 152-Sellers, 153-Personal care workers and the like, 154-Protection and security services personnel, 161-Market-oriented farmers and skilled agricultural and animal production workers, 163-Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence, 171-Skilled construction workers and the like, except electricians, 172-Skilled workers in metallurgy, metalworking and similar, 174-Skilled workers in electricity and electronics, 175-Workers in food processing, woodworking, clothing and other industries and crafts, 181-Fixed plant and machine operators, 182-Assembly workers, 183-Vehicle drivers and mobile equipment operators, 192-Unskilled workers in agriculture, animal production, fisheries and forestry, 193-Unskilled workers in extractive industry, construction, manufacturing and transport, 194-Meal preparation assistants, 195-Street vendors (except food) and street service providers

Tabella 2.4: Fattori Socio-Economici: Descrizione Features

2.5 Dati sull'Iscrizione e Accesso ai Corsi

In questa sezione vengono analizzate le informazioni degli studenti note al momento dell'immatricolazione, considerando fattori come status di studente fuori sede, bisogni educativi speciali, modalità di accesso, ordine di preferenza e corso scelto. La Tabella 2.5 fornisce una panoramica dettagliata delle variabili analizzate. Dall'analisi di questi fattori emergono le seguenti osservazioni:

- **Displaced** (Figura 2.11): La distribuzione tra studenti fuori sede e in sede è relativamente equilibrata, con una leggera prevalenza dei primi (54.8% - 2424 studenti) rispetto ai secondi (45.2% - 2000 studenti). Questo sbilanciamento non estremamente marcato tra le due categorie suggerisce che l'istituto attira sia studenti residenti nella stessa area che provenienti da altre regioni.

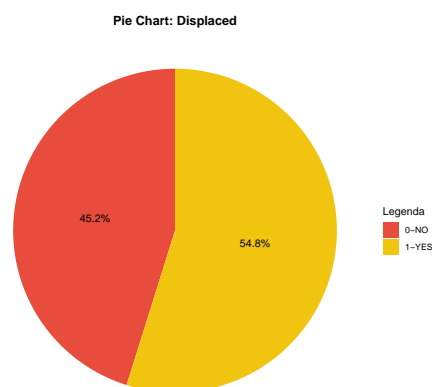


Figura 2.11: Analisi Univariata: Displaced

- **Educational Special Needs** (Figura 2.12): La quasi totalità degli studenti (98.8% - 4370 studenti) non presenta bisogni educativi speciali, mentre solo l'1.2% (53 studenti) rientra in questa categoria. La distribuzione è quindi fortemente sbilanciata, con una frequenza di studenti con bisogni educativi speciali molto bassa.

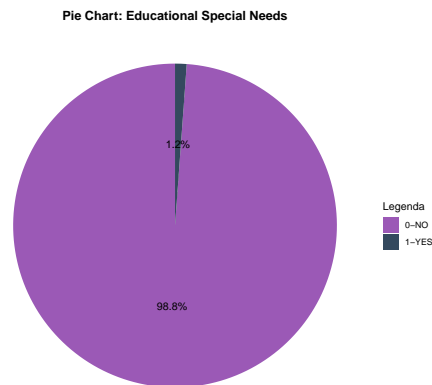


Figura 2.12: Analisi Univariata: Educational Special Needs

- **Daytime/Evening Attendance** (Figura 2.13): La maggior parte degli studenti (89.1% - 3942 studenti) frequenta corsi in modalità diurna, mentre il 10.9% (482 studenti) segue le lezioni in fascia serale. La distribuzione evidenzia uno squilibrio marcato tra le due categorie, con una netta predominanza della modalità "Daytime", suggerendo che l'istituto offra principalmente corsi con lezioni erogate durante il giorno. Tuttavia, il numero di studenti che seguono corsi serali può essere comunque considerato statisticamente significativo.

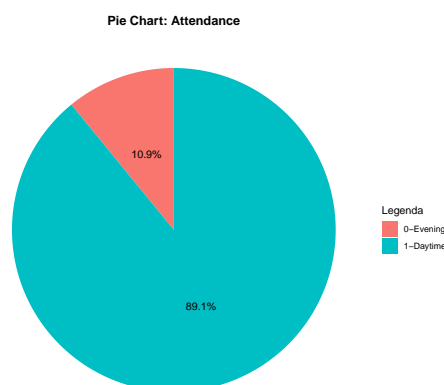


Figura 2.13: Analisi Univariata: Daytime/Evening Attendance

- **International** (Figura 2.14): La quasi totalità degli studenti (97.5% - 4313 studenti) non è iscritta ad un programma di mobilità internazionale (Erasmus), mentre solo il 2.5% (111 studenti) rientra nella categoria degli studenti Erasmus.

Un confronto con la feature "Nationality" analizzata nella Sezione 2.3, mostra che tutti gli studenti di nazionalità non Portoghese corrispondono esattamente agli studenti Erasmus.

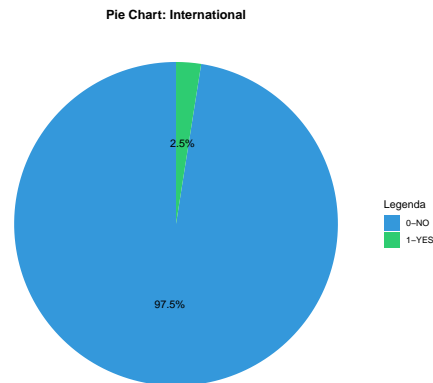


Figura 2.14: Analisi Univariata: International

- **Application Mode** (Figura 2.15): Possiamo notare una distribuzione fortemente sbilanciata, con alcune categorie nettamente più frequenti rispetto ad altre. La modalità predominante è **(1) 1st phase - general contingent** con una frequenza del 38.6% (1708 studenti). È opportuno notare che anche modalità come **(39) Over 23 years old** e **(43) Change of course** sono presenti con una frequenza piuttosto rilevante, rispettivamente pari al 17.7% (783 studenti) e al 7.1%(314 studenti). Al contrario, molte altre modalità registrano una frequenza prossima allo 0%, suggerendo che si tratta di percorsi di accesso più rari.

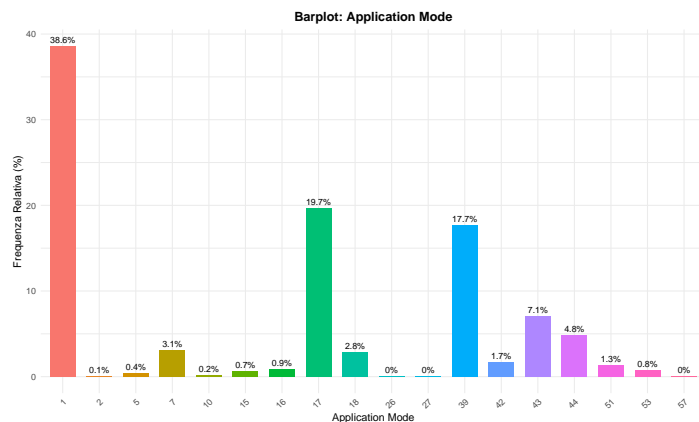


Figura 2.15: Analisi Univariata: Application Mode

- **Application Order** (Figura 2.16): Il grafico a barre mostra una netta prevalenza della modalità **(1) 1st Choiche** (68.4% - 3026 studenti), seguita a distanza dalle successive opzioni di scelta, con un progressivo calo delle frequenze all'aumentare dell'ordine di preferenza. Le categorie **0** e **9th** presentano rispettivamente 1 sola osservazione, il che suggerisce possibili errori o anomalie nei dati. La presenza della categoria **0** è particolarmente sospetta, poichè non sembra coerente con la logica dell'ordinamento delle scelte.

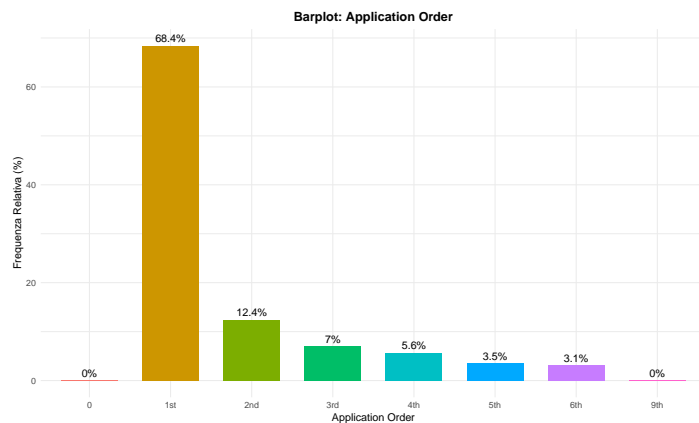


Figura 2.16: Analisi Univariata: Application Order

- **Course** (Figura 2.17): La distribuzione dei corsi a cui sono iscritti gli studenti appare relativamente bilanciata, ad eccezione di alcune categorie con valori più estremi. In particolare, il corso più frequentato è **9500-Nursing** (17.3% - 765 studenti), mentre altri corsi, come **33-Biofuel Production Technologies** (0.3% - 12 studenti) e **9556-Oral Hygiene** (1.9% - 84), registrano un numero molto inferiore di iscritti. È opportuno osservare che le modalità **33-Biofuel Production Technologies** e **171-Animation and Multimedia Design** sembrano riferirsi a corsi di diversa tipologia rispetto agli altri, come suggerisce la codifica numerica. Dalle informazioni ricavate dal sito dell'istituto è possibile che si tratti di **CTESP**, ovvero corsi tecnici superiori professionali. Inoltre, il corso **33-Biofuel Production Technologies** sembra non essere più attivo, e questo potrebbe spiegare lo scarso numero di osservazioni presenti nel dataset.

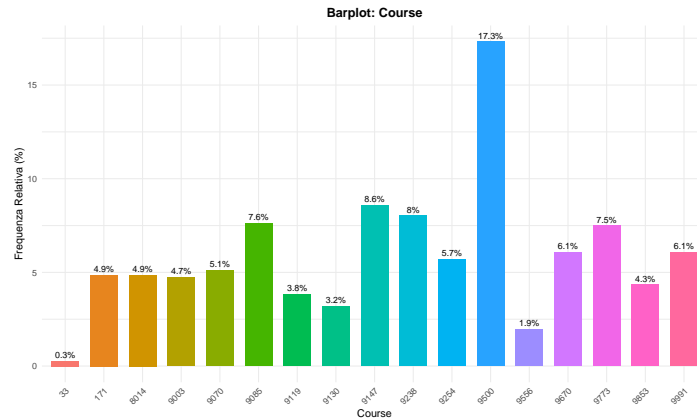


Figura 2.17: Analisi Univariata: Course

- Previous Qualification** (Figura 2.18): La distribuzione evidenzia che la modalità predominante è **1-Secondary Education** (84% - 3716 studenti), seguita da **39-Technological Specialization Course** (5% - 221 studenti). Ciò suggerisce che i corsi offerti dall'istituto, e rappresentati nel dataset, siano prevalentemente corsi di laurea che richiedono almeno il diploma di scuola secondaria (Secondary Education) per accedervi. Titoli di studio superiori, quali **3-Higher Education - degree** e **4-Higher Education - master**, risultano presenti con frequenze molto inferiori. È importante notare la presenza di numerose altre modalità con frequenze prossime allo 0%, indicative di casistiche più particolari.

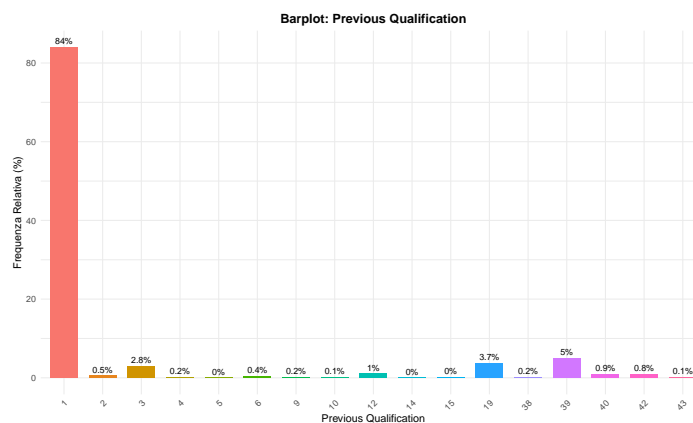


Figura 2.18: Analisi Univariata: Previous Qualification

- Previous Qualification Grade** (Figura 2.19): La distribuzione presenta una media pari a **132.61** ed una mediana pari a **133.1**, indicando che la distribuzione è quasi simmetrica, con solo una leggera asimmetria positiva (Skewness =

0.31). La Deviazione Standard (**13.19**) e l'IQR (**15**) mostrano che i valori sono abbastanza concentrati intorno alla media, senza variazioni eccessive. Il Coefficiente di Variazione (**9.94%**) conferma una relativa stabilità dei dati. Il boxplot mostra alcuni valori estremi sia nella parte inferiore che superiore, suggerendo la presenza di studenti che hanno ricevuto votazioni molto più basse o molto più alte rispetto alla media per la qualifica precedente. Un aspetto interessante riguarda la presenza di un picco massimo di densità intorno al 140 e diversi picchi più bassi in corrispondenza di valori come 110, 120, 150, ecc.. Questo pattern suggerisce che i voti potrebbero essere stati soggetti a operazioni di arrotondamento.

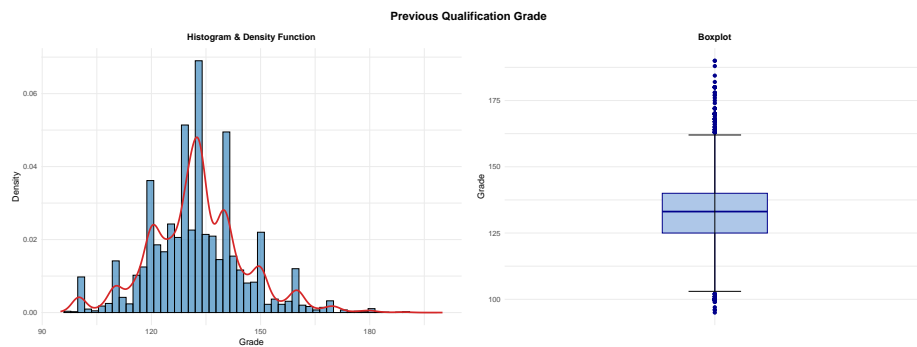


Figura 2.19: Analisi Univariata: Previous Qualification Grade

- **Admission Grade** (Figura 2.20): La distribuzione presenta una media (**126.98**) ed una mediana (**126.1**) molto vicine, indicando anche in questo caso una distribuzione relativamente simmetrica, sebbene sia presente una leggera asimmetria positiva (Skewness = **0.53**). La deviazione standard (**14.48**), l'IQR (**16.9**) e il Coefficiente di Variazione (**11.41%**) mostrano una concentrazione abbastanza stabile dei valori intorno alla media. Il boxplot evidenzia la presenza di alcuni studenti con punteggi di ammissione significativamente superiori alla norma (outlier). Inoltre, possiamo notare la presenza di due picchi massimi di densità (circa a 120 e 130) e diversi picchi più bassi in corrispondenza di valori ricorrenti (100, 140, 150, ecc..), che anche in questo caso potrebbero indicare operazioni di arrotondamento nei dati di ammissione.

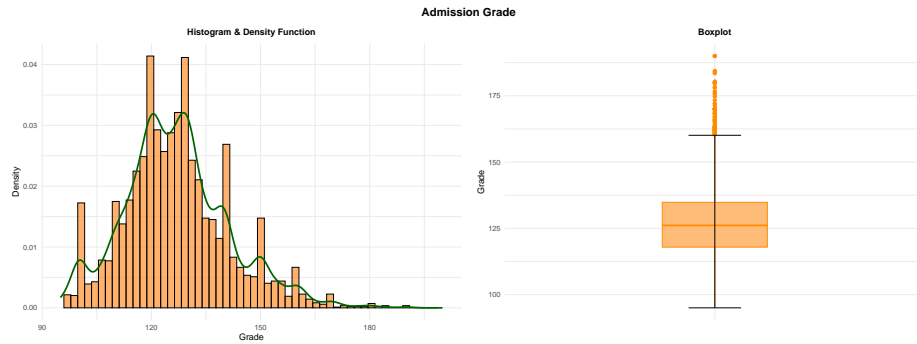


Figura 2.20: Analisi Univariata: Admission Grade

Nome	Tipologia	Descrizione	Valori
Displaced	Categorica	Studente fuori sede	0-No, 1-Yes
Daytime/Evening Attendance	Categorica	Frequenza alle lezioni dello studente	0-Evening, 1-Daytime
Educational Special Needs	Categorica	Studente con bisogni educativi speciali	0-No, 1-Yes
International	Categorica	Studente Erasmus	0-No, 1-Yes
Application Mode	Categorica	Modalità di Immatricolazione	1-1st phase - general contingent, 2-Ordinance No. 612/93, 5-1st phase - special contingent (Azores Island), 7-Holders of other higher courses, 10-Ordinance No. 854-B/99, 15-International student (bachelor), 16-1st phase - special contingent (Madeira Island), 17-2nd phase - general contingent, 18-3rd phase - general contingent, 26-Ordinance No. 533-A/99, item b2) (Different Plan), 27-Ordinance No. 533-A/99, item b3 (Other Institution), 39-Over 23 years old, 42-Transfer, 43-Change of course, 44-Technological specialization diploma holders, 51-Change of institution/course, 53-Short cycle diploma holders, 57-Change of institution/course (International)
Application Order	Categorica	Ordine di Immatricolazione (Preferenza)	1-1st Choice, 2-2nd Choice, 3-3rd Choice, 4-4th Choice, 5-5th Choice
Previous Qualification	Categorica	Qualifica precedente dello studente	1-Secondary education, 2-Higher education - bachelor's degree, 3-Higher education - degree, 4-Higher education - master's, 5-Higher education - doctorate, 6-Frequency of higher education, 9-12th year of schooling - not completed, 10-11th year of schooling - not completed, 12-Other - 11th year of schooling, 14-10th year of schooling, 15-10th year of schooling - not completed, 19-Basic education 3rd cycle (9th/10th/11th year) or equiv., 38-Basic education 2nd cycle (6th/7th/8th year) or equiv., 39-Technological specialization course, 40-Higher education - degree (1st cycle), 42-Professional higher technical course, 43-Higher education - master (2nd cycle).
Course	Categorica	Corso a cui lo studente si è immatricolato	33-Biofuel Production Technologies, 171-Animation and Multimedia Design, 8014-Social Service (evening attendance), 9003-Agronomy, 9070-Communication Design, 9085-Veterinary Nursing, 9119-Informatics Engineering, 9130-Equiculture, 9147-Management, 9238-Social Service, 9254-Tourism, 9500-Nursing, 9556-Oral Hygiene, 9670-Advertising and Marketing Management, 9773-Journalism and Communication, 9853-Basic Education, 9991-Management (evening attendance).
Previous Qualification Grade	Quantitativa Continua	Voto finale della qualifica precedente	[95 – 200]
Admission Grade	Quantitativa Continua	Voto Test di Ammissione	[95 – 200]

Tabella 2.5: Dati sull'Iscrizione e Accesso ai Corsi: Descrizione Features

2.6 Informazioni Primo Anno Accademico

In questa sezione vengono analizzate le informazioni degli studenti raccolte durante il primo anno di studi, considerando variabili come unità curriculari frequentate, valutazioni ricevute, esami superati e votazioni medie. Tutte queste informazioni sono suddivise in Primo Semestre e Secondo Semestre. La Tabella 2.6 fornisce una panoramica dettagliata delle variabili analizzate. Dall'analisi di questi fattori emergono le seguenti osservazioni:

- **Curricular Units Credited** (Figura 2.21): La distribuzione mostra una forte asimmetria positiva in entrambi i semestri, con la maggior parte degli studenti che non ha ricevuto alcun riconoscimento di Unità Curriculari. In particolare, la mediana è pari a **0** in entrambi i semestri, mentre la media è **0.71** nel primo semestre e **0.54** nel secondo, indicando che solo una minoranza degli studenti ha delle unità curriculari che gli sono state riconosciute. Questa tendenza è confermata dall'**IQR=0** in entrambi i semestri. La **Skewness** (**4.17** e **4.63**) e la **Curtosi** (**22.18** e **27.4**) indicano una distribuzione con code molto lunghe e la presenza di outlier. I grafici confermano questa evidenza, mostrando pochi casi di studenti che hanno ottenuto riconoscimenti, prevalentemente tra 1 e 5 CU. Sopra questa soglia, le frequenze calano ulteriormente. Questo potrebbe suggerire che i cambi di corso siano eventi più rari e che avvengono generalmente nelle fasi iniziali del percorso accademico.
- **Curricular Units Enrolled** (Figura 2.22): La distribuzione è fortemente concentrata intorno alle 6 CU, con una media di **6.27** nel primo semestre e **6.23** nel secondo semestre, e una mediana identica in entrambi i casi (**6 CU**). La Deviazione Standard è relativamente contenuta (**2.48** e **2.2**), indicando che la maggior parte degli studenti risulta iscritta ad un numero simile di CU. La Skewness positiva (**1.62** e **0.79**) e la Curtosi elevata (**11.93** e **10.13**) suggeriscono la presenza di una coda lunga verso destra, con alcuni studenti iscritti ad un numero significativamente maggiore di CU rispetto alla media. Il boxplot conferma la presenza di **outlier**, probabilmente dovuti a corsi con una diversa organizzazione accademica, nei quali gli studenti si iscrivono a un numero di

CU più elevato o più basso rispetto alla norma. Un elemento interessante è la presenza di una percentuale non trascurabile di studenti (circa il 5%) iscritti a **0 CU**, come si nota negli istogrammi. Questo potrebbe essere dovuto a dati mancanti o ad altri casi particolari.

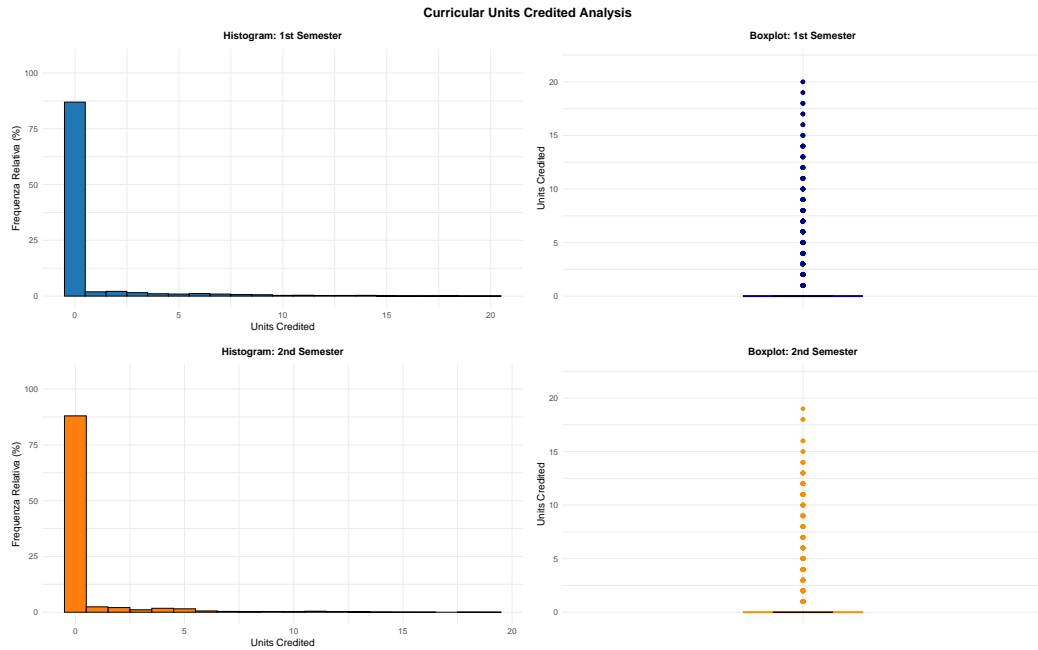


Figura 2.21: Analisi Univariata: Curricular Units Credited

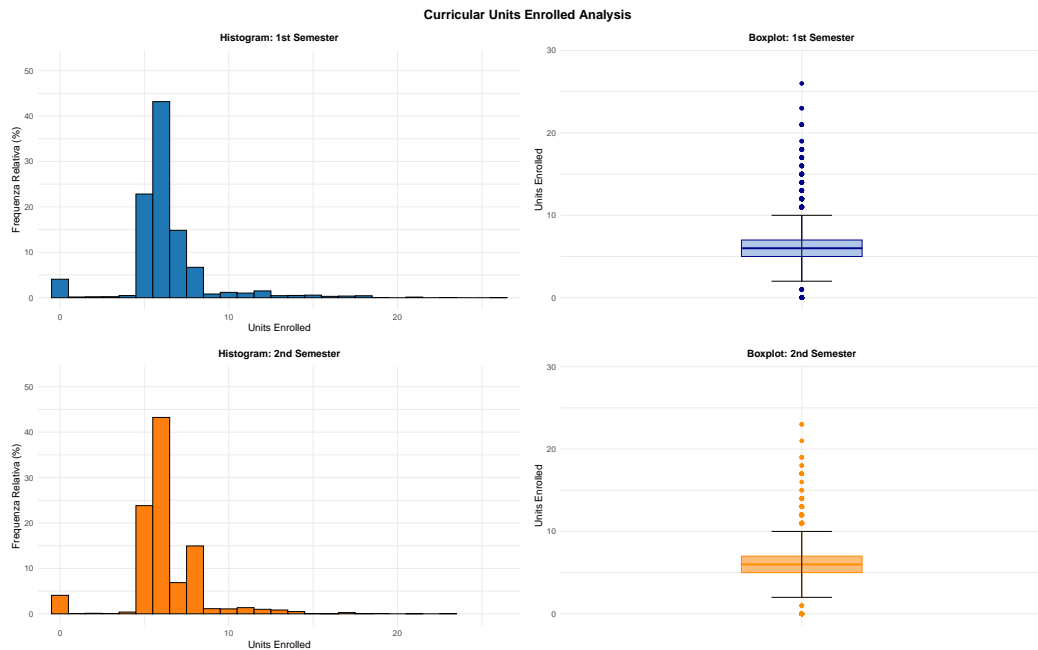


Figura 2.22: Analisi Univariata: Curricular Units Enrolled

- Curricular Units Evaluations** (Figura 2.23): La distribuzione mostra tendenze in linea con quanto osservato per i Curricular Units Enrolled. La media (8.3 nel primo semestre, 8.06 nel secondo) e la mediana (8 per entrambi i semestri) indicano che la maggior parte degli studenti riceve un numero simile di valutazioni, in linea con il numero di CU a cui sono iscritti. Il Coefficiente di Variazione (50.36% e 48.96%) evidenzia una significativa dispersione dei dati, mentre la Skewness (0.98 e 0.34) suggerisce la presenza di alcuni studenti con un numero di valutazioni superiore alla media. Un aspetto rilevante è la presenza di una percentuale significativa di studenti con **0 valutazioni**, più alta nel secondo semestre. Questo potrebbe essere dovuto a dati mancanti, a studenti che si sono iscritti ma non hanno sostenuto esami, oppure a ritiri o interruzioni del percorso di studi.

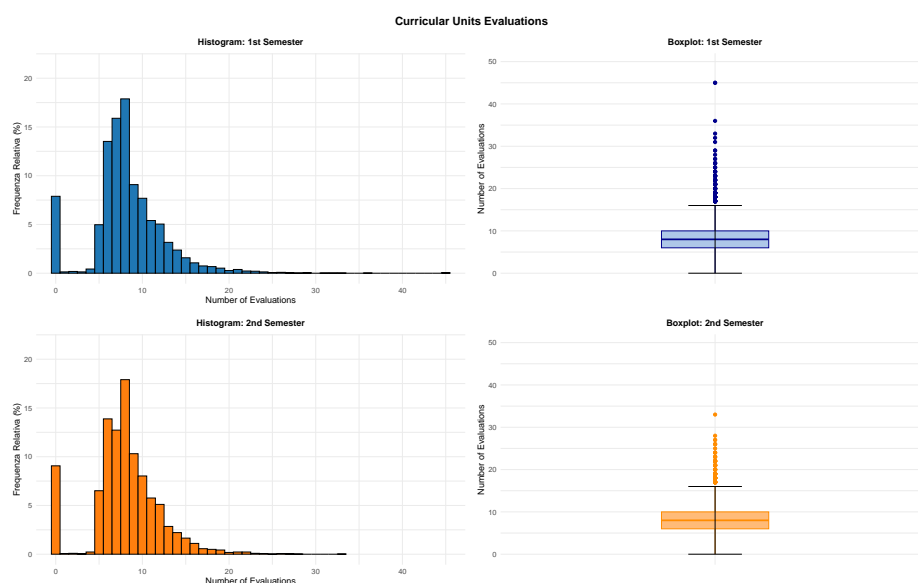


Figura 2.23: Analisi Univariata: Curricular Units Evaluations

- Curricular Units Approved** (Figura 2.24): La distribuzione mostra tendenze in linea con quelle osservate per Curricular Units Enrolled ed Evaluations con una media di 4.71 nel primo semestre e di 4.44 nel secondo, e una mediana pari a 5 in entrambi i casi. Il Coefficiente di Variazione (65.74% e 67.96%) indica una dispersione relativamente alta. La Skewness di 0.77 e 0.31 suggerisce la presenza di una leggera asimmetria positiva. Un aspetto rilevante è la presenza di una frequenza non trascurabile di studenti con **1-4 CU approvati**, suggerendo

la presenza di studenti che riescono a superare solamente poche Unità Curricolari. Inoltre, la percentuale di studenti con 0 CU Approvati è considerevole (soprattutto nel secondo semestre), segnalando anche in questo caso possibili dati mancanti, difficoltà accademiche o ritiri.

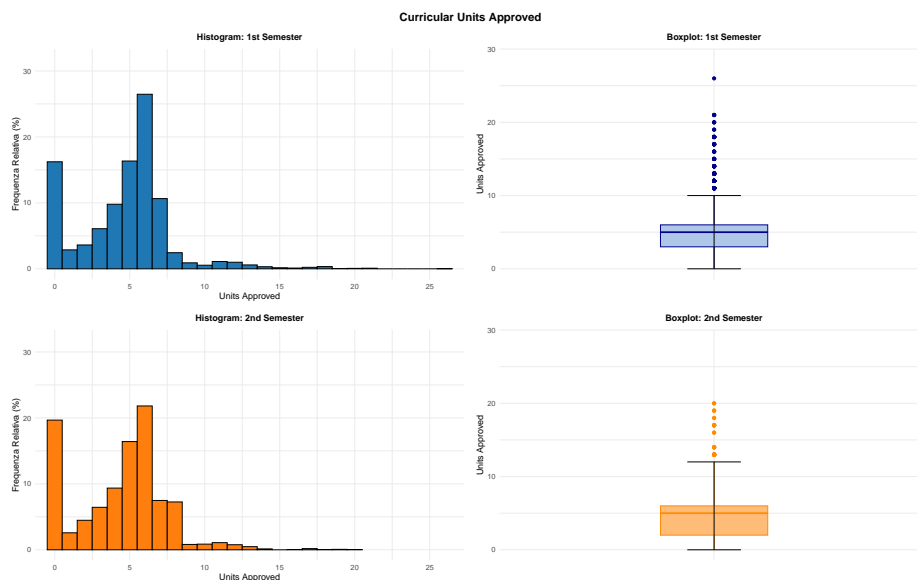


Figura 2.24: Analisi Univariata: Curricular Units Approved

- Curricular Units Grade:** La distribuzione presenta una media (10.64 nel primo semestre, 10.23 nel secondo semestre) inferiore rispetto alla mediana (12.29 e 12.2 rispettivamente), suggerendo una asimmetria negativa. Tuttavia, la media è influenzata dalla presenza di una concentrazione di outlier con un valore pari a 0, i quali indicano probabilmente studenti che non hanno superato alcun esame o dati mancanti. Gli istogrammi con le relative funzioni di densità mostrano che, considerando solo gli studenti con una votazione media diversa da 0, la distribuzione delle medie tende ad assumere una forma simile ad una curva normale, con un picco centrale (vicino al valore 12.5) ed una dispersione quasi perfettamente simmetrica. È interessante osservare che gli studenti sono più concentrati vicino alla soglia minima di superamento (10), mentre l'assenza di osservazioni vicine al massimo (20) potrebbe indicare che ottenere votazioni molto alte è un evento poco frequente.

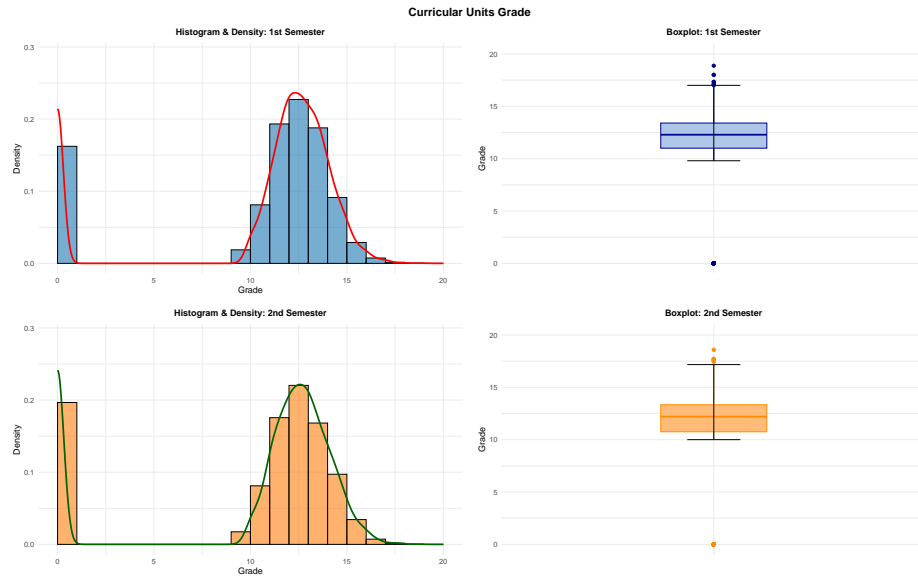


Figura 2.25: Analisi Univariata: Curricular Units Grade

- Curricular Units Without Evaluation** (Figura 2.26): La distribuzione mostra che la quasi totalità degli studenti non ha unità curriculari senza valutazione durante il primo anno accademico. Osservazione confermata anche dalla mediana pari a 0. La media (0.14 e 0.15) e la Skewness (8.2 e 7.27) confermano anche essi una concentrazione su 0 CU, e pochissimi casi di valori più alti. Dai grafici è possibile notare la presenza di outlier, ossia studenti che hanno diverse CU senza valutazione (fino a 20 CU).

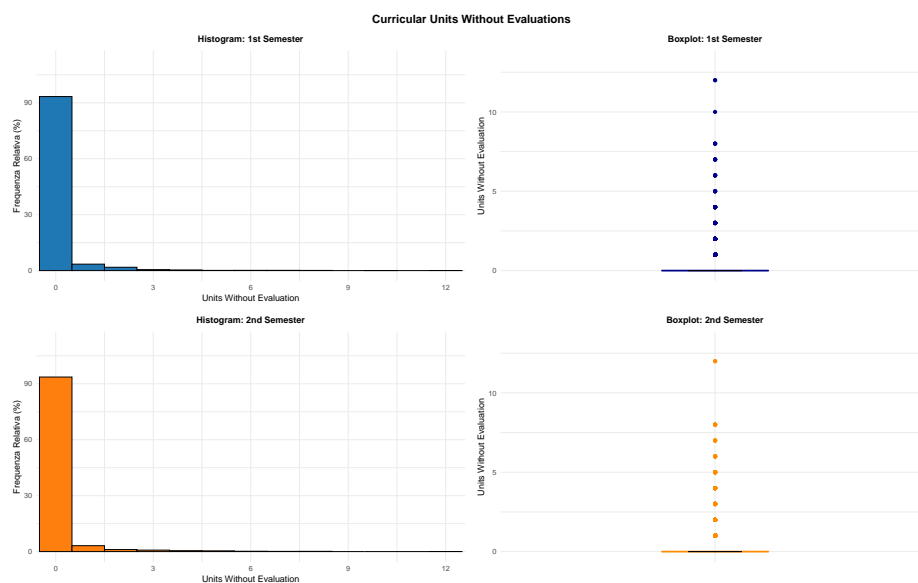


Figura 2.26: Analisi Univariata: Curricular Units Without Evaluations

Nome	Tipologia	Descrizione	Valori
Curricular Units Credited	Quantitativa Discreta	Unità Curricolari Convalidate/riconosciute nel primo e nel secondo semestre	$x \in \mathbb{N}$
Curricular Units Enrolled	Quantitativa Discreta	Unità Curricolari a cui lo studente risulta iscritto nel primo e nel secondo semestre	$x \in \mathbb{N}$
Curricular Units Evaluations	Quantitativa Discreta	Numero di valutazioni ricevute (tentativi effettuati) per le unità curricolari del primo e del secondo semestre	$x \in \mathbb{N}$
Curricular Units Approved	Quantitativa Discreta	Unità Curricolari Completate (esame superato) nel primo e nel secondo semestre	$x \in \mathbb{N}$
Curricular Units Grade	Quantitativa Continua	Votazione media delle unità curricolari del primo e del secondo semestre	$\{0\} \cup [10, 20]$
Curricular Units Without Evaluation	Quantitativa Discreta	Unità Curricolari senza alcuna votazione nel primo e nel secondo semestre	$x \in \mathbb{N}$

Tabella 2.6: Informazioni Primo Anno Accademico: Descrizione Features

2.7 Fattori Macro-Economici

In questa sezione vengono analizzati i fattori macro-economici, ovvero i fattori economici relativi al Portogallo nell'anno di immatricolazione degli studenti, includendo variabili come variazione del PIL, tasso di inflazione e tasso di disoccupazione. La Tabella 2.7 fornisce una panoramica dettagliata delle variabili analizzate. Innanzitutto è opportuno osservare che a differenza di altre variabili continue, le features macro-economiche, in questo caso assumo un insieme discreto di soli 10 valori distinti, riflettendo caratteristiche macro-economiche del Portogallo nel periodo in cui sono stati raccolti i dati (2008/2009 - 2018/2019). Dall'analisi di questi fattori emergono le seguenti osservazioni:

- **GDP** (Figura 2.27): La media è **0%**, mentre la mediana e la moda sono entrambe **0.32%**, suggerendo quindi un maggiore concentrazione di iscritti in anni in cui il PIL era stabile o in leggera crescita. La Deviazione Standard del (**2.27%**) indica una variabilità moderata delle variazioni del PIL nei diversi anni accademici analizzati. La Skewness (**-0.39**) e la Curtosi (**2**) indicano una leggera asimmetria negativa e una distribuzione senza picchi troppo estremi.

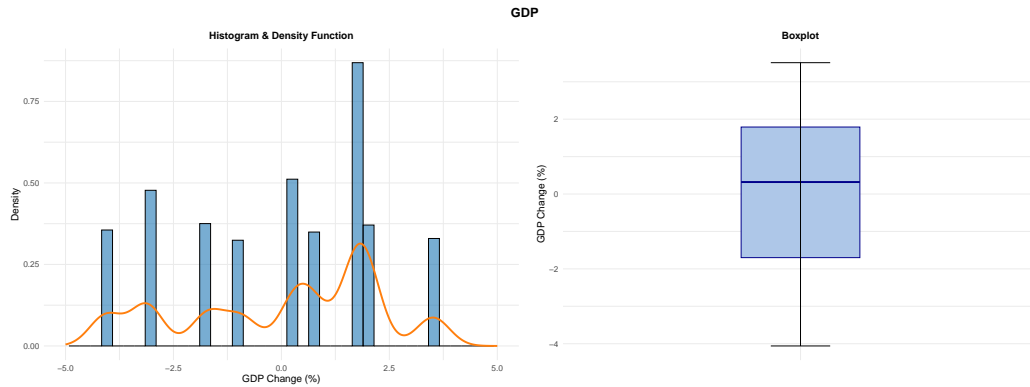


Figura 2.27: Analisi Univariata: GDP

- Inflation Rate** (Figura 2.28): La media è **1.23%**, mentre la mediana e la moda coincidono a **1.4%**, indicando una distribuzione leggermente concentrata attorno a questi valori. La Deviazione Standard di **1.38%** evidenzia una discreta variabilità tra i diversi anni considerati. La Skewness (**0.25**) suggerisce una leggera asimmetria positiva, mentre la Curtosi (**1.96**) indica una distribuzione piuttosto piatta, senza picchi estremi. Infatti, anche dall'istogramma è possibile osservare la presenza di diversi picchi di densità, tutti relativamente equilibrati tra loro.

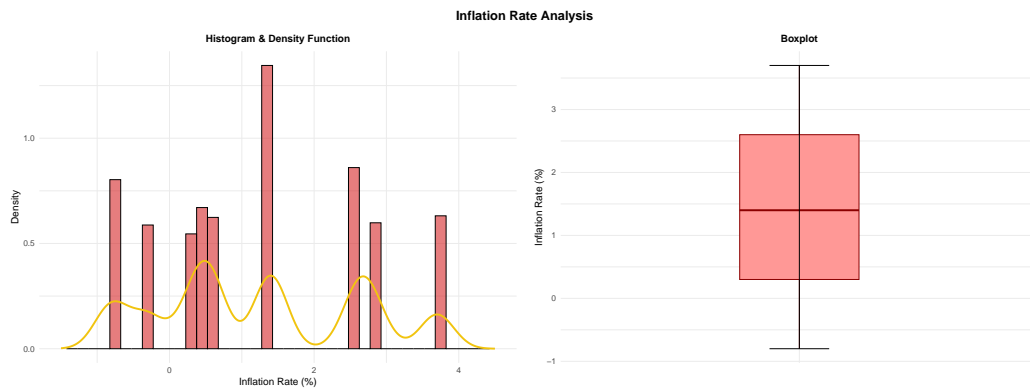


Figura 2.28: Analisi Univariata: Inflation Rate

- Unemployment Rate** (Figura 2.29): La media è **11.57%**, mentre la mediana è leggermente inferiore, pari a **11.1%**, indicando una distribuzione abbastanza equilibrata. La moda (**7.6%**) evidenzia che il valore più frequente corrisponde al tasso di disoccupazione più basso del periodo considerato. La deviazione standard è **2.66%**, indicando una variabilità moderata tra gli anni. La Skewness

(0.21) suggerisce una leggera asimmetria positiva. Anche in questo caso, la Curtosi (2) indica una distribuzione senza picchi estremi (Come si può notare anche dal grafico, in cui sono presenti diversi picchi di densità, ma tutti relativamente equilibrati tra loro).

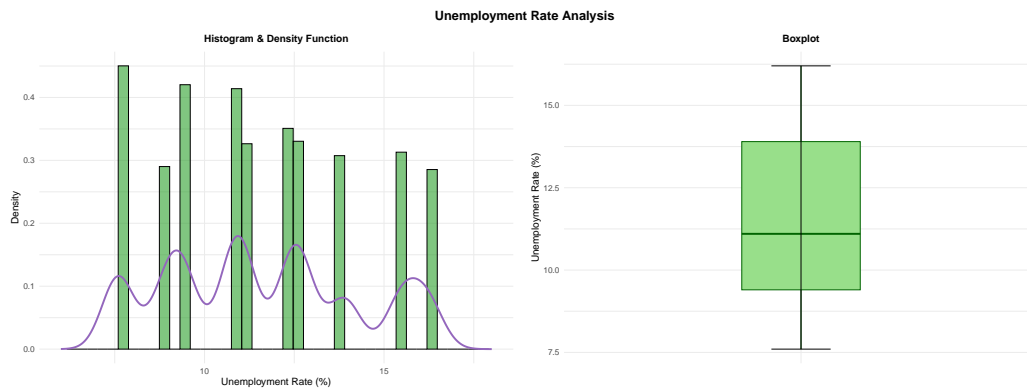


Figura 2.29: Analisi Univariata: Unemployment Rate

Nome	Tipologia	Descrizione	Valori
GDP	Quantitativa Continua	Variazione percentuale del PIL del Portogallo nell'anno di immatricolazione	$x \in \mathbb{R}$
Inflation Rate	Quantitativa Continua	Tasso di inflazione del Portogallo nell'anno di immatricolazione	$x \in \mathbb{R}$
Unemployment Rate	Quantitativa Continua	Tasso di disoccupazione del Portogallo nell'anno di immatricolazione	$x \in \mathbb{R}$

Tabella 2.7: Fattori Macro-Economici: Descrizione Features

2.8 Insight Analisi Univariata

L'analisi univariata ha evidenziato diversi aspetti chiave relativi alla distribuzione delle variabili nel dataset:

- **Distribuzione sbilanciata delle variabili categoriche:** Molte variabili categoriche presentano un numero elevato di categorie, ma la distribuzione delle frequenze risulta spesso fortemente sbilanciata. In diversi casi, una singola categoria domina nettamente sulle altre, mentre alcune categorie hanno una frequenza molto bassa, rendendole statisticamente poco rilevanti. Potrebbe

essere utile aggregare alcune modalità meno rappresentate per rendere più efficace le analisi successive.

- **Codifica numerica delle variabili categoriche poco chiara:** Nella maggior parte dei casi, non è esplicitamente chiaro quale criterio sia stato utilizzato dagli autori del dataset per trasformare le variabili categoriche non binarie in formato numerico. Sebbene questa codifica permetta di trattare le variabili in modo più agevole nei modelli quantitativi, non è detto che i valori numerici assegnati abbiano un significato ordinato o semantico. Questo aspetto potrebbe richiedere una rivalutazione della codifica prima di utilizzarle in analisi più avanzate, per evitare interpretazioni fuorvianti o problemi nei modelli statistici. L'unico caso in cui la codifica numerica sembra avere un criterio logico chiaro riguarda la variabile "Course", dove i valori corrispondono ai codici ufficiali dei corsi di studio dell'istituto.
- **Presenza diffusa di outlier nelle variabili numeriche:** La maggior parte delle variabili numeriche mostra valori estremi che si discostano significativamente dalla distribuzione centrale. La presenza di outlier suggerisce la necessità di un'analisi più approfondita per determinare se tali valori siano errori, dati anomali ma validi o indicatori di pattern specifici.
- **Presenza di valori sospetti nelle informazioni del primo anno accademico:** Molte delle variabili relative al primo anno di studi contengono un numero elevato di osservazioni con valore pari a 0. Questo potrebbe essere dovuto anche ad errori o dati mancanti codificati come zero. Sarà necessaria un'analisi più fine per distinguere i dati mancanti dalle osservazioni reali, evitando interpretazioni errate nelle analisi successive.
- **Bassa variabilità nei fattori macro-economici:** Le variabili relative a PIL, inflazione e tasso di disoccupazione assumono un numero limitato di valori, poiché si riferiscono al contesto nazionale (Portogallo) e i dati sono stati raccolti in un intervallo di soli 10 anni. Questo ridotto livello di variabilità potrebbe limitarne l'impatto nelle analisi successive, poiché i dati macroeconomici rimangono relativamente stabili rispetto ai singoli percorsi accademici degli studenti.

CAPITOLO 3

Analisi Bivariata

In questo capitolo viene condotta l'analisi bivariata con l'obiettivo di studiare la relazione tra ciascuna feature del dataset e la variabile Target. Questo permette di individuare eventuali associazioni tra le caratteristiche degli studenti e il loro esito accademico, nonché possibili fattori di rischio legati al successo o all'abbandono degli studi. A seconda della tipologia di variabile, sono stati utilizzati strumenti adeguati per valutare e rappresentare efficacemente tali relazioni.

Per le **Variabili Categoriche** sono stati utilizzati:

- **Tabelle di Contingenza:** utilizzate per confrontare la distribuzione delle categorie della variabile rispetto ai tre gruppi della variabile target.
- **Grouped BarPlot:** impiegati per rappresentare graficamente la distribuzione delle categorie rispetto ai tre esiti accademici.

Per le **Variabili Numeriche** sono stati utilizzati:

- **Boxplot:** utilizzati per evidenziare differenze nella distribuzione della variabile numerica tra i tre gruppi della variabile Target, permettendo di osservare variazioni nella tendenza centrale e nella dispersione.

- **Istogrammi:** adottati per analizzare e confrontare la distribuzione delle variabili numeriche discrete nei tre gruppi della variabile Target.
- **Kernel Density Plot:** adottati come alternativa agli istogrammi per le variabili numeriche continue.
- **ViolinPlot:** utilizzati in alcuni casi per mostrare le stesse informazioni dei boxplot e dei kernel density plot ma in maniera più compatta.

Nei capitoli successivi verranno inizialmente illustrate le operazioni di analisi delle correlazioni tra features numeriche e le operazioni di aggregazione delle feature categoriche, effettuate per ottimizzare l'analisi bivariata e migliorare l'interpretazione dei dati. Successivamente, verranno presentate le osservazioni relative alla relazione tra le diverse features e la variabile target, evidenziando in particolare le feature che mostrano associazioni più forti e quelle che, invece, non presentano differenze rilevanti tra i diversi gruppi.

3.1 Correlazione tra Feature Numeriche

Prima di procedere con l'analisi bivariata tra le variabili del dataset e la variabile target, è stata calcolata la matrice di correlazione per verificare la presenza di relazioni significative tra le feature numeriche. Sebbene questa analisi non fornisca informazioni direttamente utili per rispondere alla RQ 1 (poiché la variabile target non è numerica), permette di individuare pattern rilevanti tra le altre variabili. Questi pattern possono essere utili per comprendere la struttura del dataset e per supportare analisi successive. Dall'analisi della matrice mostrata in Figura 3.1 emergono alcune osservazioni interessanti:

- **Correlazione tra Previous Qualification Grade e Admission Grade:** Si osserva una correlazione positiva moderatamente forte ($r = 0.58$), suggerendo che gli studenti con voti più alti per la qualifica precedente tendono a ottenere punteggi più elevati anche nei test di ammissione.
- **Correlazione tra le Medie dei Voti nei due Semestri:** Esiste una correlazione molto forte ($r = 0.84$) tra la media dei voti del primo e del secondo semestre.

Questo implica che gli studenti con buoni risultati nel primo semestre tendono a mantenere un buon rendimento anche nel secondo.

- **Correlazione tra le Variabili Accademiche dei Due Semestri:** Le feature relative ai dati degli studenti nei due semestri del primo anno accademico mostrano correlazioni positive di grado moderato o forte. Questo risultato è plausibile, considerando che tipicamente i due semestri accademici tendono a essere organizzati in modo simile, offrendo un numero comparabile di unità curriculari. Di conseguenza, gli studenti generalmente mantengono un carico di studio coerente tra i due periodi, iscrivendosi a un numero simile di corsi e sostenendo/superando un numero analogo di esami.

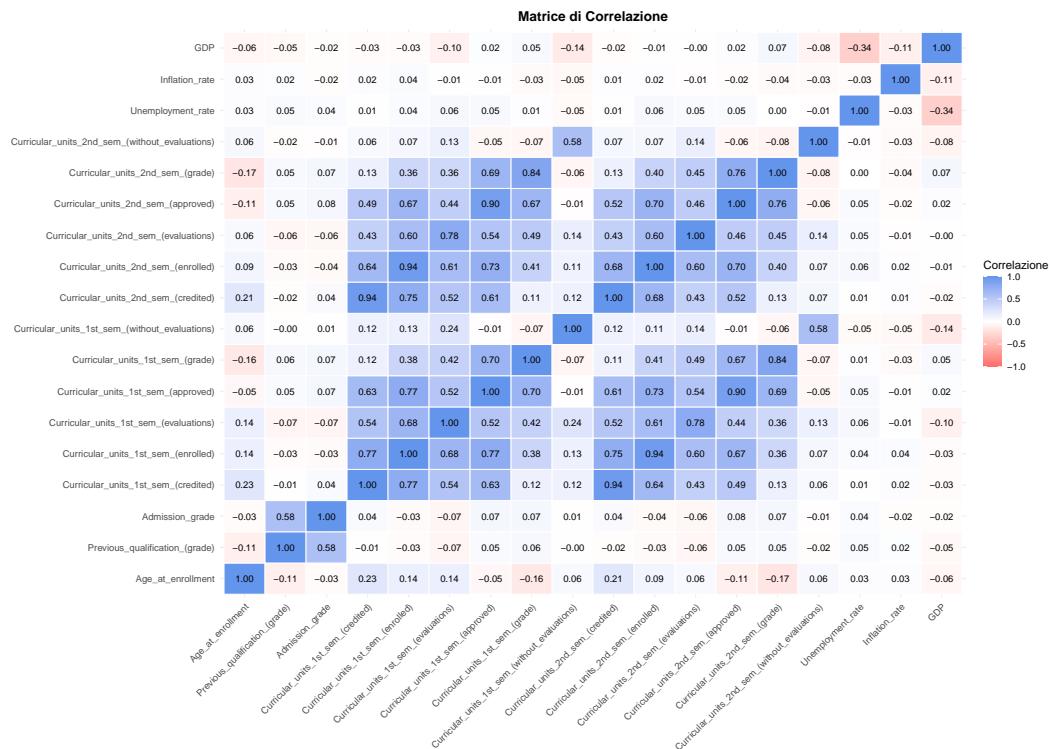


Figura 3.1: Matrice di Correlazione tra features quantitative

Dagli scatterplot mostrati in Figura 3.3 e Figura 3.2, possiamo notare alcuni pattern strutturali interessanti all'interno dei dati relativi alle votazioni degli studenti. In particolare, gli scatterplot presentano linee fitte di punti orizzontali e verticali in corrispondenza di valori interi, confermando maggiormente l'ipotesi fatta in precedenza riguardo possibili arrotondamenti nelle valutazioni degli studenti. Inoltre,

nello Scatterplot che mostra la relazione tra le medie dei voti del primo e del secondo semestre (Figura 3.3) si nota una chiara linea diagonale, la quale potrebbe suggerire che in precedenza siano già state utilizzate tecniche di imputazione dei dati attraverso modelli di regressione lineare.

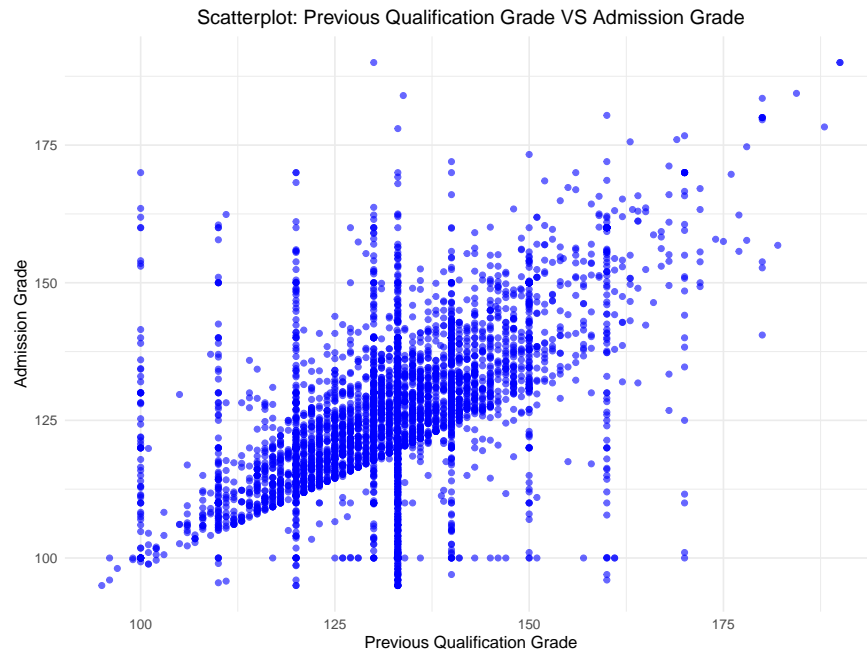


Figura 3.2: Scatterplot: Previous Qualification VS Admission Grade

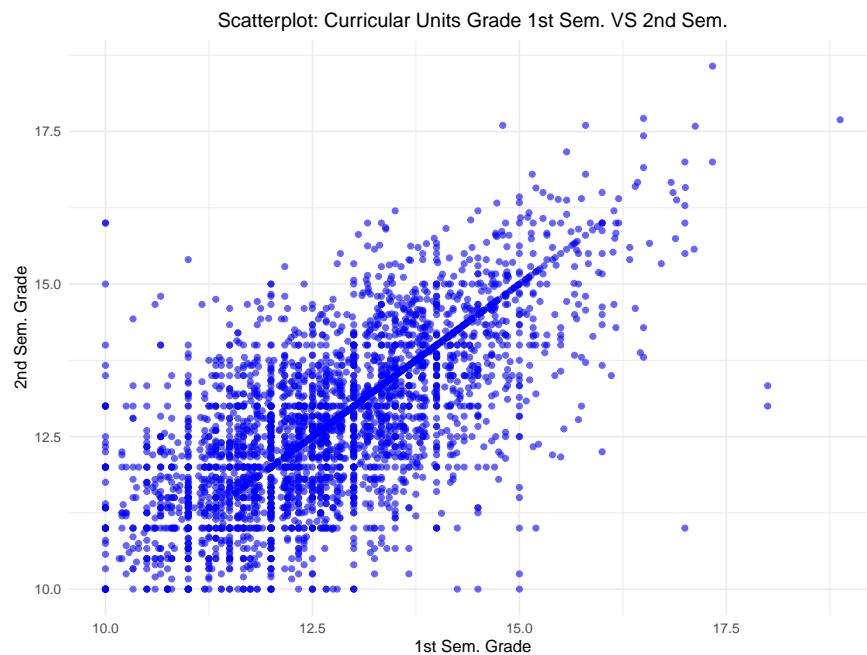


Figura 3.3: Scatterplot: CU Grade 1st sem. VS 2nd sem.

Sebbene queste osservazioni non rispondano direttamente alla nostra Research Question, esse forniscono informazioni cruciali sulla struttura del dataset e sulle relazioni interne tra le variabili. In particolare, la forte correlazione tra alcune feature potrebbe essere sfruttata per l'imputazione di dati mancanti, riducendo così la perdita di informazioni. Tuttavia, la presenza di numerose variabili fortemente correlate tra loro, specialmente tra le feature accademiche relative ai due semestri, evidenzia un problema di multicollinearità. Questo fenomeno può influenzare negativamente i modelli predittivi e le analisi inferenziali, portando, ad esempio, a stime instabili dei coefficienti nei modelli di regressione e riducendo la capacità di interpretare correttamente il contributo di ogni variabile. Per questo motivo, nelle fasi successive dell'analisi, potrebbe essere fondamentale valutare strategie di selezione delle feature o riduzione della dimensionalità, al fine di mitigare questi effetti e garantire una rappresentazione equilibrata delle informazioni.

3.2 Aggregazione Feature Categorie

L'analisi univariata ha evidenziato la presenza di variabili categoriche con diverse modalità, spesso caratterizzate da frequenze molto basse. Per gestire queste situazioni, è stato effettuato un processo di aggregazione mirato, con due obiettivi principali:

- **Migliorare l'interpretabilità Statistica:** La presenza di categorie con poche osservazioni può portare a interpretazioni errate nell'analisi bivariata. Ad esempio, se un'unica osservazione appartiene a una certa modalità ed è associata a un esito accademico positivo (es. Graduate), si potrebbe erroneamente concludere che tutti gli individui con quella caratteristica ottengano lo stesso risultato. Questo problema è ancora più rilevante nei test statistici e nei modelli predittivi, dove basse frequenze possono causare risultati distorti o non generalizzabili.
- **Migliorare la Generazione del Dataset Sintetico con LLM:** L'aggregazione delle feature categoriche è stata effettuata anche per semplificare la generazione del dataset sintetico, riducendo la complessità del prompt e minimizzando il rumore dovuto alla presenza di troppe categorie poco rappresentative. Questo approccio permette di evitare ambiguità e incoerenze nei dati generati, oltre che

consentire un maggiore controllo sulle distribuzioni delle variabili, garantendo che il dataset sintetico rispecchi più fedelmente le proprietà statistiche del dataset originale.

L'aggregazione è stata condotta prestando attenzione a **preservare l'informazione rilevante** senza perdere dettagli significativi e **bilanciare la distribuzione delle frequenze** quanto meglio possibile, riducendo il numero di categorie con poche osservazioni. Nei paragrafi successivi verranno descritti i criteri specifici adottati per ciascuna variabile categorica.

Marital Status

La variabile *Marital Status* presentava 6 modalità, alcune delle quali con frequenze prossime allo 0%. Per garantire una maggiore robustezza nell'analisi bivariata, queste modalità sono state accorpate secondo la similarità concettuale, come riportato nella Tabella 3.1.

Nuova Categoria	Codici Originali	Descrizione
1 - Single	1 (Single)	Studenti non coniugati
2 - Married or Partner	2 (Married), 5 (Facto Union)	Coniugati o in unione di fatto
3 - Previously Married	3 (Widower), 4 (Divorced), 6 (Legally Separated)	Vedovi, divorziati o legalmente separati

Tabella 3.1: Aggregazione: "Marital Status"

Il risultato di questo accorpamento, illustrato nella Figura 3.4, ha ridotto il numero di modalità a tre, garantendo delle frequenze più significative, anche se la distribuzione resta fortemente sbilanciata.

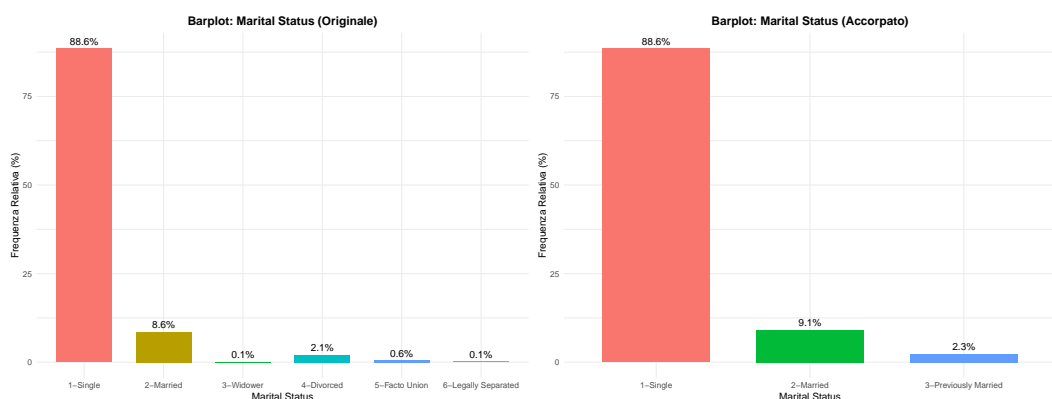


Figura 3.4: Effetto Aggregazione "Marital Status"

Nationality

La variabile *Nationality* comprendeva numerose modalità, ma una di esse risultava nettamente predominante: *Portuguese* (97.5%). Per questo motivo è stata convertita in una variabile categorica binaria, come mostrato nella Tabella 3.2.

Nuova Categoria	Codici Originali	Descrizione
1 - Portuguese	1 (Portuguese)	Studenti di nazionalità portoghese
2 - Other Nationality	2 (German), 6 (Spanish), 11 (Italian), 13 (Dutch), 14 (English), 17 (Lithuanian), 21 (Angolan), 22 (Cape Verdean), 24 (Guinean), 25 (Mozambican), 26 (Santomean), 32 (Turkish), 41 (Brazilian), 62 (Romanian), 100 (Moldovan), 101 (Mexican), 103 (Ukrainian), 105 (Russian), 108 (Cuban), 109 (Colombian)	Studenti di altre nazionalità

Tabella 3.2: Aggregazione: "Nationality"

L'aggregazione ha ridotto la variabile a due sole categorie, distinguendo tra studenti portoghesi e stranieri. Gli effetti della nuova distribuzione possono essere osservati nella Figura 3.5

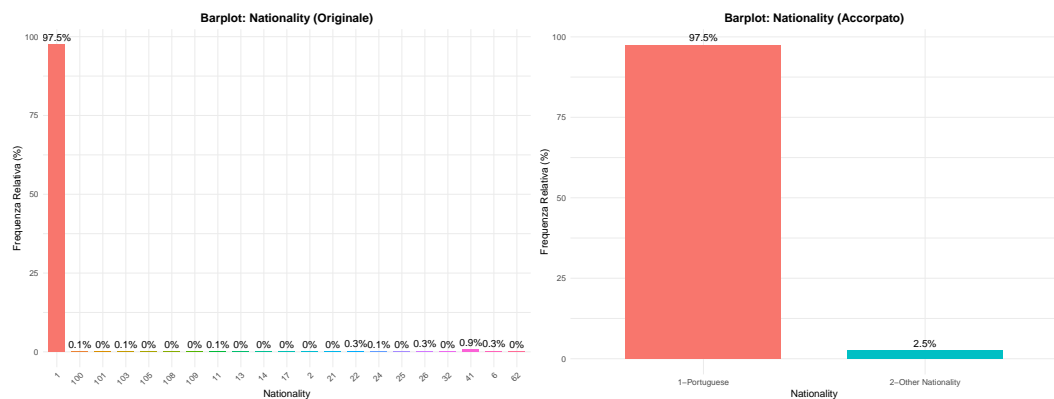


Figura 3.5: Effetto Aggregazione "Nationality"

Come già evidenziato in precedenza, gli studenti non portoghesi corrispondono esattamente agli studenti Erasmus. Di conseguenza, la nuova variabile *Nationality* è diventata equivalente alla feature *International*. Nelle analisi successive si terrà conto di questa ridondanza e, se necessario, si potrebbe eliminare una delle due features.

Mother's / Father's Qualification

Le variabili *Mother's Qualification* e *Father's Qualification* rappresentano il livello di istruzione dei genitori degli studenti nel campione. Molte delle modalità presenti descrivevano percorsi di studio interrotti o incompleti con frequenze molto basse. Per migliorare l'analisi, è stato effettuato un accorpamento basato sul funzionamento del sistema scolastico portoghese, considerando esclusivamente la qualifica effettivamente raggiunta. La Tabella 3.3 riassume la nuova categorizzazione adottata.

Nuova Categoria	Codici Originali	Descrizione
1 - Higher Education	2 (Bachelor's), 3 (Degree), 4 (Master's), 5 (Doctorate), 40 (Degree - 1st Cycle), 41 (Specialized Studies), 42 (Technical Course), 43 (Master - 2nd Cycle), 44 (Doctorate - 3rd Cycle)	Genitori con istruzione universitaria
2 - Secondary Education	1 (Secondary Education), 6 (Frequency of Higher Education), 18 (General Commerce Course), 22 (Technical-Professional Course), 27 (2nd Cycle General High School), 39 (Technological Specialization Course), 13, 20, 25, 31, 33 (Complementary High School Courses)	Genitori con diploma di scuola superiore o titolo equivalente
3 - Basic Education (3° ciclo)	9 (12th Year Not Completed), 10 (11th Year Not Completed), 12 (Other - 11th Year), 14 (10th Year), 19 (Basic Education 3rd Cycle)	Genitori con istruzione media inferiore
4 - Basic Education (2° ciclo)	11 (7th Year Old), 26 (7th Year), 29 (9th Year Not Completed), 30 (8th Year), 38 (Basic Education 2nd Cycle)	Genitori con istruzione elementare superiore
5 - Basic Education (1° ciclo)	37 (Basic Education 1st Cycle - 4th/5th Year)	Genitori con istruzione elementare inferiore
6 - No Qualification	34 (Unknown), 35 (Can't Read or Write), 36 (Can Read Without 4th Year)	Genitori senza istruzione formale

Tabella 3.3: Aggregazione "Mother's / Father's Qualification"

Come illustrato nella Figura 3.6, questa aggregazione ha ridotto sensibilmente il numero di categorie, migliorando il bilanciamento delle distribuzioni e facilitando l'interpretazione dei dati.

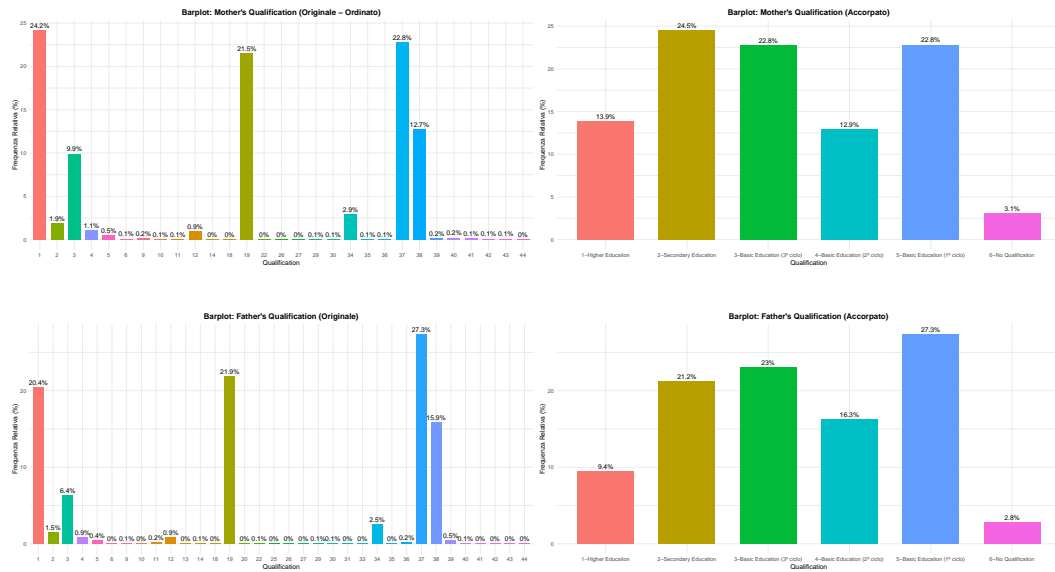


Figura 3.6: Effetto Aggregazione "Mother's / Father's Qualification"

Mother's / Father's Occupation

Le variabili *Mother's Occupation* e *Father's Occupation* descrivono la professione dei genitori degli studenti. Tuttavia, molte categorie rappresentavano occupazioni molto rare con una frequenza prossima allo 0%. Per migliorare l'interpretabilità dei dati, le professioni sono state accorpate in macro-categorie professionali, come mostrato nella Tabella 3.4.

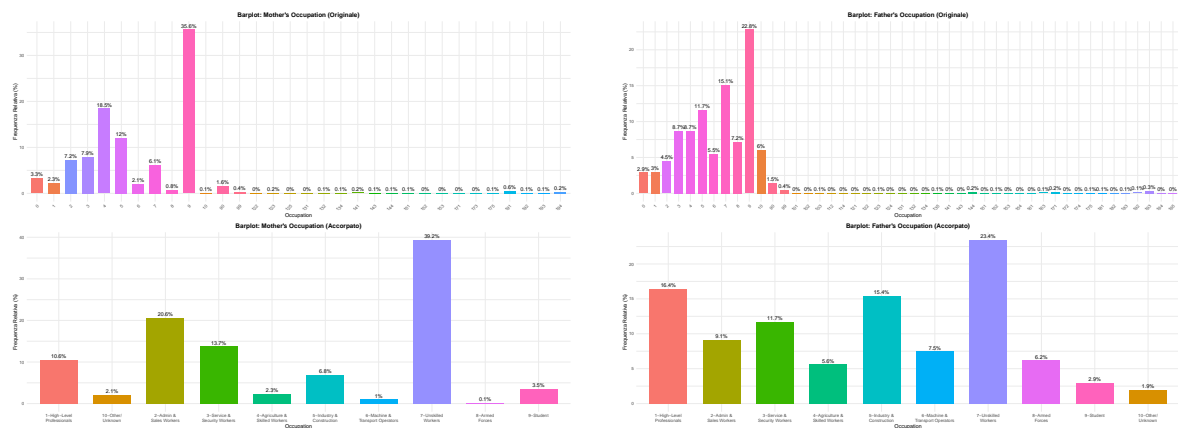


Figura 3.7: Effetto Aggregazione "Mother's / Father's Occupation"

Come illustrato nella Figura 3.7, l'aggregazione anche in questo caso ha ridotto sensibilmente il numero di categorie, migliorando il bilanciamento delle frequenze. Sebbene alcune delle nuove categorie per Mother's Occupation presentino ancora

frequenze molto basse, si è deciso di mantenerle per garantire coerenza tra le due variabili. In ogni caso, questo aspetto verrà tenuto in considerazione nelle analisi successive.

Nuova Categoria	Codici Originali	Descrizione
1 - High-Level Professionals	1 (Legislative/Executive Roles), 2 (Intellectual and Scientific Specialists), 3 (Intermediate Level Technicians), 112 (Admin & Commercial Directors), 114 (Hotel, Trade Directors), 121 (Physical Sciences Specialists), 122 (Health Professionals), 123 (Teachers), 124 (Finance & Admin Specialists), 132 (Health Technicians)	Professionisti di alto livello, dirigenti e specialisti
2 - Admin & Sales Workers	4 (Administrative Staff), 141 (Secretaries), 143 (Accounting & Registry Operators), 144 (Admin Support), 151 (Personal Service Workers), 152 (Sellers)	Impiegati amministrativi e addetti alle vendite
3 - Service & Security Workers	5 (Personal Services & Security), 153 (Personal Care Workers), 154 (Protection & Security), 134 (Legal, Social & Sports Technicians)	Addetti ai servizi e sicurezza
4 - Agriculture & Skilled Workers	6 (Farmers & Agriculture Workers), 161 (Market-Oriented Farmers), 163 (Subsistence Farmers)	Lavoratori agricoli e forestali
5 - Industry & Construction	7 (Skilled Industry & Construction Workers), 171 (Construction Workers), 172 (Metalworkers), 174 (Electricians), 175 (Food, Wood & Textile Workers)	Operai dell'industria e delle costruzioni
6 - Machine & Transport Operators	8 (Machine & Assembly Workers), 181 (Fixed Plant Operators), 182 (Assembly Workers), 183 (Vehicle Operators), 131 (Engineering Technicians), 135 (ICT Technicians)	Operatori di macchinari e trasporti
7 - Unskilled Workers	9 (Unskilled Laborers), 192 (Unskilled Agriculture Workers), 193 (Unskilled Industry Workers), 194 (Meal Prep Assistants), 195 (Street Vendors)	Lavoratori non qualificati
8 - Armed Forces	10 (Military), 101 (Armed Forces Officers), 102 (Sergeants), 103 (Other Military Personnel)	Forze armate
9 - Student	0 (Student)	Studenti
10 - Other / Unknown	90 (Other Situation), 99 (Blank)	Occupazione non specificata / disoccupati

Tabella 3.4: Aggregazione "Mother's / Father's Occupation"

Application Mode

La variabile *Application Mode* presentava diverse categorie, tra le quali molte rappresentavano situazioni particolari o regolamenti specifici con frequenze estremamente basse. Per semplificare l'analisi, le modalità di ammissione sono state raggruppate in cinque macro-categorie, come mostrato nella Tabella 3.5.

Nuova Categoria	Codici Originali	Descrizione
1 - General Contingent	1 (1st Phase - General Contingent), 17 (2nd Phase - General Contingent), 18 (3rd Phase - General Contingent)	Candidature standard attraverso il sistema di accesso generale
2 - Transfers & Changes	42 (Transfer), 43 (Change of Course), 51 (Change of Institution/Course), 57 (Change of Institution/Course - International)	Studenti che hanno cambiato corso o istituzione
3 - Older Students	39 (Over 23 Years Old)	Studenti ammessi tramite percorsi per adulti
4 - Special Diplomas	7 (Holders of Other Higher Courses), 44 (Technological Specialization Diploma Holders), 53 (Short Cycle Diploma Holders)	Studenti con titoli accademici superiori o percorsi formativi tecnici
5 - Special Contingent & Ordinances	2 (Ordinance No. 612/93), 5 (1st Phase - Special Contingent Azores), 10 (Ordinance No. 854-B/99), 15 (International Student - Bachelor), 16 (1st Phase - Special Contingent Madeira), 26 (Ordinance No. 533-A/99, item b2 - Different Plan), 27 (Ordinance No. 533-A/99, item b3 - Other Institution)	Categorie speciali di ammissione e regolamenti specifici

Tabella 3.5: Aggregazione "Application Mode"

Come mostrato nella Figura 3.8, questa suddivisione non ha migliorato molto il bilanciamento ma ha reso la variabile più chiara e interpretabile.

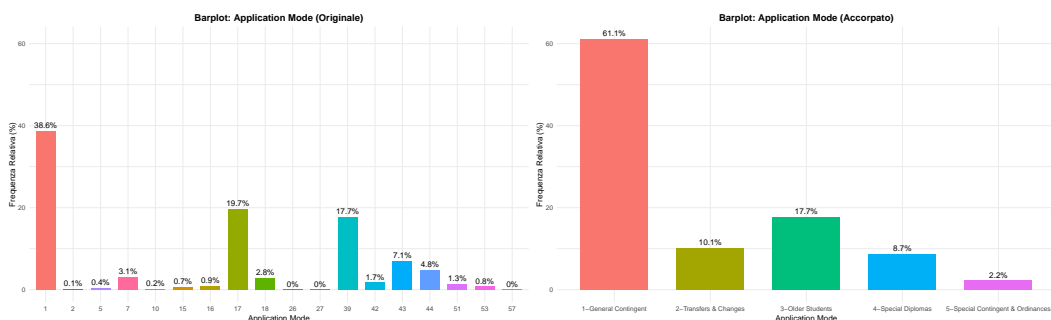


Figura 3.8: Effetto Aggregazione "Application Mode"

Previous Qualification

La variabile *Previous Qualification* rappresenta il titolo di studio posseduto dallo studente prima dell'iscrizione all'istituto. Tuttavia, solo una modalità risultava predominante sulle altre. Per migliorare l'interpretabilità e ridurre la presenza di categorie con frequenze molto basse, è stata effettuata un'aggregazione, come riportato nella Tabella 3.6.

Nuova Categoria	Codici Originali	Descrizione
1 - Higher Education	2 (Bachelor's Degree), 3 (Degree), 4 (Master's), 5 (Doctorate), 40 (Degree - 1st Cycle), 42 (Professional Higher Technical Course), 43 (Master - 2nd Cycle)	Titoli accademici universitari
2 - Secondary Education	1 (Secondary Education - 12th Year), 6 (Frequency of Higher Education)	Diploma di scuola superiore
3 - Technological Specialization	39 (Technological Specialization Course)	Diploma tecnico-specialistico post-secondario
4 - Basic Education	9 (12th Year Not Completed), 10 (11th Year Not Completed), 12 (Other - 11th Year), 14 (10th Year), 15 (10th Year Not Completed), 19 (Basic Education 3rd Cycle), 38 (Basic Education 2nd Cycle)	Percorsi di istruzione di base

Tabella 3.6: Aggregazione "Previous Qualification"

Come mostrato nella Figura 3.9, questa aggregazione ha permesso di ottenere una suddivisione più semplice, con categorie dalle frequenze più significative, seppur ancora fortemente sbilanciate.

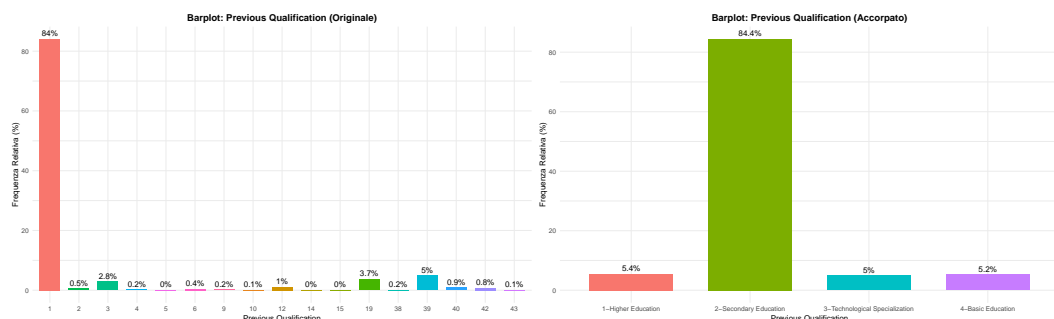


Figura 3.9: Effetto Aggregazione "Previous Qualification"

Course

La variabile *Course* identifica il corso di laurea a cui lo studente è iscritto. In questo caso la distribuzione non era eccessivamente sbilanciata, con solo alcune modalità caratterizzate da frequenze più estreme. Tuttavia, anche in questo caso è stata effettuata un'aggregazione in otto macro-aree disciplinari, come riportato nella Tabella 3.7, principalmente per semplificare il processo di generazione dei dati sintetici e migliorare l'interpretabilità.

Nuova Categoria	Codici Originali	Descrizione
1 - Health Sciences	9500 (Nursing), 9556 (Oral Hygiene)	Corsi di area sanitaria
2 - Veterinary Sciences	9085 (Veterinary Nursing), 9130 (Equin-culture)	Corsi di scienze veterinarie e zootecnia
3 - Social Sciences	8014 (Social Service - Evening), 9238 (So-cial Service), 9773 (Journalism and Com-munication), 9853 (Basic Education)	Corsi di scienze sociali e umanistiche
4 - Business & Management	9147 (Management), 9670 (Advertising and Marketing Management), 9991 (Ma-nagement - Evening)	Corsi di economia, marketing e gestione aziendale
5 - Communication & Design	171 (Animation and Multimedia Design), 9070 (Communication Design)	Corsi di design e comunicazione visiva
6 - Engineering & Technolo-gy	9119 (Informatics Engineering), 33 (Bio-fuel Production Technologies)	Corsi di ingegneria e tecnologia
7 - Agriculture & Environ-ment	9003 (Agronomy)	Corsi di agronomia e scienze ambientali
8 - Tourism & Services	9254 (Tourism)	Corsi di turismo e servizi

Tabella 3.7: Aggregazione "Course"

Come mostrato nella Figura 3.10, questa aggregazione ha reso la variabile più interpretabile, mantenendo comunque una distribuzione non estremamente sbilanciata tra le categorie.

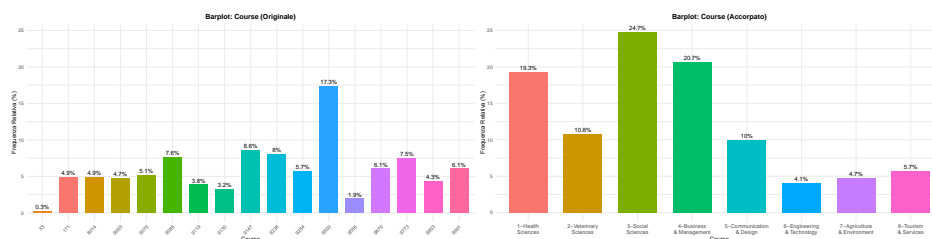


Figura 3.10: Effetto Aggregazione "Course"

Application Order

La variabile *Application Order* indica la preferenza con cui lo studente ha scelto il corso di laurea al momento dell'iscrizione, con valori compresi tra 1 (prima scelta) e 6 (sesta scelta). Tuttavia, erano presenti due valori anomali:

- **0:** non rappresentava una preferenza valida ed è stato accorpato a 1 (prima scelta);
- **9:** era un valore fuori scala ed è stato accorpato a 6 (sesta scelta).

Come si può osservare nella Figura 3.11, questa modifica ha avuto un impatto trascurabile sulla distribuzione complessiva, poiché le osservazioni con valori anomali erano solo due.

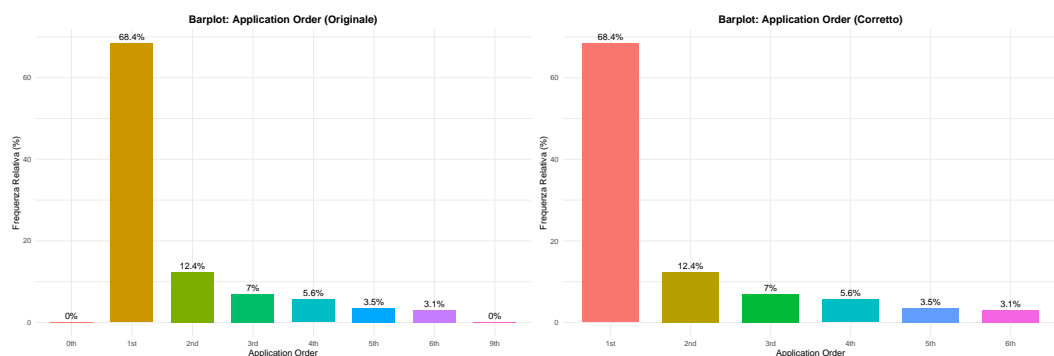


Figura 3.11: Correzione Anomalia di Application Order

3.3 Feature Construction

Dopo l'aggregazione delle feature categoriche, sono state create due nuove variabili con l'obiettivo di migliorare l'analisi delle performance accademiche degli studenti. L'intento è quello di sintetizzare informazioni chiave sulle unità curriculari in modo più interpretabile rispetto alle singole variabili originali. Infatti, variabili come il numero di valutazioni ricevute o il numero di unità curriculari superate (*approved*) potrebbero non essere indicatori ideali per confrontare gli studenti, poiché la struttura dei corsi potrebbe variare, rendendo difficile un confronto diretto. Per questo motivo, le nuove variabili proposte forniscono metriche standardizzate che permettono di valutare il progresso accademico degli studenti in maniera più oggettiva. Queste nuove

variabili verranno utilizzate nelle fasi successive per migliorare l'analisi bivariata, verificando se tali metriche siano associate agli esiti accademici, e nei capitoli successivi, per identificare possibili gruppi di studenti attraverso tecniche di clustering, permettendo di individuare pattern distintivi nelle loro performance accademiche.

Completed Exams Ratio

La prima variabile creata misura la percentuale di unità curriculari completate con successo rispetto a quelle a cui lo studente risulta iscritto nel primo anno accademico. Viene calcolata come:

$$\text{Completed Exams Ratio} = \left(\frac{\text{CU Approved}_{1\text{sem}} + \text{CU Approved}_{2\text{sem}}}{\text{CU Enrolled}_{1\text{sem}} + \text{CU Enrolled}_{2\text{sem}}} \right) \times 100$$

Questa metrica fornisce un'indicazione diretta del progresso accademico degli studenti rispetto al loro piano di studi, evidenziando eventuali ritardi nel completamento delle unità curriculari.

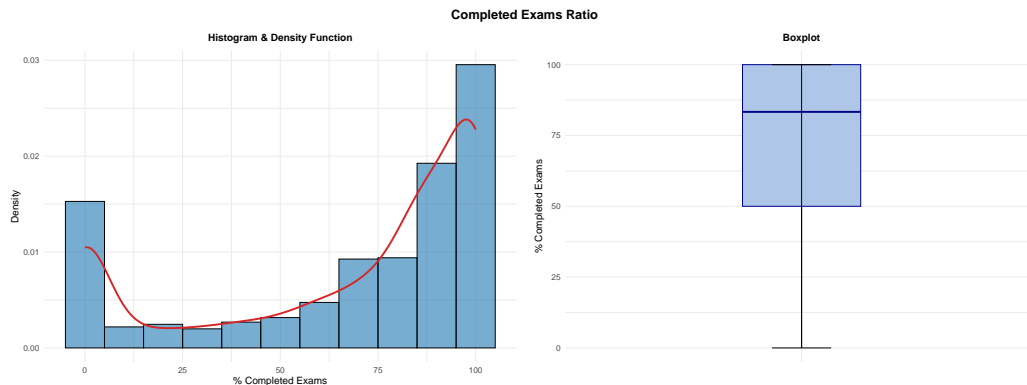


Figura 3.12: Analisi Univariata: Completed Exams Ratio

I grafici della variabile (Figura 3.12) mostrano una distribuzione quasi bimodale, con una parte significativa degli studenti concentrata vicino al valore massimo (**100%**), indicando che molti studenti completano con successo tutte le unità curriculari a cui si iscrivono, ed un picco vicino allo **0%**, suggerendo che alcuni studenti si iscrivono ai corsi ma non riescono a completarli. La distribuzione è fortemente asimmetrica, evidenziando una netta separazione tra studenti con prestazioni elevate e quelli con difficoltà accademiche.

Passed Exams Ratio

La seconda variabile misura l'efficacia degli studenti nel superare gli esami sostenuti. Viene calcolata come:

$$\text{Passed Exams Ratio} = \left(\frac{\text{CU Approved}_{1\text{sem}} + \text{CU Approved}_{2\text{sem}}}{\text{CU Evaluations}_{1\text{sem}} + \text{CU Evaluations}_{2\text{sem}}} \right) \times 100$$

Questa metrica permette di distinguere tra studenti che tentano molti esami senza successo e quelli che riescono a superare la maggior parte delle prove sostenute. Un valore basso indica che lo studente ha difficoltà nel superare gli esami, mentre un valore elevato suggerisce una maggiore efficacia nel completamento delle prove.

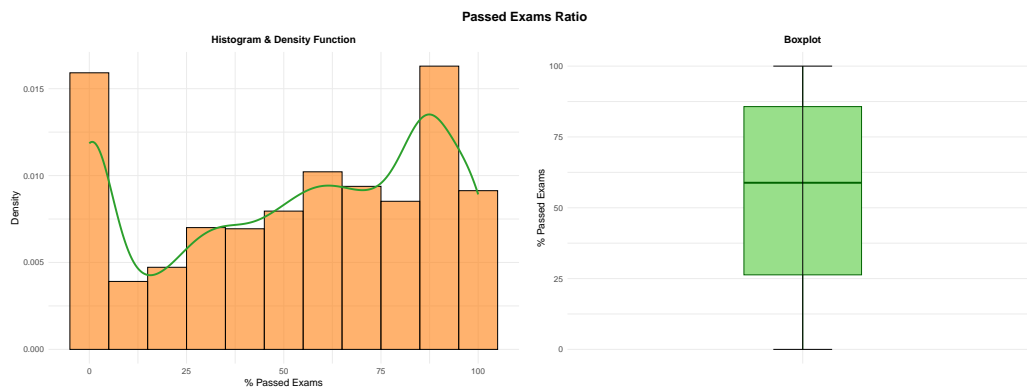


Figura 3.13: Analisi Univariata: Passed Exams Ratio

I grafici della variabile (Figura 3.13) mostrano una distribuzione più uniforme rispetto alla metrica precedente. Tuttavia, è presente un picco significativo in corrispondenza dello 0%, indicando che un numero consistente di studenti non è riuscito a superare alcun esame nonostante li abbia sostenuti. La distribuzione è più dispersa, con valori che si distribuiscono più uniformemente tra 0% e 100%, suggerendo una maggiore variabilità nella capacità degli studenti di superare gli esami sostenuti.

3.4 Fattori Demografici

L'analisi bivariata ha evidenziato come alcune caratteristiche demografiche sembrano essere associate agli esiti accademici degli studenti. Di seguito, vengono riportate le

principali osservazioni:

- **Gender** (Figura 3.14): L'analisi della relazione tra **Gender** e **Target** evidenzia una differenza significativa nei tassi di successo accademico tra studenti maschi e femmine. Le studentesse mostrano una maggiore probabilità di completare gli studi nei tempi previsti, con una percentuale di **Graduate** pari al **57.9%**, rispetto al **35.2%** degli studenti maschi. Al contrario, gli studenti di sesso maschile presentano un tasso di **Dropout** nettamente superiore (**45.1%**) rispetto alle studentesse (**25.1%**). Il tasso di **Enrolled** è simile tra i due gruppi, con una leggera prevalenza nei maschi (**19.7%** contro **17%**).

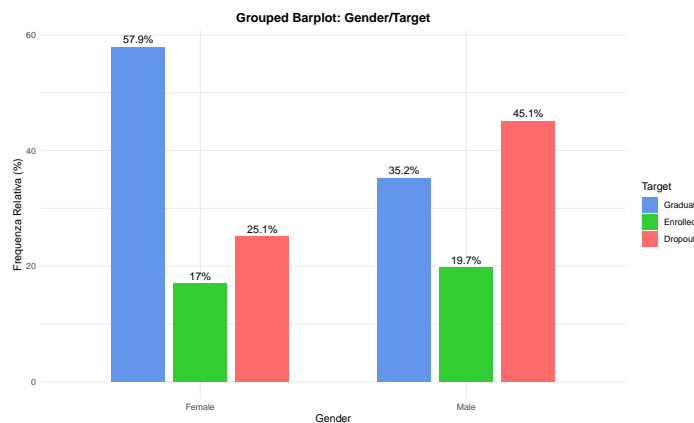


Figura 3.14: Analisi Bivariata: Gender VS Target

- **Marital Status** (Figura 3.15): Gli studenti **Single** mostrano il tasso di laurea più alto (**51.4%**) e il dropout più basso (**30.2%**), suggerendo una maggiore continuità accademica. Gli studenti **Married** hanno una percentuale inferiore di laureati (**39.4%**) e un tasso di **Dropout** più elevato (**47%**), mentre quelli **Previously Married** presentano una dinamica simile (**34.7%** di laureati, **46.5%** di dropout), ma con una quota leggermente maggiore di studenti ancora iscritti (**18.8%**). Questi dati suggeriscono che gli impegni familiari possano influenzare negativamente la carriera accademica.

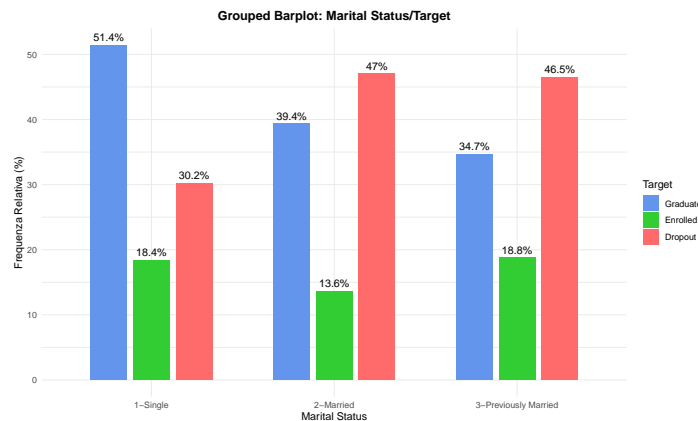


Figura 3.15: Analisi Bivariata: Marital Status VS Target

- **Nationality** (Figura 3.16): Non emergono differenze significative tra studenti **Portoghesi** e **Internazionali** in termini di esiti accademici. La percentuale di laureati è quasi identica (**50%** e **49.1%**), così come il tasso di **Dropout** (rispettivamente **32.2%** e **29.1%**). Anche la quota di studenti ancora iscritti senza laurearsi varia di poco (**17.8%** e **21.8%**). Questi risultati indicano che la nazionalità non sembra avere un impatto rilevante sulla carriera accademica.

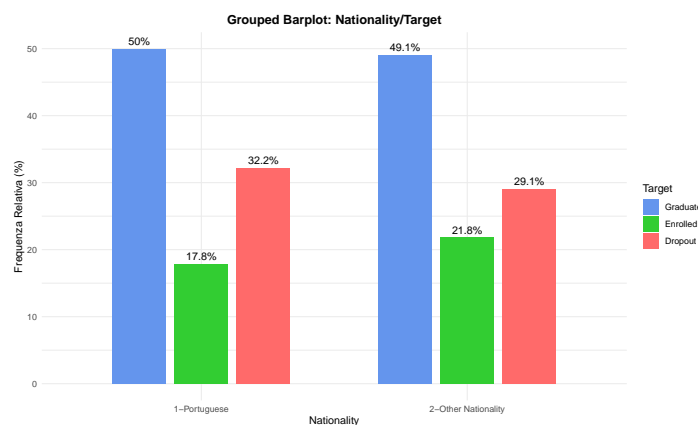


Figura 3.16: Analisi Bivariata: Nationality VS Target

- **Age at Enrollment** (Figura 3.17): L'età di immatricolazione sembrerebbe mostrare una relazione con l'esito accademico. Gli studenti che si immatricolano in giovane età sembrano avere una maggiore tendenza a laurearsi nei tempi previsti. Al contrario, gli studenti **Dropout** presentano una distribuzione più ampia, con un'età media più elevata e una coda lunga che si estende fino ai **46+ anni**. Gli studenti ancora iscritti senza laurearsi mostrano una distribuzione

intermedia, più simile a quella dei laureati. Questi risultati suggeriscono che un'età avanzata all'immatricolazione sia associata a un rischio maggiore di abbandono. Infine, è importante notare la presenza di numerosi **outlier** nei gruppi **Graduate** ed **Enrolled**, il che indica che, sebbene meno frequente, ci sono studenti che riescono a completare o continuare gli studi anche iscrivendosi in età più avanzata.

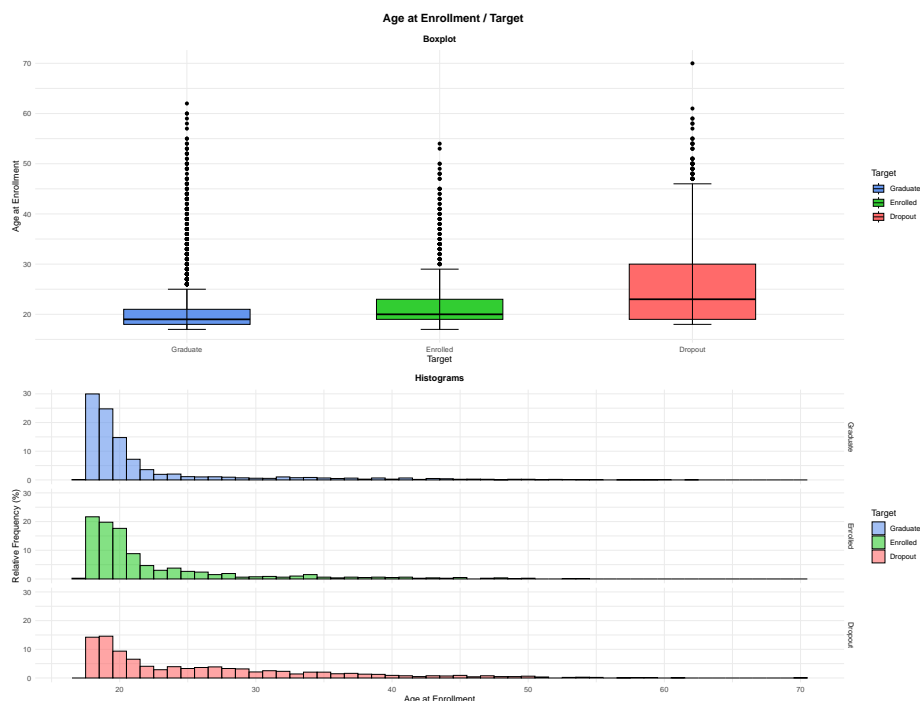


Figura 3.17: Analisi Bivariata: Age At Enrollment VS Target

Un'analisi incrociata tra **Marital Status** e **Age at Enrollment** ha evidenziato che gli studenti sposati o precedentemente sposati tendono a immatricolarsi ad un'età più avanzata. Questo potrebbe suggerire che il maggiore tasso di **Dropout** tra gli studenti più adulti sia almeno in parte attribuibile agli impegni familiari, che potrebbero rendere più difficile il completamento del percorso accademico.

3.5 Fattori Socio-Economici

L'analisi bivariata dei fattori socio-economici ha evidenziato che le condizioni economiche sembrano essere particolarmente associate al successo e al fallimento accademico degli studenti. Di seguito, vengono riportate le principali osservazioni:

- **Debtor** (Figura 3.18): Gli studenti senza debiti economici mostrano una probabilità di laurearsi significativamente più alta (**53.8%**), mentre solo il **20.1%** degli studenti con debiti riesce a completare il percorso di studi. Inoltre, il tasso di **Dropout** tra gli studenti debitori (**62%**) è più che doppio rispetto a quello degli studenti senza debiti (**28.3%**). Il tasso di studenti ancora iscritti (**Enrolled**) è simile tra i due gruppi (**18%** contro **17.9%**). Questi risultati suggeriscono che difficoltà economiche possano costituire un fattore di rischio significativo per l'abbandono universitario.

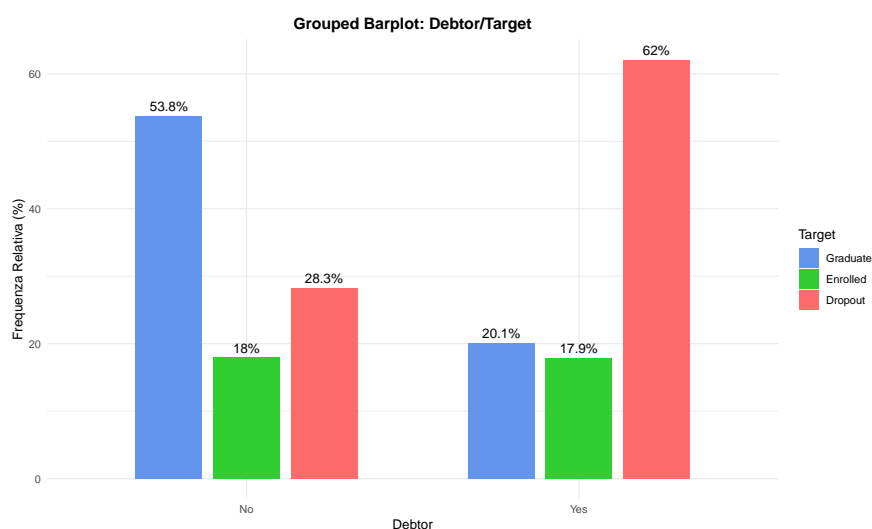


Figura 3.18: Analisi Bivariata: Debtor VS Target

- **Tuition Fees Up To Date** (Figura 3.19): Il pagamento regolare delle tasse universitarie sembra essere fortemente associato al successo accademico. Gli studenti che sono in regola con i pagamenti mostrano una percentuale di **Graduate** significativamente più alta (**56%**) rispetto a quelli che non lo sono (**5.5%**). Al contrario, il **Dropout** è predominante tra coloro che non sono in regola, con un tasso dell'**86.6%**, mentre tra chi ha pagato regolarmente le tasse universitarie la percentuale di abbandono scende drasticamente al **24.7%**. Questo dato evidenzia il legame tra difficoltà economiche e interruzione del percorso accademico, suggerendo che il mancato pagamento delle tasse possa essere un indicatore precoce del rischio di abbandono.

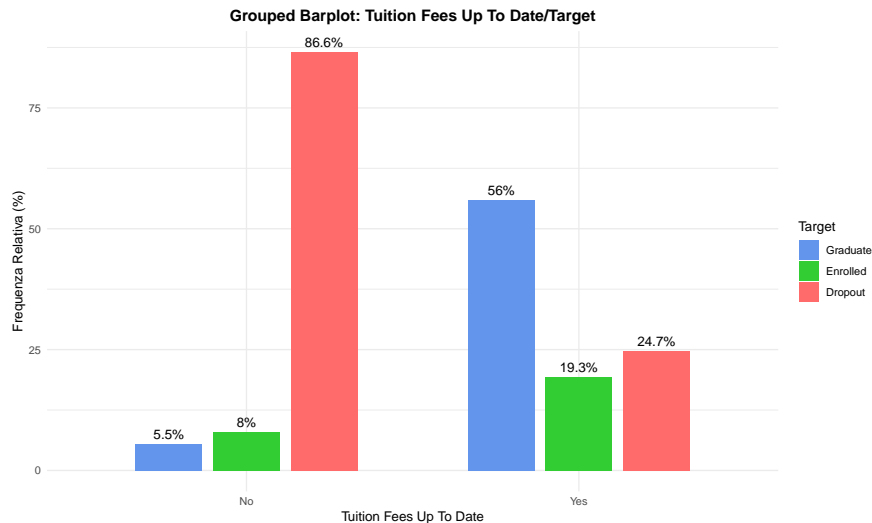


Figura 3.19: Analisi Bivariata: Tuition Fees Up to Date VS Target

- Scholarship Holder** (Figura 3.20): Il possesso di una borsa di studio sembra avere un impatto positivo sulla probabilità di completare il percorso accademico. Gli studenti borsisti mostrano un tasso di **Graduate** significativamente più alto (76%) rispetto a chi non beneficia di una borsa di studio (41.3%). Inoltre, il tasso di **Dropout** tra i borsisti è notevolmente inferiore (12.2%) rispetto a chi non riceve supporto finanziario (38.7%). Questi dati suggeriscono che il supporto economico fornito dalle borse di studio possa contribuire a ridurre il rischio di abbandono e favorire il completamento degli studi.

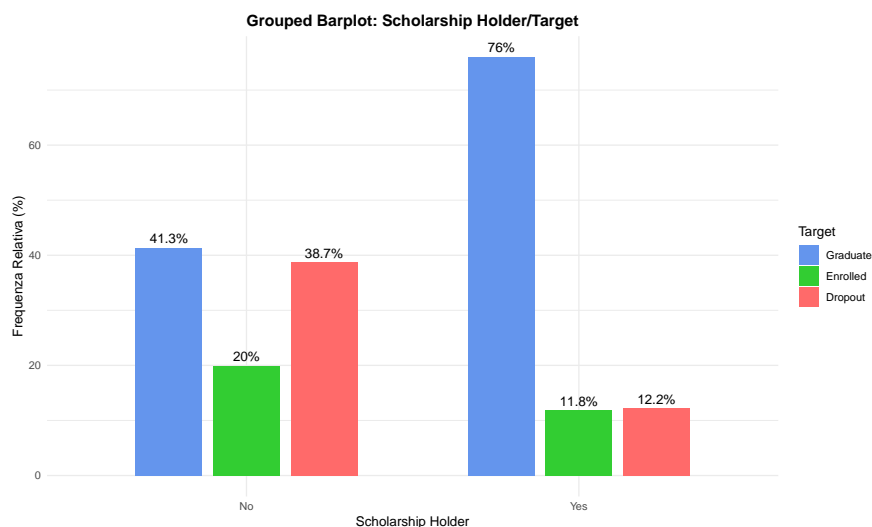


Figura 3.20: Analisi Bivariata: Scholarship Holder VS Target

- Mother's / Father's Qualification** (Figura 3.21): Dall'analisi bivariata delle qualifiche dei genitori, non si osservano differenze sostanziali tra i vari livelli di istruzione, ad eccezione della categoria **6-No Qualification**, che presenta una percentuale di **Dropout** nettamente superiore (73.5% per le madri e 72.1% per i padri), suggerendo che un basso livello di istruzione familiare possa rappresentare un fattore di rischio per l'abbandono universitario. Per le altre categorie, i tassi di successo accademico rimangono relativamente stabili, senza differenze marcate tra i diversi livelli di istruzione. Da un'analisi incrociata tra l'età di immatricolazione e il livello di istruzione dei genitori, è emerso che gli studenti con genitori privi di qualifiche accademiche tendono ad immatricolarsi ad un'età mediamente più elevata. Di conseguenza, l'apparente associazione tra il basso livello di istruzione dei genitori e il rischio di abbandono potrebbe essere influenzata principalmente dall'età di immatricolazione più avanzata, che sembra essere ancora una volta un fattore rilevante nella determinazione del successo o del fallimento accademico.

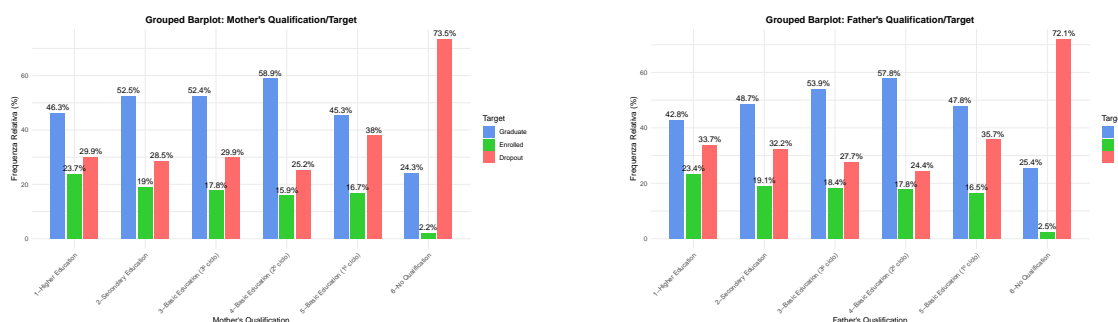


Figura 3.21: Analisi Bivariata: Parent's Qualification VS Target

- Mother's / Father's Occupation** (Figura 3.22): L'occupazione dei genitori non sembra mostrare una relazione forte con l'esito accademico dello studente, con distribuzioni relativamente simili tra le varie categorie. Tuttavia, emergono due eccezioni: gli studenti con genitori che rientrano nella categoria **Student** o **Other/Unknown** mostrano tassi di **Dropout** nettamente superiori rispetto agli altri gruppi. In particolare, per i padri classificati come **Other/Unknown**, il tasso di abbandono raggiunge il **70.2%**, e per le madri nella stessa categoria arriva al **73.6%**. Mentre gli studenti con genitori classificati come **Student** presentano una probabilità di dropout del **64.1%** per il padre e del **68.8%** per la madre.

Al contrario, nelle categorie occupazionali più strutturate (ad esempio, **High-Level Professionals** e **Industry & Construction**), i tassi di laurea risultano più elevati e il dropout più contenuto. Tuttavia, la differenza tra questi gruppi è meno marcata rispetto alle categorie precedentemente discusse. Questi dati potrebbero suggerire che gli studenti le cui famiglie hanno entrate economiche minori, perché i genitori sono essi stessi ancora studenti o non lavorano, siano più a rischio. Infine, è opportuno segnalare che le categorie **6 - Machine & Transport Operators** e **8 - Armed Forces** per la madre presentano frequenze relative molto basse (come già evidenziato nella Sezione 3.2) ed è bene non considerarle in questa analisi, poiché non permettono di trarre osservazioni statisticamente significative.

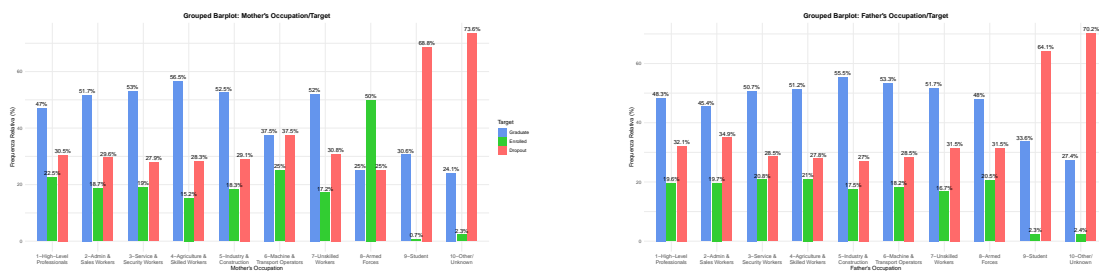


Figura 3.22: Analisi Bivariata: Parent's Occupation VS Target

Dopo aver analizzato più nel dettaglio la condizione degli studenti con **debiti finanziari** e non in regola con il pagamento delle tasse, è emerso che tra gli studenti che rientrano in questa categoria e che abbandonano gli studi, solo il **14%** riceve una borsa di studio. Al contrario, tra coloro che, pur avendo difficoltà economiche, riescono a laurearsi nei tempi previsti, la percentuale di borsisti sale al **46%**. Questo conferma che il supporto economico e motivazionale di una borsa di studio potrebbe ridurre significativamente il rischio di abbandono.

3.6 Dati sull'Iscrizione e Accesso ai Corsi

L'analisi bivariata dei dati relativi all'iscrizione e all'accesso ai corsi ha evidenziato alcune differenze significative tra i gruppi di studenti, suggerendo che fattori come

la modalità di accesso, la frequenza e la tipologia stessa del corso possano essere associati agli esiti accademici. Di seguito vengono riportate le principali osservazioni:

- **Displaced** (Figura 3.23): Gli studenti **fuori sede** mostrano una percentuale di **Graduate** più alta (**54.6%**) rispetto agli studenti **in sede** (**44.3%**). Inoltre, il tasso di **Dropout** tra gli studenti **in sede** è superiore (**37.6%** contro **27.6%** dei fuori sede). La percentuale di studenti ancora iscritti (**Enrolled**) è pressoché equivalente nei due gruppi (**18.1%** per gli studenti in sede e **17.8%** per i fuori sede). Questi dati suggeriscono che gli studenti **fuori sede** possano avere una maggiore motivazione o beneficiare di un ambiente più favorevole al completamento del percorso accademico.

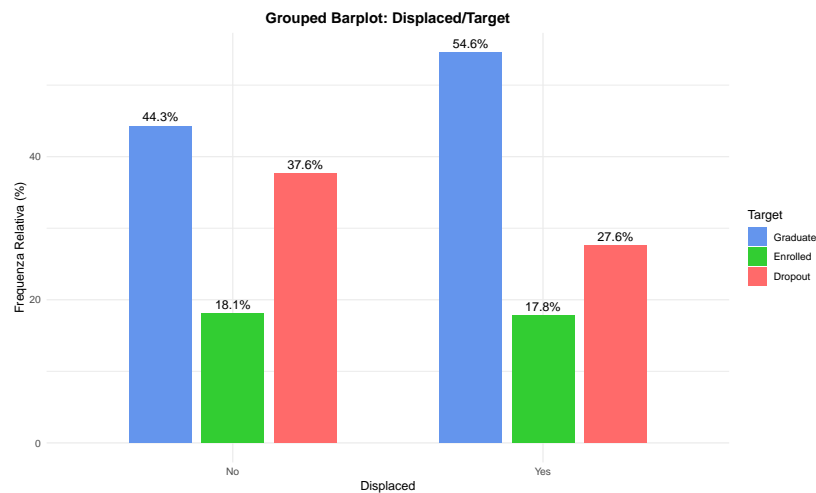


Figura 3.23: Analisi Bivariata: Displaced VS Target

- **Educational Special Needs** (Figura 3.24): L'analisi della presenza di Bisogni Educativi Speciali (BES) non evidenzia differenze marcate negli esiti accademici. Gli studenti senza BES presentano un tasso di laurea leggermente superiore (**50%** contro **45.1%**) e un tasso di dropout leggermente inferiore (**32.1%** rispetto a **33.3%**) rispetto a quelli con BES. Tuttavia, la quota di studenti ancora iscritti senza laurearsi risulta leggermente più alta tra gli studenti con BES (**21.6%** rispetto a **17.9%**). Questo suggerisce che, pur non avendo un impatto significativo sull'abbandono universitario, la presenza di bisogni educativi speciali potrebbe essere associata a una maggiore probabilità di prolungare la durata degli studi.

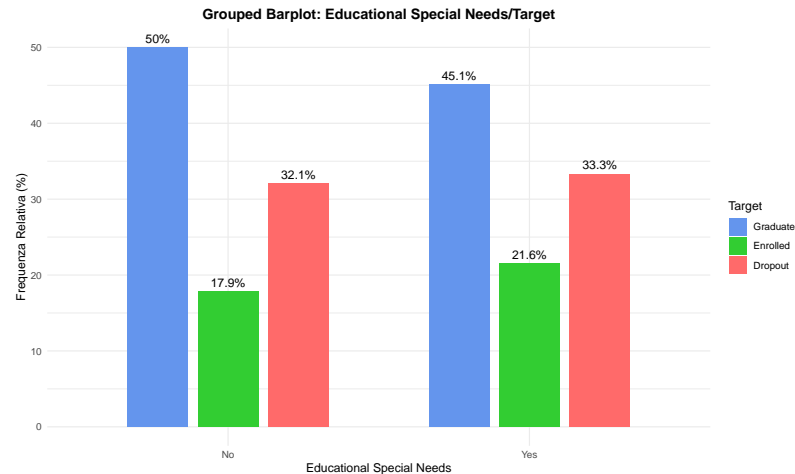


Figura 3.24: Analisi Bivariata: BES VS Target

- Daytime / Evening Attendance** (Figura 3.25): L'orario di frequenza dei corsi mostra una certa influenza sull'esito accademico. Gli studenti che frequentano i corsi diurni presentano un tasso di laurea più elevato (51%) rispetto a quelli che frequentano in orario serale (41.6%). Al contrario, il tasso di dropout è più alto tra gli studenti serali (42.9%) rispetto ai corsisti diurni (30.8%). La percentuale di studenti ancora iscritti senza laurearsi risulta invece simile tra i due gruppi (15.5% per gli studenti serali e 18.2% per quelli diurni). Questi dati suggeriscono che gli studenti serali possano incontrare maggiori difficoltà nel completare il percorso accademico, probabilmente a causa di impegni lavorativi o personali che interferiscono con il rendimento universitario.

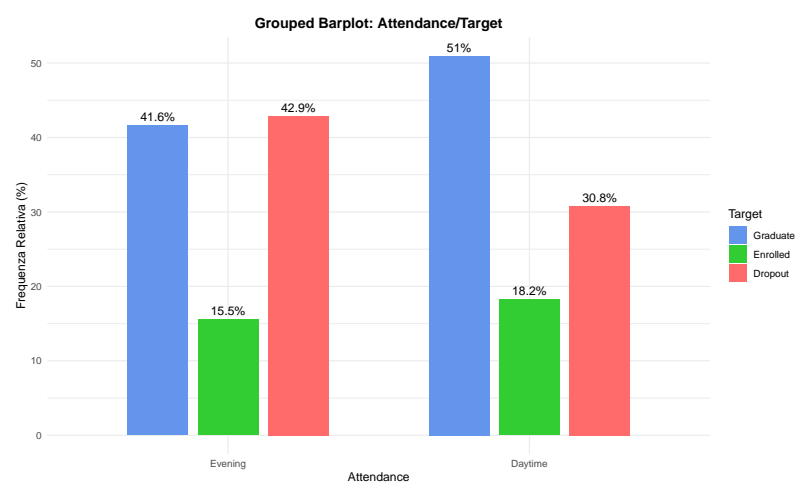


Figura 3.25: Analisi Bivariata: Attendance VS Target

- **International** (Figura 3.26): L'analisi della variabile International conferma quanto già osservato in precedenza con la variabile Nationality, che dopo l'aggregazione è risultata equivalente a questa. Non emergono differenze significative tra studenti internazionali e locali in termini di esiti accademici. Il tasso di laurea è pressoché identico tra i due gruppi (50% per gli studenti locali e 49.1% per gli studenti internazionali), così come la percentuale di **Dropout** (32.2% contro 29.1%). Anche la quota di studenti ancora iscritti senza laurearsi è molto simile (17.8% per gli studenti locali e 21.8% per gli internazionali). Questi risultati suggeriscono che la condizione di studente Erasmus o internazionale non rappresenta un fattore determinante per il successo accademico

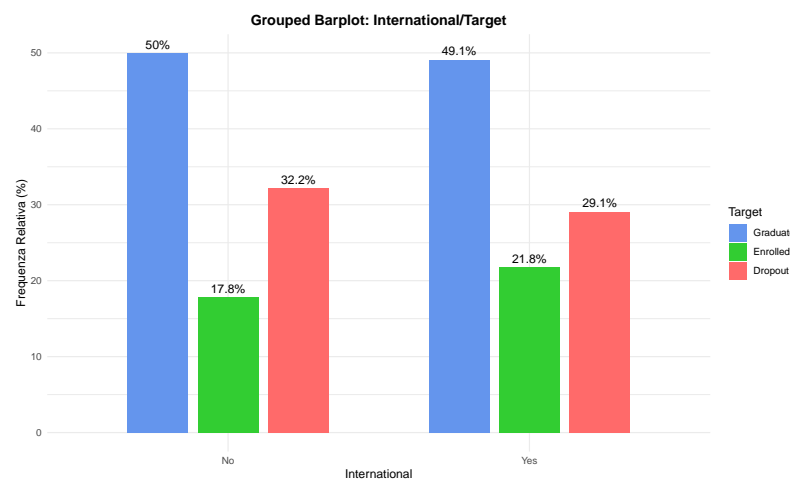


Figura 3.26: Analisi Bivariata: International VS Target

- **Application Mode** (Figura 3.27): L'analisi della modalità di iscrizione evidenzia differenze significative nei tassi di successo accademico. Gli studenti che si sono immatricolati tramite il **General Contingent** mostrano il tasso più alto di **Graduate** (58.5%) e il dropout più basso (23.9%). Anche la categoria **Special Contingent & Ordinances** presenta un tasso di laurea elevato (54.5%) e un dropout relativamente contenuto (19.2%). Al contrario, gli studenti immatricolati nella categoria **Over 23 years old** presentano la percentuale di **Dropout** più alta (55.4%), confermando ancora una volta che l'età di immatricolazione può essere un fattore determinante per il rischio di abbandono. Il loro tasso di laurea è nettamente inferiore rispetto agli altri gruppi (29.2%), in linea con quanto osservato nell'analisi di *Age At Enrollment*. Anche gli studenti che si

immatricolano attraverso le categorie **Transfers & Changes** o **Special Diplomas** mostrano tassi di dropout superiori rispetto al **General Contingent**, ma meno marcati rispetto alla categoria Over 23.

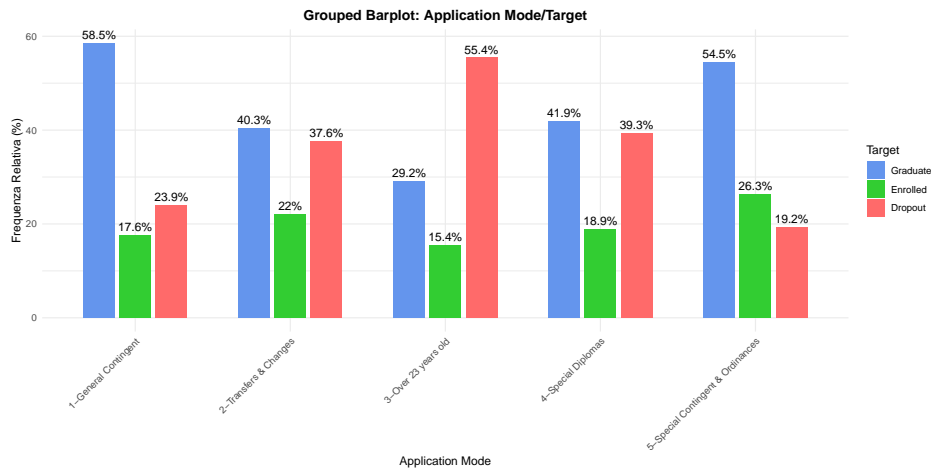


Figura 3.27: Analisi Bivariata: Application Mode VS Target

- **Application Order** (Figura 3.28): L'analisi bivariata di **Application Order** mostra che la probabilità di laurearsi tende ad aumentare con l'aumentare dell'ordine di preferenza, mentre la probabilità di **Dropout** diminuisce. Gli studenti che hanno scelto il loro corso come **6^a opzione** mostrano il tasso di laurea più alto (**65.2%**) e il tasso di abbandono più basso (**22.5%**), mentre coloro che hanno scelto il corso come **1^a opzione** presentano una percentuale inferiore di laureati (**46.5%**) e un dropout più elevato (**34.8%**). Tuttavia, la categoria relativa alla **5^a opzione** si discosta da questa tendenza, mostrando un tasso di dropout più alto (**34.4%**) rispetto alle altre opzioni avanzate e un tasso di laurea più basso (**49.4%**). Questa deviazione non sembra avere una spiegazione logica evidente e potrebbe essere dovuta a fluttuazioni casuali nei dati piuttosto che a un vero fenomeno strutturale. Inoltre, come osservato durante l'analisi univariata, con il diminuire dell'ordine di preferenza, diminuisce anche il numero di osservazioni nel dataset. Questo potrebbe contribuire alle variazioni nei tassi di successo accademico, rendendo alcune differenze meno robuste dal punto di vista statistico.

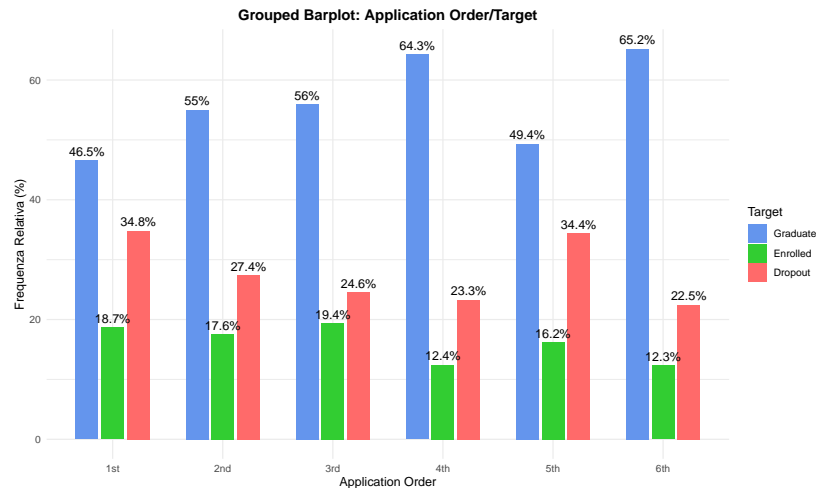


Figura 3.28: Analisi Bivariata: Application Order VS Target

- Previous Qualification** (Figura 3.29): Gli studenti con una qualifica di **Basic Education** mostrano il tasso di **Dropout** più elevato (**64.2%**), probabilmente a causa di lacune nelle competenze accademiche. Anche gli studenti con una qualifica di **Higher Education** pregressa presentano un tasso di abbandono relativamente alto (**49.2%**), il che potrebbe essere influenzato dall'età più elevata o dalla difficoltà dei percorsi avanzati che richiedono titoli accademici precedenti. Le categorie **Secondary Education** e **Technological Specialization** mostrano tassi di **Graduate** simili (**52.2%** e **43.4%** rispettivamente), ma gli studenti con una formazione tecnologica tendono ad avere un tasso più alto di studenti ancora iscritti senza laurearsi (**25.1%**), suggerendo possibili difficoltà nel completare il percorso nei tempi previsti.

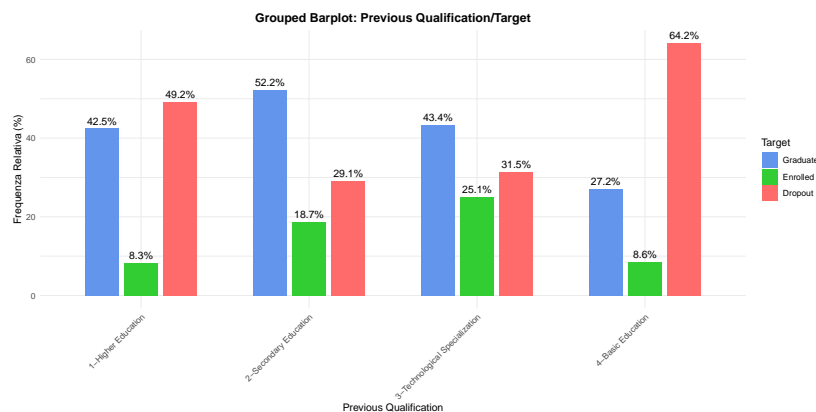


Figura 3.29: Analisi Bivariata: Previous Qualification VS Target

- Course** (Figura 3.30): L'area disciplinare del corso di studi mostra una forte associazione con l'esito accademico. Gli studenti iscritti ai corsi di **Health Sciences** presentano il tasso di **Graduate** più alto (68.5%) e il tasso di **Dropout** più basso (13.7%). Al contrario, **Engineering & Technology** mostra la percentuale di dropout più elevata (54.9%) e il tasso di laurea più basso (8.2%), suggerendo difficoltà accademiche più marcate in questo settore. Le facoltà di **Veterinary Sciences**, **Social Sciences** e **Business & Management** mostrano una distribuzione più bilanciata tra successo accademico e abbandono, con tassi di dropout compresi tra il 29.5% e il 39.8%. **Tourism & Services** e **Agriculture & Environment** registrano tassi di dropout relativamente elevati (38.1% e 41% rispettivamente), ma mantengono una quota discreta di laureati. Questi dati evidenziano differenze significative tra le diverse discipline accademiche, suggerendo che alcuni corsi possano presentare maggiori difficoltà, influenzando così la probabilità di completamento degli studi.

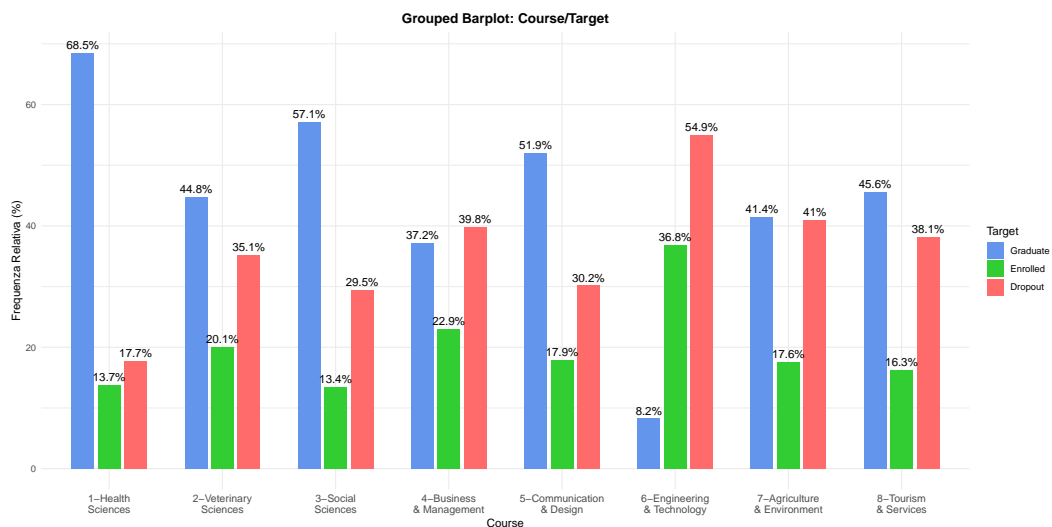


Figura 3.30: Analisi Bivariata: Course VS Target

- Previous Qualification Grade** (Figura 3.31): Il voto della qualifica precedente non sembra essere un fattore determinante negli esiti accademici. L'analisi dei **boxplot** mostra distribuzioni molto simili tra i tre gruppi (**Graduate**, **Enrolled** e **Dropout**), con mediane e intervalli interquartili pressoché sovrapponibili. Anche la funzione di densità nei **violin plot** non evidenzia grandi differenze tra i gruppi, ad eccezione degli studenti **Dropout**, che mostrano un picco di

densità più accentuato attorno alla votazione di 130. Questo potrebbe indicare una maggiore concentrazione di studenti che abbandonano in quella fascia di voti, ma non sembra esserci una relazione chiara con i risultati accademici.

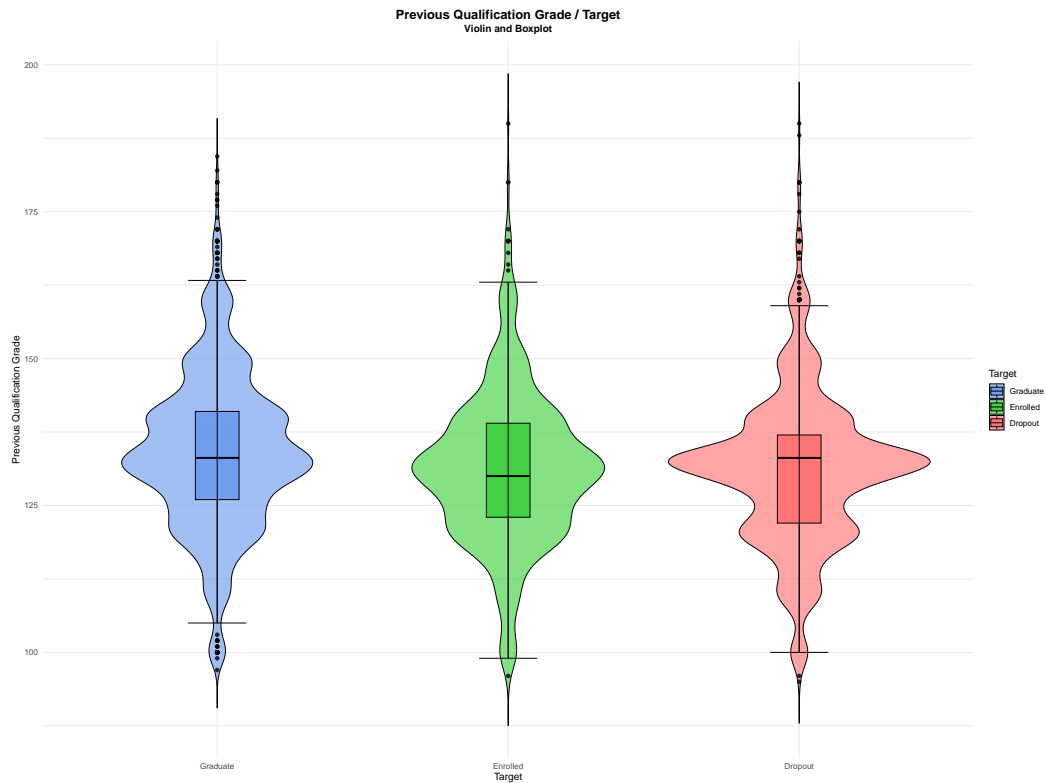


Figura 3.31: Analisi Bivariata: Previous Qualification Grade VS Target

- **Admission Grade** (Figura 3.32): Anche il voto di ammissione non sembra essere un fattore distintivo significativo tra i diversi esiti accademici. L'analisi dei **boxplot** mostra distribuzioni molto simili per i tre gruppi (**Graduate**, **Enrolled** e **Dropout**), con mediane e intervalli interquartili quasi sovrapponibili. Tuttavia, osservando la funzione di densità nei **violin plot**, si nota che gli studenti **Graduate** tendono ad avere una densità leggermente più alta per votazioni elevate, mentre i gruppi **Dropout** ed **Enrolled** presentano una maggiore concentrazione attorno ai punteggi più bassi.

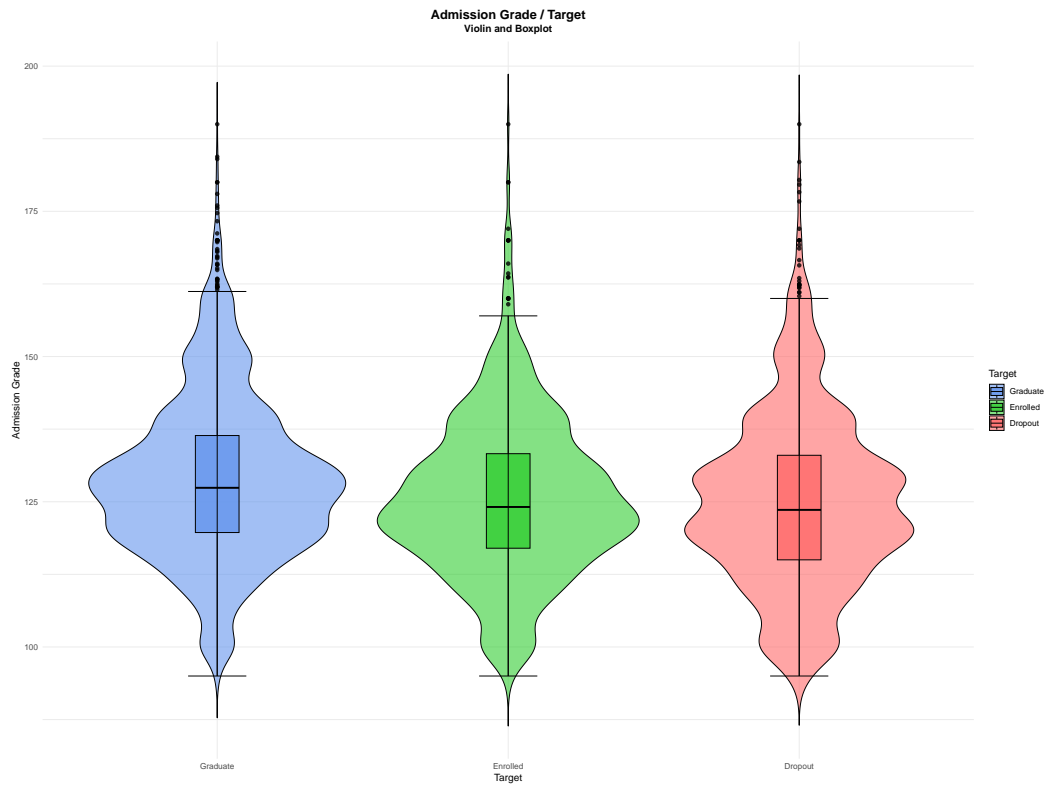


Figura 3.32: Analisi Bivariata: Admission Grade VS Target

3.7 Informazioni Primo Anno Accademico

Prima di procedere con l'analisi bivariata delle variabili relative al primo anno accademico, il dataset è stato opportunamente ripulito. In particolare, sono stati rimossi **180** studenti che presentavano valori pari a **0** per tutte le variabili relative ai due semestri, poiché questi dati sembravano indicare la mancanza di informazioni piuttosto che valori validi. Inoltre, sono stati eliminati **164** studenti che risultavano iscritti a delle unità curriculari ma non avevano ricevuto alcuna valutazione in entrambi i semestri. Poiché non era chiaro se questi dati rappresentassero effettivamente informazioni mancanti o meno, si è deciso di escluderli per evitare distorsioni nell'analisi. L'analisi bivariata di questi dati ha permesso di osservare che alcune delle informazioni raccolte durante il primo anno accademico sembrano essere particolarmente associate alla variabile Target. Di seguito vengono riportate le principali osservazioni:

- **Curricular Units Credited** (Figura 3.33): Il numero di unità curriculari riconosciute nei primi due semestri non sembra differire significativamente tra i

tre gruppi (**Graduate**, **Enrolled**, **Dropout**). I **boxplot** evidenziano una forte concentrazione di valori intorno allo zero per tutti i gruppi, con la presenza di **outlier** che si estendono fino a circa 20 unità. Anche le distribuzioni degli **istogrammi** risultano molto simili, con la maggior parte degli studenti che ha accreditato poche o nessuna unità in entrambi i semestri.

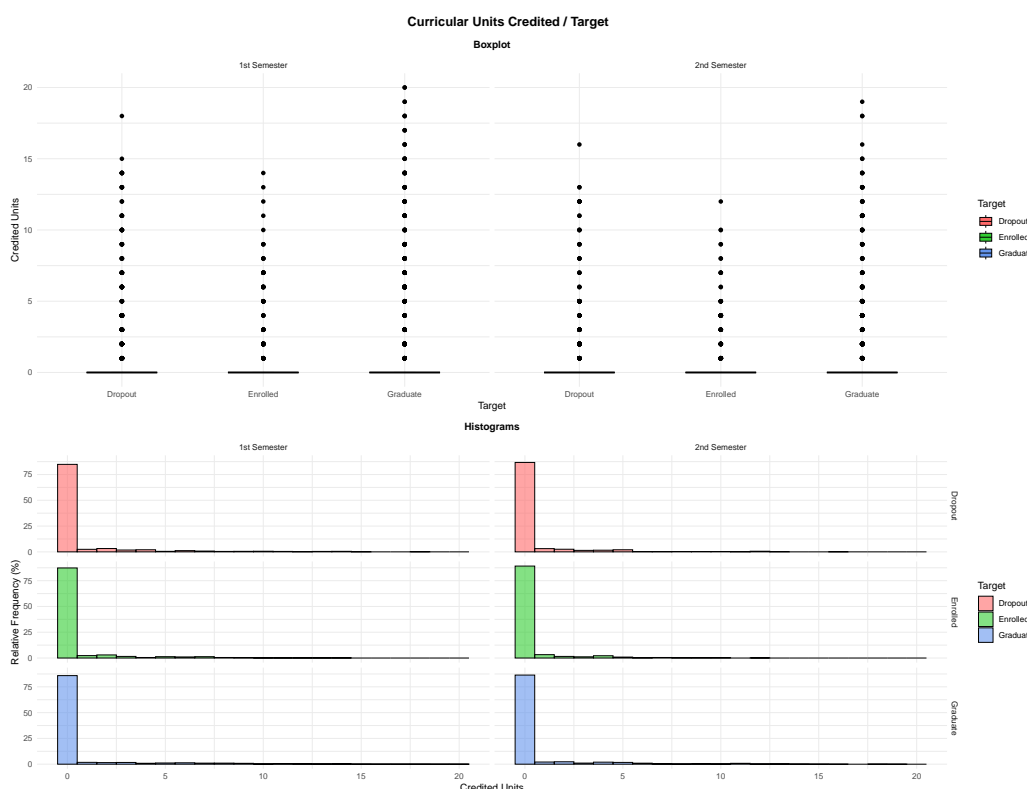


Figura 3.33: Analisi Bivariata: CU Credited VS Target

- **Curricular Units Enrolled** (Figura 3.34): Il numero di unità curriculari a cui gli studenti si sono iscritti nei primi due semestri non sembra essere un forte discriminante tra i tre gruppi (**Graduate**, **Enrolled**, **Dropout**). I **boxplot** mostrano mediane perfettamente sovrapposte, indicando una distribuzione simile delle iscrizioni ai corsi. Tuttavia, analizzando gli **istogrammi**, si nota che gli studenti **Enrolled** e **Dropout** tendono a concentrarsi su valori leggermente più bassi rispetto ai **Graduate**, la cui distribuzione appare più centrata su un numero maggiore di unità curriculari. Questo potrebbe suggerire che studenti iscritti a un minor numero di corsi abbiano una maggiore tendenza a finire fuori corso o ad abbandonare gli studi.

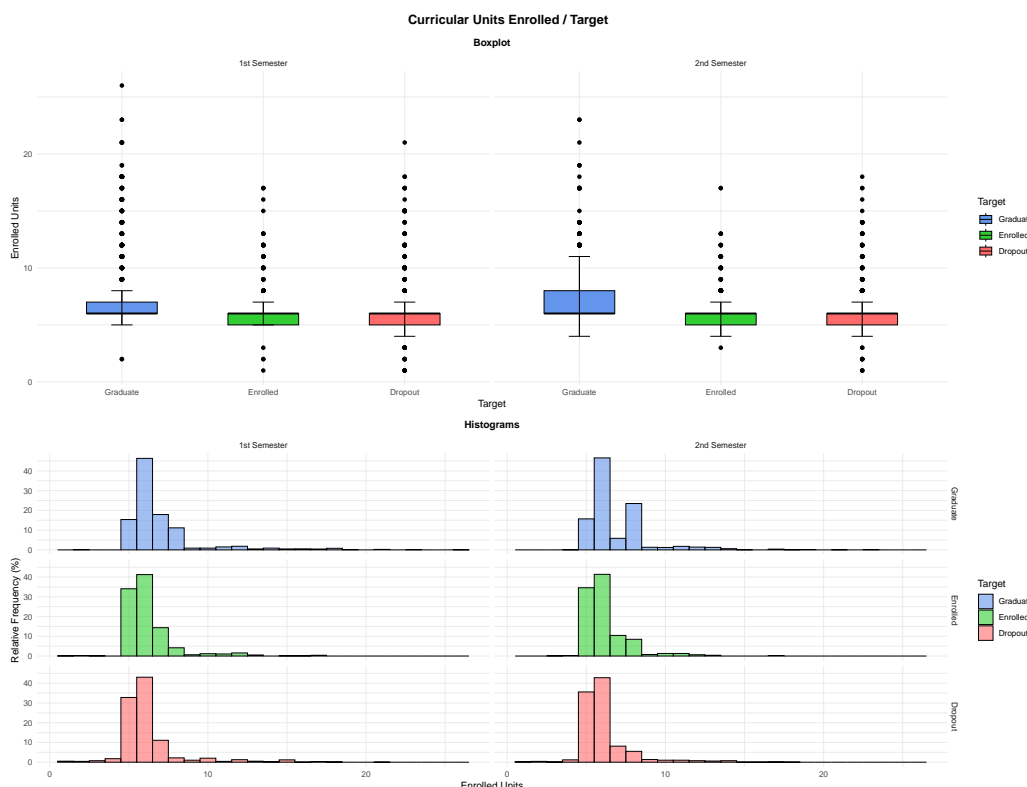


Figura 3.34: Analisi Bivariata: CU Enrolled VS Target

- Curricular Units Evaluations** (Figura 3.33): Il numero di valutazioni ricevute dagli studenti nei primi due semestri mostra alcune differenze tra i gruppi. I **boxplot** indicano che gli studenti **Graduate** tendono ad avere un numero di valutazioni più concentrato verso valori più bassi rispetto agli **Enrolled** e ai **Dropout**, suggerendo che questi ultimi potrebbero tentare gli esami più volte senza riuscire a superarli. Inoltre, nel **secondo semestre**, è presente un leggero picco di studenti **Dropout** con **zero valutazioni**, il che suggerisce che coloro che non sostengono alcuna prova d'esame siano particolarmente a rischio di abbandono.
- Curricular Units Approved** (Figura 3.33): L'analisi del numero di unità curriculari superate mostra differenze significative tra i tre gruppi. Gli studenti **Graduate** tendono a concentrarsi su valori più elevati, evidenziando un numero maggiore di esami superati. Gli **Enrolled** presentano una distribuzione più ampia con valori generalmente più bassi rispetto ai **Graduate**, includendo anche studenti che non hanno superato alcun esame. Il gruppo **Dropout**

è prevalentemente concentrato su valori bassi, con un picco particolarmente pronunciato in corrispondenza dello 0, suggerendo una forte associazione tra il mancato superamento di esami e l'abbandono universitario.

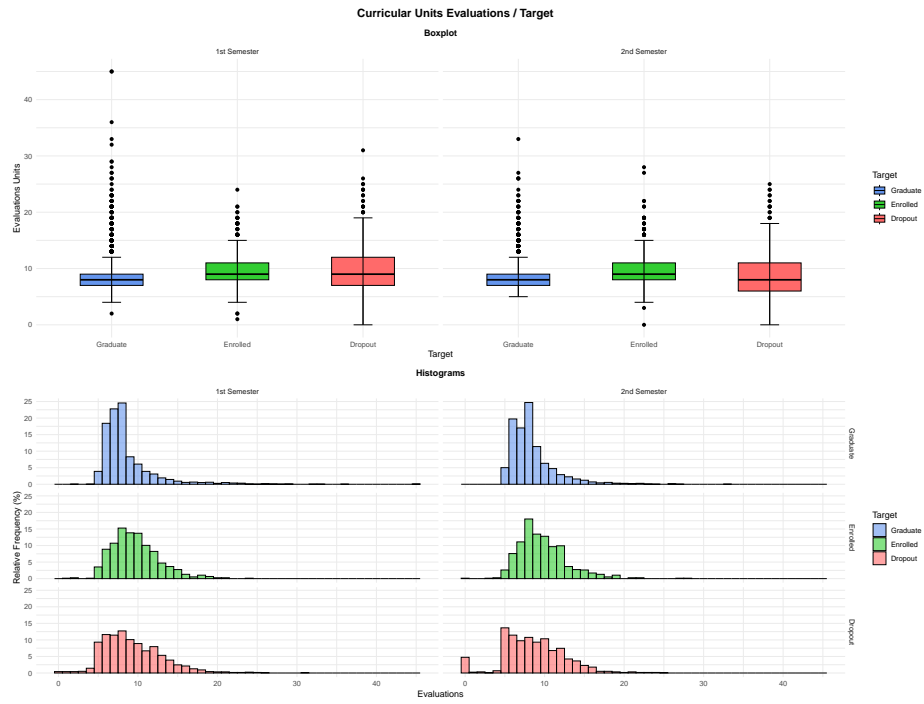


Figura 3.35: Analisi Bivariata: CU Evaluations VS Target

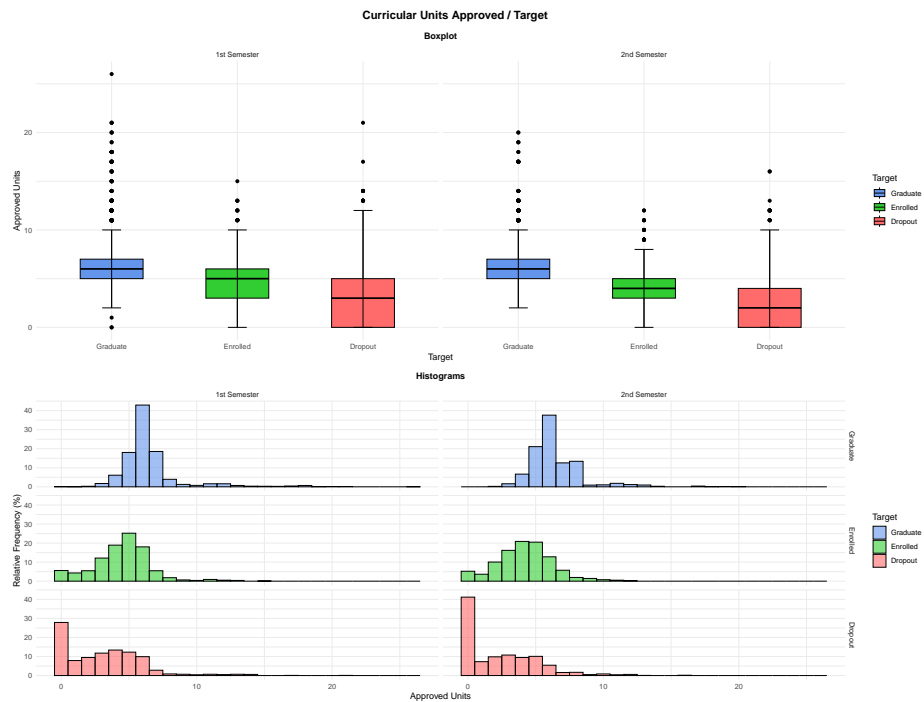


Figura 3.36: Analisi Bivariata: CU Approved VS Target

- Curricular Units Grade** (Figura 3.33): Anche l'analisi della distribuzione dei voti ottenuti nei primi due semestri mostra alcune differenze tra i tre gruppi. I **Graduate** tendono ad avere una media dei voti più alta, seguiti dagli **Enrolled**, mentre i **Dropout** presentano voti medi più bassi. Inoltre, si osserva una leggera concentrazione di studenti **Enrolled** con una media pari a 0 (Nessun esame superato), e una concentrazione molto più marcata di studenti **Dropout** con media 0, in particolare nel **secondo semestre**. Questo rafforza le osservazioni fatte in precedenza sulla distribuzione degli **Approved Curricular Units**, suggerendo che gli studenti con prestazioni accademiche molto basse fin dal primo anno siano più propensi all'abbandono universitario.

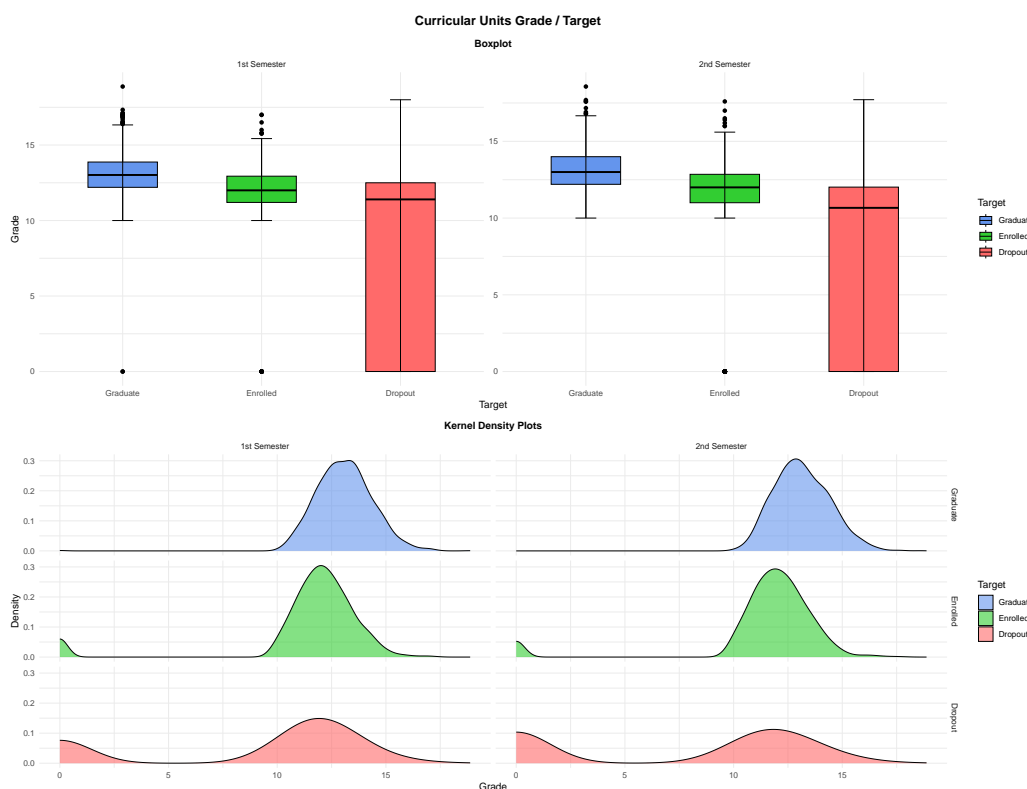


Figura 3.37: Analisi Bivariata: CU Grade VS Target

- Curricular Units Without Evaluations** (Figura 3.33): Il numero di unità curriculari senza valutazioni non mostra differenze significative tra i tre gruppi. La maggior parte degli studenti, indipendentemente dal loro esito accademico, è concentrata sul valore 0. Tuttavia, gli studenti **Enrolled** e **Dropout** presentano una percentuale leggermente più alta anche per valori superiori a 0, suggeren-

do che alcuni studenti in questi gruppi non abbiano ricevuto valutazioni per alcune delle unità curriculari frequentate.

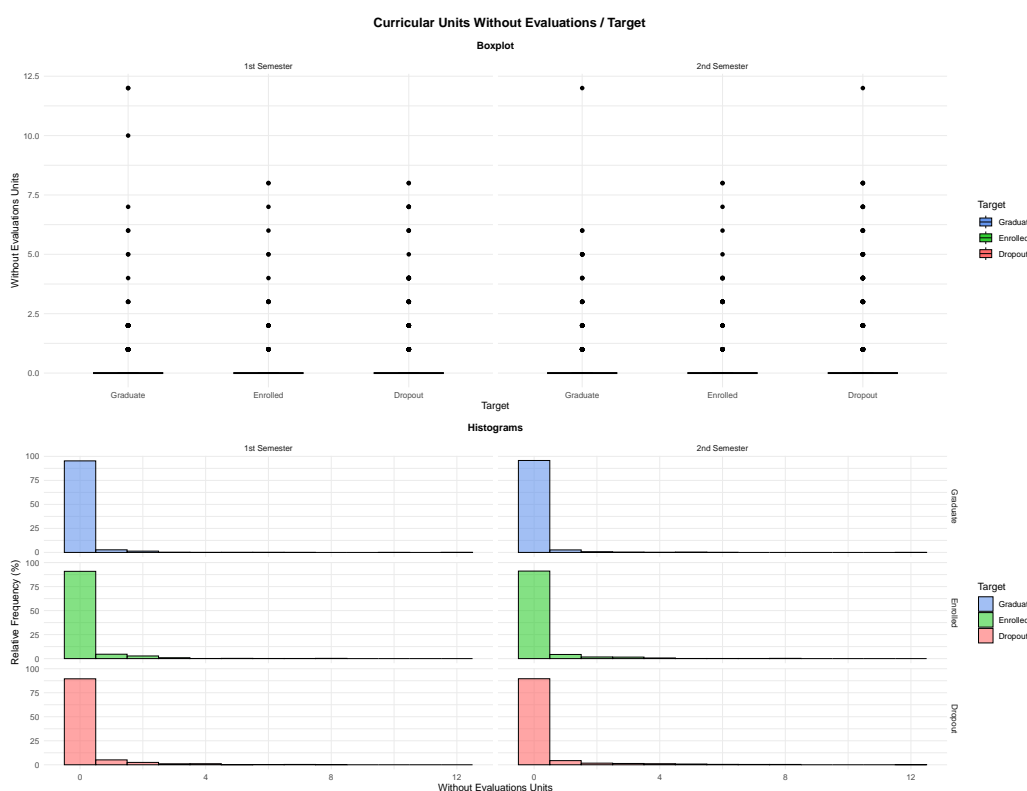


Figura 3.38: Analisi Bivariata: CU Without Evaluations VS Target

- Completed Exams Ratio** (Figura 3.39): L'analisi bivariata di questa variabile mostra differenze sostanziali tra i tre gruppi di studenti. I Boxplot e i Kernel Density Plot indicano che gli studenti **Graduate** hanno una distribuzione fortemente concentrata intorno al valore massimo (**100%**), suggerendo che la maggior parte di essi completa quasi tutte le unità curriculari a cui si iscrive. Al contrario, il gruppo **Dropout** presenta una distribuzione significativamente più ampia, con una mediana molto più bassa e un numero consistente di studenti con percentuali inferiori al 50%, evidenziando una forte associazione tra un basso tasso di esami completati e il rischio di abbandono accademico. Gli studenti **Enrolled** mostrano una distribuzione intermedia, con una mediana più elevata rispetto ai **Dropout**, ma inferiore a quella dei **Graduate**, suggerendo che una percentuale di esami non completati potrebbe essere un indicatore di un possibile ritardo nel percorso accademico. Questi risultati supportano l'idea

che il completamento di tutti o di buona parte degli esami previsti durante il primo anno sia un fattore fortemente associato al successo accademico.

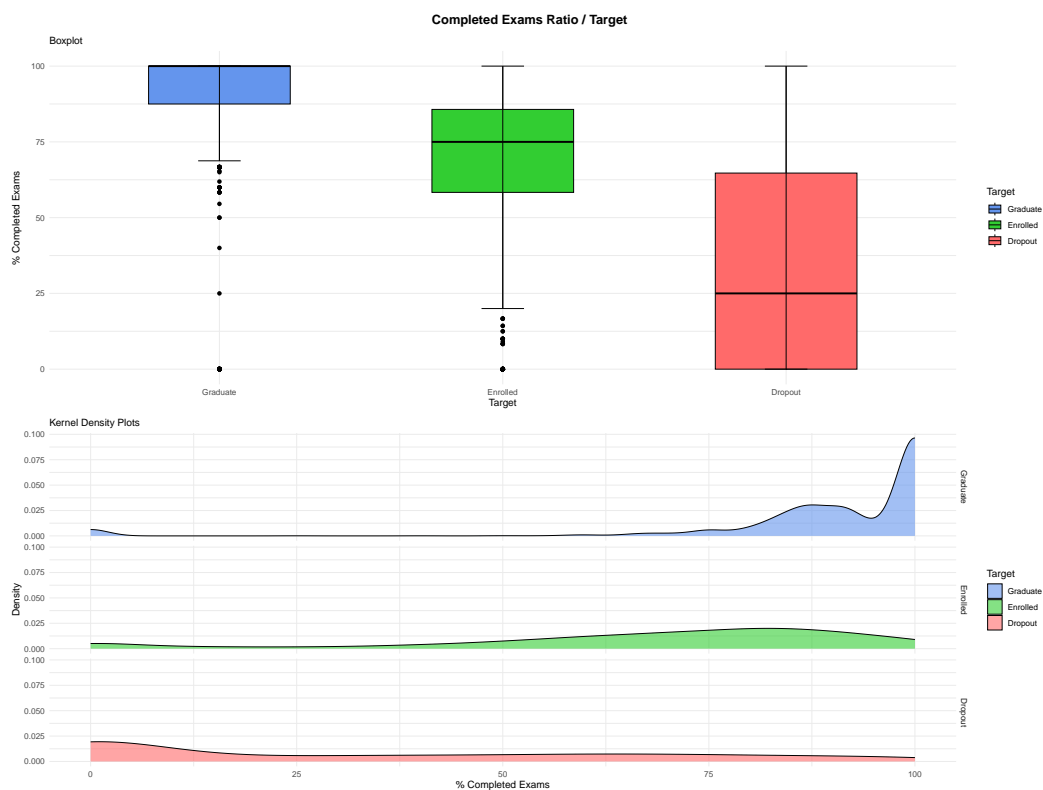


Figura 3.39: Analisi Bivariata: Completed Exams Ratio VS Target

- Passed Exams Ratio** (Figura 3.40): Anche l'analisi bivariata di questa variabile evidenzia differenze significative tra i tre gruppi di studenti. I Boxplot indicano che gli studenti **Graduate** tendono ad avere un valore mediano elevato, suggerendo che riescono a superare la maggior parte degli esami che sostengono. Al contrario, il gruppo **Dropout** presenta una distribuzione con valori tipicamente più bassi, con una mediana inferiore al 50%, suggerendo che questi studenti hanno difficoltà nel superare gli esami, il che potrebbe contribuire al loro abbandono universitario. Gli studenti **Enrolled** mostrano una distribuzione intermedia, con una mediana maggiore rispetto ai **Dropout** ma inferiore ai **Graduate**, suggerendo che la loro capacità di superare gli esami non è ottimale e potrebbe contribuire a ritardi nel completamento del percorso accademico. I **Kernel Density Plots** confermano queste osservazioni, mostrando una concentrazione maggiore di studenti **Graduate** nei valori più alti della metrica, mentre

i **Dropout** presentano una distribuzione con un picco significativo intorno a 0%, la quale indica la presenza di studenti che non hanno superato alcun esame pur avendoli sostenuti. Gli **Enrolled** si distribuiscono più uniformemente tra 25% e 75%, suggerendo una maggiore eterogeneità nel loro rendimento accademico. Questi dati potrebbero indicare che la capacità di superare gli esami con successo gli sostenuti possa essere un fattore discriminante tra chi completa con successo il percorso universitario e chi invece abbandona, con il gruppo degli **Enrolled** che si colloca in una situazione intermedia tra i due estremi.

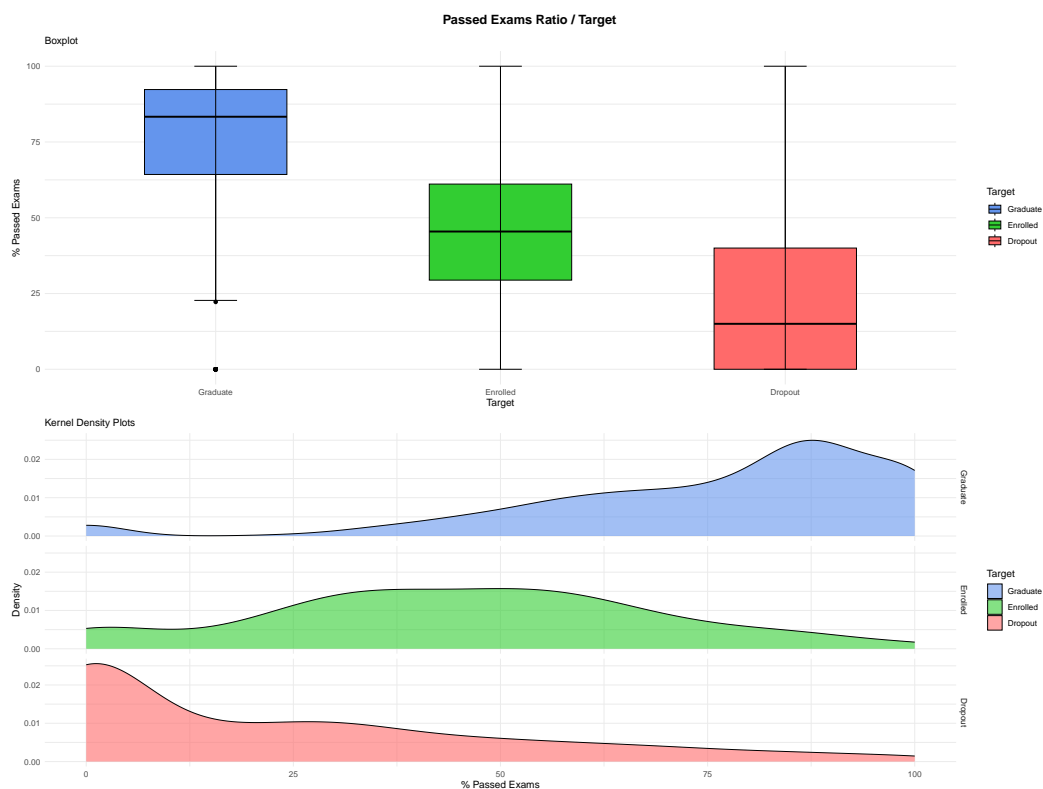


Figura 3.40: Analisi Bivariata: Passed Exams Ratio VS Target

3.8 Fattori Macro-Economici

L'analisi bivariata dei fattori macro-economici non suggerisce una forte relazione tra essi e l'esito accademico degli studenti. Di seguito vengono riportate le principali osservazioni:

- **GDP:** La variazione del **PIL** al momento dell'immatricolazione non sembra influenzare in modo significativo l'esito accademico. L'analisi dei **boxplot** mostra

distribuzioni pressoché identiche per i tre gruppi, con intervalli interquartili sovrapponibili e differenze minime nella dispersione dei dati. Tuttavia, osservando la funzione di densità nei **violin plot**, si nota che gli studenti **Graduate** presentano una leggera tendenza verso valori di GDP più elevati rispetto agli altri gruppi, con una **mediana leggermente più alta**. Inoltre, si evidenzia un **picco di densità più marcato** intorno al valore 1.7, che risulta meno pronunciato nei gruppi **Enrolled** e **Dropout**. Nonostante queste leggere differenze, l'effetto della variazione del PIL sul successo accademico appare trascurabile.

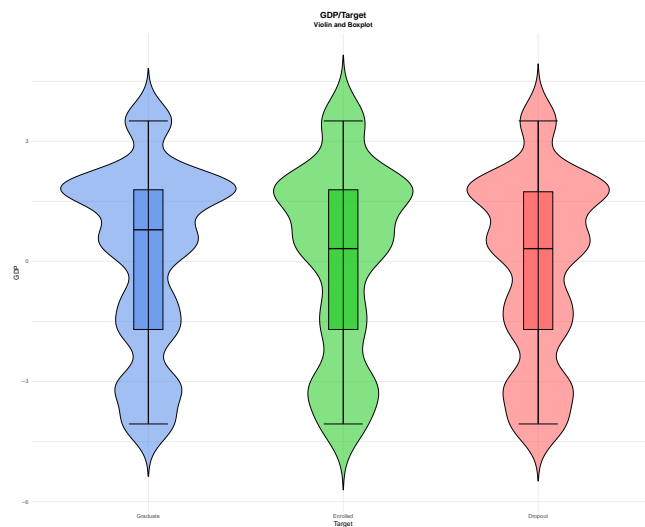


Figura 3.41: Analisi Bivariata: GDP VS Target

- **Inflation Rate:** Il tasso di inflazione al momento dell'immatricolazione non sembra avere un impatto rilevante sull'esito accademico. I **boxplot** mostrano distribuzioni molto simili tra i tre gruppi, con intervalli interquartili pressoché sovrapponibili. Tuttavia, emerge che gli studenti **Graduate** presentano una **mediana più bassa** rispetto agli altri due gruppi, suggerendo che una maggiore concentrazione di laureati si trovi in periodi caratterizzati da tassi di inflazione leggermente più bassi. Nonostante ciò, le differenze tra i gruppi rimangono contenute e non indicano un'influenza significativa dell'inflazione sugli esiti accademici.

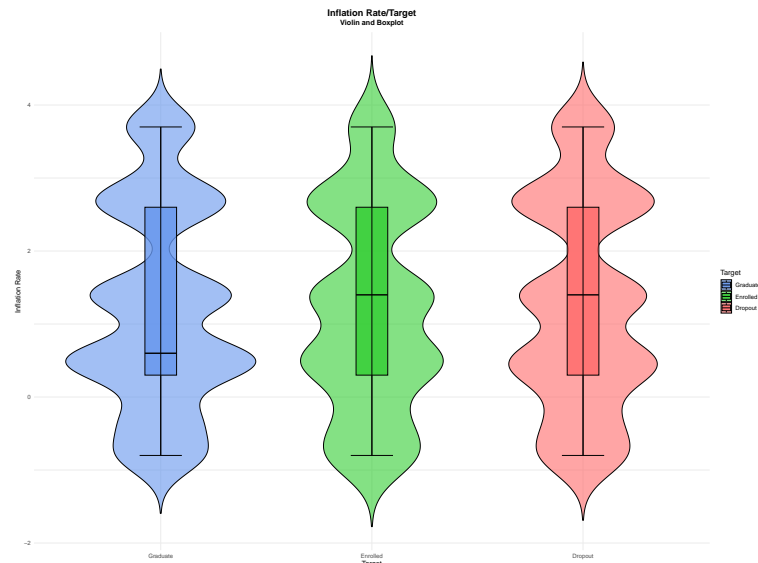


Figura 3.42: Analisi Bivariata: Inflation Rate VS Target

- Unemployment Rate:** Anche il tasso di disoccupazione al momento dell'immatricolazione non appare un fattore determinante per l'esito accademico. I **boxplot** mostrano distribuzioni simili tra i tre gruppi, con **mediane sovrapponibili** e intervalli interquartili comparabili. Tuttavia, il gruppo **Enrolled** presenta un **IQR più ristretto** rispetto agli altri due gruppi, ma nel complesso non emergono differenze tali da suggerire un effetto significativo del tasso di disoccupazione sull'esito accademico.

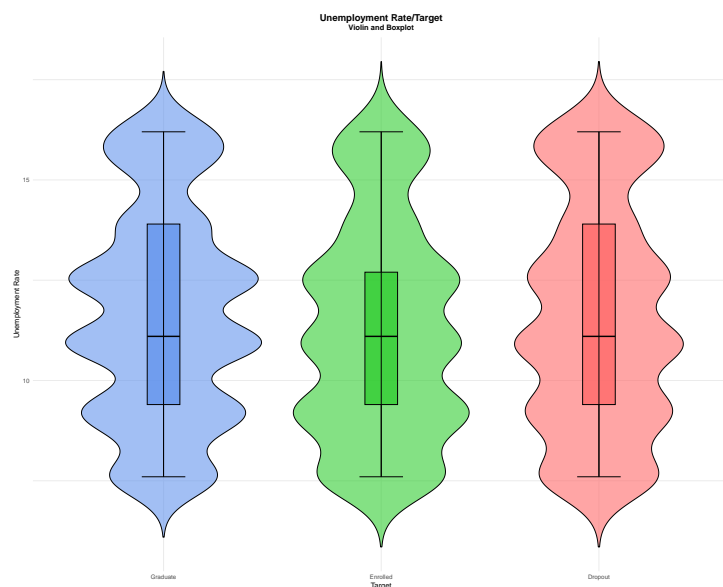


Figura 3.43: Analisi Bivariata: Unemployment Rate VS Target

3.9 Insight Analisi Bivariata

Dall'analisi bivariata condotta emergono alcune considerazioni rilevanti che meritano di essere evidenziate:

- **Correlazione tra feature numeriche:** L'analisi delle correlazioni ha evidenziato relazioni significative tra diverse variabili numeriche. Alcune correlazioni, come ad esempio quella tra le medie dei voti ottenuti nei due semestri, possono essere utilmente sfruttate nelle analisi successive. Tuttavia, è necessario prestare particolare attenzione al fenomeno della multicollinearità, che potrebbe influenzare negativamente le performance dei modelli a causa della presenza di variabili altamente correlate tra loro.
- **Fattori economici e successo accademico:** L'analisi ha rivelato un'associazione significativa tra variabili economiche degli studenti e il loro esito accademico. In particolare, fattori quali lo stato di borsista e la regolarità nei pagamenti delle tasse universitarie sono risultati significativamente associati con l'esito finale dello studente.
- **Performance accademiche e successo accademico:** Le metriche legate al rendimento accademico nel primo anno risultano fortemente associate all'esito finale dello studente. La percentuale di esami superati, i voti ottenuti e il tasso di completamento delle unità curriculari emergono come elementi determinanti nel distinguere i diversi esiti accademici degli studenti. Queste evidenze confermano l'importanza di monitorare attentamente le performance accademiche nei primi semestri per identificare studenti a rischio di abbandono.
- **Fattori macroeconomici e debole associazione:** Come previsto, le variabili macroeconomiche quali PIL, tasso di inflazione e disoccupazione hanno mostrato una debole associazione con l'esito accademico degli studenti. Tale risultato può essere attribuito alla scarsa variabilità dei dati macroeconomici disponibili. Per condurre analisi più significative in questo ambito, sarebbe opportuno disporre di dati relativi a studenti provenienti da istituti situati in differenti regioni geografiche, garantendo così una maggiore variabilità nei fattori macro-economici analizzati.

CAPITOLO 4

Test Statistici

L'analisi bivariata condotta in precedenza ha permesso di individuare possibili relazioni tra le caratteristiche degli studenti e i loro esiti accademici. Tuttavia, per determinare se tali associazioni siano statisticamente significative e non frutto di variazioni casuali, è stata necessaria l'applicazione di specifici test statistici. L'obiettivo di questi test è verificare se esista una relazione concreta tra le variabili analizzate e l'esito accademico (**Graduate, Enrolled, Dropout**), identificando i fattori che possono influenzare realmente il successo o l'abbandono degli studenti e permettendo, con un certo livello di confidenza, di estendere i risultati all'intera popolazione da cui è stato estratto il campione. In base alla natura delle variabili, sono stati applicati due approcci differenti:

- Per le **Variabili Categoricali** è stato utilizzato il **Test di Indipendenza del chi-quadro**, che permette di verificare se esiste una associazione significativa tra due variabili qualitative, andando a confrontare le distribuzioni osservate con quelle attese sotto l'ipotesi di indipendenza.
- Per le **Variabili Numeriche** è stato utilizzato il **test di Kruskal-Wallis**, un test non parametrico che rappresenta un'estensione del test di Wilcoxon per più di due gruppi. Questo test consente di confrontare le distribuzioni dei gruppi per

capire se ci sono differenze significative, però senza assumere la normalità dei dati o l'omogeneità delle varianze come l'ANOVA.

4.1 Test di Indipendenza del Chi-Quadro

Il **Test di Indipendenza del Chi-Quadro** è un test statistico che verifica l'esistenza di una relazione significativa tra due variabili categoriche. Questo test confronta le frequenze osservate e attese in una **tabella di contingenza** per determinare se la distribuzione delle categorie di una variabile dipenda dalla distribuzione delle categorie dell'altra variabile.

Ipotesi del test

Il test si basa sulle seguenti ipotesi:

- **Ipotesi Nulla (H_0):** Le due variabili sono indipendenti, ovvero la distribuzione delle categorie di una variabile non è influenzata dall'altra.
- **Ipotesi Alternativa (H_1):** Le due variabili non sono indipendenti, ovvero esiste una relazione tra di esse.

Calcolo della statistica χ^2

Per applicare il test, i dati vengono organizzati in una **tabella di contingenza**, in cui le righe rappresentano le categorie della prima variabile e le colonne rappresentano le categorie della seconda variabile. La statistica del test è calcolata come:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

dove:

- O_{ij} rappresenta la frequenza osservata nella cella corrispondente alla riga i e alla colonna j .
- E_{ij} rappresenta la frequenza attesa nella stessa cella, calcolata come:

$$E_{ij} = \frac{(\text{Totale riga } i) \times (\text{Totale colonna } j)}{\text{Totale osservazioni } N}$$

Interpretazione dei risultati

Il valore della statistica χ^2 viene confrontato con la distribuzione del chi-quadro con un numero di **gradi di libertà** pari a:

$$df = (\text{Numero di righe } r - 1) \times (\text{Numero di colonne } c - 1)$$

Se il **p-value** associato alla statistica χ^2 è inferiore alla soglia di significatività (nel nostro caso $\alpha = 0.05$), si **rifiuta l'ipotesi nulla** e si conclude che esiste una relazione significativa tra le due variabili categoriche analizzate. Al contrario, se il **p-value** è superiore alla soglia, non si può rifiutare H_0 , indicando che non ci sono prove sufficienti per dire che le due variabili siano dipendenti.

4.1.1 V di Cramér

Il test del chi-quadro permette di determinare se due variabili categoriche sono associate, ma non quantifica la forza di tale associazione. Per questo motivo, quando il test del chi-quadro indica una relazione significativa, è utile calcolare la **V di Cramér**, che misura l'intensità dell'associazione tra le due variabili. La formula per calcolare la V di Cramér è la seguente:

$$V = \sqrt{\frac{\chi^2}{N \times (k - 1)}}$$

dove:

- χ^2 è il valore del test di indipendenza del chi-quadro;
- N è il numero totale di osservazioni;
- k è il minore tra il numero di righe (r) e il numero di colonne (c) della tabella di contingenza, quindi $k = \min(r - 1, c - 1)$.

Il valore di **V** è sempre compreso tra **0 e 1**, dove **V=0** indica che non c'è nessuna associazione tra le variabili e **V=1** indica invece un'associazione perfetta. Per interpretare la forza dell'associazione, si possono utilizzare i seguenti intervalli indicativi:

V di Cramér	Forza dell'Associazione
$V \leq 0.10$	Associazione Molto Debole
$0.10 < V \leq 0.30$	Associazione Debole
$0.30 < V \leq 0.50$	Associazione Moderata
$V > 0.50$	Associazione Forte

4.2 Test di Kruskal-Wallis

Il **Test di Kruskal-Wallis** è un test non parametrico utilizzato per confrontare le distribuzioni di due o più gruppi indipendenti. Questo test rappresenta un'alternativa all'**ANOVA** nei casi in cui non sia possibile assumere la normalità della distribuzione dei dati e l'omogeneità delle varianze tra i gruppi.

Ipotesi del test

Il test si basa sulle seguenti ipotesi:

- **Ipotesi Nulla (H_0):** Le distribuzioni dei gruppi sono identiche.
- **Ipotesi Alternativa (H_1):** Almeno un gruppo presenta una distribuzione significativamente diversa dagli altri.

Calcolo della statistica del test

Il test di Kruskal-Wallis si basa sulla trasformazione dei dati in ranghi e sulla somma dei ranghi nei diversi gruppi. La statistica test è definita come:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

dove:

- N è il numero totale di osservazioni,
- k è il numero di gruppi,

- R_i è la somma dei ranghi assegnati alle osservazioni nel gruppo i ,
- n_i è la dimensione del gruppo i .

Per calcolare i ranghi:

1. Si ordinano tutte le osservazioni in ordine crescente, ignorando temporaneamente l'appartenenza ai gruppi.
2. A ciascuna osservazione viene assegnato un **rango** in base alla posizione nell'ordine crescente (la più piccola ha rango 1, la seconda ha rango 2, ecc.);
3. Se due o più osservazioni hanno lo stesso valore (*ties*), si assegna loro il **rango medio** corrispondente alle posizioni occupate. Ad esempio, se due valori uguali si trovano nelle posizioni 5 e 6, entrambi ricevono il rango medio:

$$\frac{5 + 6}{2} = 5.5.$$

4. Dopo l'assegnazione dei ranghi, si suddividono nuovamente le osservazioni nei rispettivi gruppi e si calcola la **somma dei ranghi** R_i per ciascun gruppo.

Interpretazione dei risultati

Il valore di H viene confrontato con una distribuzione **chi-quadro** con $k - 1$ **gradi di libertà**. Se il **p -value** è inferiore alla soglia di significatività (nel nostro caso $\alpha = 0.05$), si **rifiuta** H_0 , indicando differenze significative tra almeno due gruppi. Al contrario, se il **p -value** è superiore ad α , non si può rifiutare H_0 , suggerendo che non ci sono evidenze sufficienti per affermare che ci sono differenze reali tra i gruppi.

4.2.1 Epsilon-Quadrato (ϵ^2)

L'epsilon quadrato (ϵ^2) è una misura della grandezza dell'effetto per il test di Kruskal-Wallis, che quantifica la proporzione della variabilità totale nei dati spiegata dalle differenze tra i gruppi. In altre parole, indica quanto le differenze tra i gruppi influenzano la distribuzione complessiva delle osservazioni.

La formula per calcolare l'epsilon quadrato (ϵ^2) è la seguente:

$$\epsilon^2 = \frac{H - (k - 1)}{N - 1}$$

dove:

- H è il valore della statistica di Kruskal-Wallis,
- k è il numero di gruppi,
- N è il numero totale di osservazioni.

I valori di ϵ^2 variano tra **0 e 1** e vengono tipicamente interpretati seguendo questi intervalli indicativi:

Epsilon-Quadrato	Dimensione dell'Effetto
$\epsilon^2 < 0.01$	Effetto Trascurabile
$0.01 \leq \epsilon^2 < 0.06$	Effetto Debole
$0.06 \leq \epsilon^2 < 0.14$	Effetto Moderato
$\epsilon^2 \geq 0.14$	Effetto Forte

4.3 Esecuzione e Risultati dei Test

Dopo aver descritto la metodologia dei test statistici utilizzati, in questa sezione vengono presentati i risultati ottenuti dall'applicazione del **Test di Indipendenza del Chi-Quadro** per le variabili categoriche e del **Test di Kruskal-Wallis** per le variabili numeriche. L'obiettivo è di confermare ulteriormente quanto osservato con i grafici realizzati nei capitoli precedenti e verificare quali caratteristiche degli studenti mostrano un'associazione statisticamente significativa con l'esito accademico (**Graduate, Enrolled, Dropout**).

4.3.1 Analisi dei Risultati: Variabili Categoriche

I risultati del **Test di Indipendenza del Chi-Quadro**, riportati nella Tabella 7.1, confermano che i **fattori economici** sono le features maggiormente associate dell'e-

sito accademico degli studenti. In particolare, la variabile **Tuition Fees Up To Date** ($V = 0.4315$) mostra un'associazione **moderata**, indicando che il regolare pagamento delle tasse universitarie è strettamente legato al successo accademico. Analogamente, la variabile **Scholarship Holder** ($V = 0.3044$) presenta un'associazione significativa e moderata, suggerendo che il possesso di una borsa di studio può contribuire a ridurre il rischio di abbandono universitario. La variabile **Debtor** ($V = 0.2421$) mostra un'associazione più debole, ma comunque rilevante, rafforzando l'idea che le difficoltà finanziarie possano influenzare negativamente l'esito accademico.


Oltre agli aspetti finanziari, altre variabili mostrano associazioni significative con il successo accademico. La variabile **Gender** ($V = 0.2296$) suggerisce che il genere gioca un ruolo importante, confermando che gli studenti maschi tendono ad avere un tasso di dropout più elevato rispetto alle studentesse. Anche la variabile **Course** ($V = 0.2003$) ha un impatto significativo, indicando che l'area disciplinare di studio può influenzare le probabilità di laurea nei tempi previsti.

Feature	p-value	V di Cramér	Significativo
Tuition Fees Up To Date	1.47×10^{-179}	0.4315	Sì
Scholarship Holder	9.59×10^{-90}	0.3044	Sì
Debtor	4.86×10^{-57}	0.2421	Sì
Gender	2.22×10^{-51}	0.2296	Sì
Course	3.65×10^{-67}	0.2003	Sì
Application Mode	3.08×10^{-68}	0.1955	Sì
Previous Qualification	3.23×10^{-34}	0.1389	Sì
Mother's Qualification	8.31×10^{-31}	0.1377	Sì
Father's Qualification	8.31×10^{-31}	0.1377	Sì
Father's Occupation	8.71×10^{-23}	0.1300	Sì
Mother's Occupation	3.48×10^{-18}	0.1192	Sì
Displaced	2.88×10^{-13}	0.1143	Sì
Marital Status	4.48×10^{-12}	0.0817	Sì
Application Order	9.21×10^{-10}	0.0845	Sì
Attendance	5.74×10^{-07}	0.0806	Sì
Nationality	0.5273	0.0170	No
International	0.5273	0.0170	No
Educational Special Needs	0.7254	0.0120	No

Tabella 4.1: Risultati del Test di Indipendenza del Chi-Quadro

Altre variabili con un effetto significativo, ma di intensità minore, includono la **modalità di ammissione** e le qualifiche e occupazioni dei genitori. In fase di analisi bivariata, queste variabili non sembravano mostrare un'associazione evidente con l'esito accademico, ad eccezione di alcune categorie in cui il tasso di dropout risultava particolarmente elevato. Per approfondire questa relazione, è stata condotta un'analisi dei **residui standardizzati**, la quale ha confermato che le categorie che contribuiscono maggiormente alla significatività dell'associazione tra queste variabili e l'esito accademico sono: **6 - No Qualification** per la qualifica dei genitori e **9 - Student** e **10 - Other/Unknown** per l'occupazione. Questo suggerisce che gli studenti con genitori privi di titolo di studio, disoccupati o ancora studenti possano essere maggiormente esposti al rischio di abbandono universitario, anche se l'effetto complessivo di queste variabili rimane inferiore rispetto ai fattori economici.

D'altra parte, alcune variabili non risultano significativamente associate all'esito accademico. In particolare, le variabili **Nationality** e **International** mostrano un'associazione molto debole e non significativa, suggerendo che gli studenti locali e internazionali hanno probabilità simili di successo accademico. Anche la variabile **Educational Special Needs** non è significativamente associata all'esito accademico.

 **Finding RQ1.** Con un livello di significatività del 95%, i fattori economici, quali pagamento delle tasse, possesso di borse di studio e debiti economici, emergono come le variabili maggiormente associate all'esito accademico. Anche il genere e il corso di studi influenzano significativamente l'esito, mentre qualifiche e occupazioni dei genitori mostrano un impatto minore, concentrato in specifiche categorie.

4.3.2 Analisi dei Risultati: Variabili Numeriche

I risultati del **Test di Kruskal-Wallis**, riportati nella Tabella 4.2, confermano che il **Completed Exams Ratio** e il **Passed Exams Ratio** sono le variabili con la maggiore influenza sull'esito accademico, mostrando un effetto particolarmente forte ($\epsilon^2 > 0.45$). Anche il numero di **Unità Curricolari Approvate** e la **media dei voti** nei primi due semestri presentano un impatto rilevante sulla variabile target, con un effetto altrettanto elevato ($\epsilon^2 > 0.30$). Questi risultati evidenziano come il rendimento

accademico nel primo anno rappresenti un fattore determinante per il successo o l'abbandono degli studi. Inoltre, l'**età in fase di immatricolazione** mostra un effetto moderato, confermando che gli studenti immatricolati in età più avanzata tendono ad avere una probabilità maggiore di dropout rispetto ai più giovani.

Feature	p-value	Epsilon Squared	Significativo
Completed Exams Ratio	$< 2.2 \times 10^{-308}$	0.5243	Si
Passed Exams Ratio	$< 2.2 \times 10^{-308}$	0.4815	Si
Curricular Units 2nd Sem (Approved)	$< 2.2 \times 10^{-308}$	0.4297	Si
Curricular Units 1st Sem (Approved)	3.07×10^{-308}	0.3472	Si
Curricular Units 2nd Sem (Grade)	4.36×10^{-261}	0.2939	Si
Curricular Units 1st Sem (Grade)	1.23×10^{-198}	0.2234	Si
Age at Enrollment	1.67×10^{-81}	0.0912	Si
Curricular Units 2nd Sem (Enrolled)	3.51×10^{-62}	0.0694	Si
Curricular Units 1st Sem (Enrolled)	1.08×10^{-54}	0.0609	Si
Curricular Units 1st Sem (Evaluations)	3.34×10^{-36}	0.0401	Si
Curricular Units 2nd Sem (Evaluations)	5.18×10^{-33}	0.0365	Si
Admission Grade	1.11×10^{-15}	0.0169	Si
Previous Qualification (Grade)	6.48×10^{-14}	0.0149	Si
Curricular Units 2nd Sem (Without Evaluations)	4.96×10^{-12}	0.0128	Si
Curricular Units 1st Sem (Without Evaluations)	1.96×10^{-10}	0.0110	Si
Unemployment Rate	1.46×10^{-3}	0.0032	Si
GDP	1.69×10^{-3}	0.0031	Si
Curricular Units 2nd Sem (Credited)	0.0844	0.0012	No
Curricular Units 1st Sem (Credited)	0.2593	0.0007	No
Inflation Rate	0.4580	0.0004	No


Tabella 4.2: Risultati del Test di Kruskal-Wallis

Per approfondire le differenze tra i gruppi, è stato condotto un **test post-hoc di Dunn**, il quale ha evidenziato che, nella maggior parte dei casi, vi è una differenza significativa tra **Graduate** e **Dropout**, suggerendo un contrasto netto tra gli studenti di successo e coloro che abbandonano. Il gruppo **Enrolled**, invece, mostra un comportamento intermedio: in molte variabili non presenta differenze significative rispetto ai Dropout, suggerendo che questi studenti potrebbero trovarsi in una fase critica del percorso accademico, con difficoltà simili a chi ha già abbandonato, ma senza aver formalmente interrotto gli studi. Questo aspetto è cruciale per l'implementazione

di strategie di supporto, in quanto un'attenzione mirata a questi studenti potrebbe prevenire un futuro abbandono nei semestri successivi.

I **fattori macroeconomici**, invece, mostrano un effetto estremamente ridotto. In particolare, il **tasso di inflazione** non risulta nemmeno significativo, suggerendo che le condizioni economiche generali al momento dell'immatricolazione abbiano un impatto trascurabile sull'esito accademico. Anche il **voto di ammissione** e il **voto della qualifica precedente** presentano un effetto trascurabile, indicando che la performance accademica pre-universitaria non è un forte predittore del successo universitario.

Tra le variabili prive di un'associazione significativa con l'esito accademico, oltre all'**Inflation Rate**, si include il numero di **unità curriculari riconosciute** (*credited*). Questo risultato suggerisce che il riconoscimento di unità curriculari non rappresenti un vantaggio determinante nel percorso universitario e non influisca significativamente sulle probabilità di completamento del corso di studi.

 **Finding RQ2.** Con un livello di significatività del 95%, il rendimento accademico nel primo anno, misurato attraverso il numero di esami completati, la percentuale di esami superati e la media dei voti, emerge come uno dei fattori che influenzano maggiormente gli esiti accademici degli studenti.

Analisi dei Cluster

In questo capitolo, mediante tecniche di clustering, si analizzerà la possibilità di identificare gruppi di studenti con comportamenti e performance simili durante il primo anno accademico. L'obiettivo è comprendere meglio le caratteristiche di questi gruppi al fine di sviluppare strategie di supporto mirate. Nelle sezioni successive verranno descritte le operazioni di preprocessing dei dati, la metodologia di clustering adottata e i risultati ottenuti dall'analisi.

5.1 Feature Selection

Per condurre un'analisi di clustering efficace, è fondamentale selezionare le variabili più rilevanti che descrivono il comportamento accademico degli studenti nel primo anno di studi. Sulla base delle osservazioni emerse durante l'analisi esplorativa del dataset, sono state selezionate le seguenti variabili:

- **Curricular Units Enrolled (1st Semester, 2nd Semester):** Numero di unità curriculari a cui lo studente risulta iscritto in ciascun semestre.
- **Curricular Units Evaluations (1st Semester, 2nd Semester):** Numero di valutazioni ricevute in ciascun semestre.

- **Curricular Units Approved (1st Semester, 2nd Semester):** Numero di unità curriculari superate con successo in ciascun semestre.
- **Curricular Units Grade (1st Semester, 2nd Semester):** Media dei voti ottenuti nelle unità curriculari superate in ciascun semestre.
- **Completed Exams Ratio:** Percentuale di unità curriculari completate con successo rispetto al totale previsto nel primo anno accademico.
- **Passed Exams Ratio:** Percentuale di esami superati rispetto ai tentativi effettuati nel primo anno accademico.

Sebbene queste feature forniscano una rappresentazione completa del rendimento accademico nel primo anno, quattro features non sono state incluse nell'analisi di clustering a causa della loro scarsa variabilità:

- **Curricular Units Credited (1st Semester, 2nd Semester):** La maggior parte degli studenti presenta un valore pari a zero per queste feature, rendendole poco informative per la segmentazione dei gruppi.
- **Curricular Units Without Evaluations (1st Semester, 2nd Semester):** Anche in questo caso, la distribuzione dei valori mostra una forte concentrazione attorno allo zero, suggerendo che questa variabile potrebbe non apportare informazioni significative al clustering.

5.2 Preprocessing

Prima di procedere con il clustering, è stato necessario applicare una serie di operazioni di **preprocessing** per garantire che i dati fossero puliti, coerenti e pronti per l'analisi. Le principali operazioni effettuate in questa fase sono le seguenti:

- **Gestione dei Valori Mancanti:** È stata verificata la presenza di eventuali valori mancanti nelle feature selezionate andando ad analizzare nel dettaglio i valori sospetti individuati in fase di analisi univariata. Questi valori sono stati opportunamente gestiti, eliminando le righe o imputandoli.

- **Gestione degli Outlier:** Gli outlier individuati durante la fase di analisi univariata con i Boxplot, sono stati gestiti con tecniche apposite come la Winsorization per evitare che influenzassero negativamente il calcolo delle distanze e i risultati del clustering.
- **Riduzione della Dimensionalità:** Come visto in fase di analisi bivariata, molte feature tra quelle selezionate per il clustering risultano correlate tra loro, indicando una possibile ridondanza delle stesse. Pertanto, è stato valutato l'uso di tecniche di riduzione della dimensionalità come la **PCA (Principal Component Analysis)**, al fine di migliorare la separazione dei cluster e ottimizzare la rappresentazione dei dati nello spazio delle feature.

5.2.1 Gestione dei Valori Mancanti

Il primo controllo effettuato riguarda gli studenti che presentavano valori pari a 0 per tutte le feature relative al primo anno accademico. Queste informazioni potrebbero effettivamente rappresentare dati mancanti, in quanto, dalla documentazione del dataset, è emerso che le informazioni relative al primo anno accademico sono state integrate solo in un secondo momento. Di conseguenza, la mancanza di tali valori potrebbe derivare da un'assenza di registrazione piuttosto che da una performance accademica reale. Poiché questi studenti rappresentano 180 osservazioni su un totale di 4424, si è deciso semplicemente di rimuoverli.

Successivamente, è stato individuato un gruppo di 164 studenti che risultavano iscritti ad alcune unità curriculari ($CU\ Enrolled > 0$) ma che avevano tutti i valori delle altre feature pari a 0. Anche in questo caso, la mancanza di valutazioni potrebbe indicare dati effettivamente mancanti piuttosto che un comportamento accademico valido. Pertanto, anche queste osservazioni sono state rimosse dal dataset.

Un'altra osservazione riguarda un numero significativo di studenti che, pur risultando iscritti a unità curriculari e avendo tentato degli esami, non sono riusciti a superarne alcuno in entrambi i semestri o solo in uno dei due, presentando quindi una media dei voti pari a 0. In questo caso, tali dati non possono essere considerati

mancanti, bensì rappresentano una condizione accademica reale che potrebbe essere indicativa di difficoltà nello studio o scarsa partecipazione agli esami. Per questo motivo, si è deciso di non eliminare queste osservazioni, ma di trattarle opportunamente nella fase di gestione degli outlier. Infatti, come osservato nell'analisi univariata, gli studenti con una media pari a 0 risultano outlier e potrebbero influenzare negativamente il calcolo delle distanze durante il clustering, rendendo necessaria una gestione specifica nella fase successiva.

5.2.2 Gestione degli Outlier

La presenza di outlier può influenzare negativamente i risultati del clustering, alterando il calcolo delle distanze tra le osservazioni e generando gruppi poco rappresentativi. Per questo motivo, è stata adottata una strategia di gestione degli outlier che combina due approcci principali:

- **Winsorization:** per limitare l'influenza dei valori estremi mantenendoli entro soglie più ragionevoli.
- **Imputazione:** per gestire i casi di studenti con una media dei voti pari a 0 in uno dei due semestri.

La **Winsorization** è una tecnica statistica che riduce l'impatto degli outlier sostituendo i valori estremi con soglie più conservative. In questo caso, è stata applicata una **Winsorization al 5%**, sostituendo i valori inferiori al 1° percentile con il valore del 1° percentile stesso e quelli superiori al 99° percentile con il valore del 99° percentile. Questo metodo consente di preservare la distribuzione generale dei dati evitando che valori anomali distorcano il clustering.

Un caso particolare riguarda gli studenti che hanno riportato una media dei voti pari a 0 in uno dei due semestri. Poiché il voto medio rappresenta un'informazione chiave nel clustering e valori pari a 0 potrebbero influenzare negativamente il calcolo delle distanze, si è deciso di imputare tali valori sfruttando la correlazione tra la media del primo e quella del secondo semestre, individuata durante l'analisi bivariata. A tal fine, è stato sviluppato un **modello di regressione lineare semplice** (Figura 5.1),

in cui la media del secondo semestre (y) è stata stimata sulla base della media del primo semestre (x), tramite la seguente equazione:

$$y = 0.686 \cdot x + 3.986$$

Le prestazioni del modello sono state valutate attraverso due indicatori principali:

- **R-squared (R^2):** 0.4327
- **Residual Standard Error (RSE):** 0.9983

Sebbene l' R^2 non sia particolarmente elevato, questo è un risultato atteso in contesti di scienze sociali, dove il comportamento umano è influenzato da molteplici fattori non catturabili da un semplice modello di regressione. Tuttavia, il valore del **RSE** indica che le previsioni del modello si discostano mediamente di 1 punto dal voto reale, il che è più che accettabile per l'obiettivo di questa analisi, dove lo scopo principale è ottenere un valore plausibile e non troppo estremo per il clustering piuttosto che una stima altamente precisa.

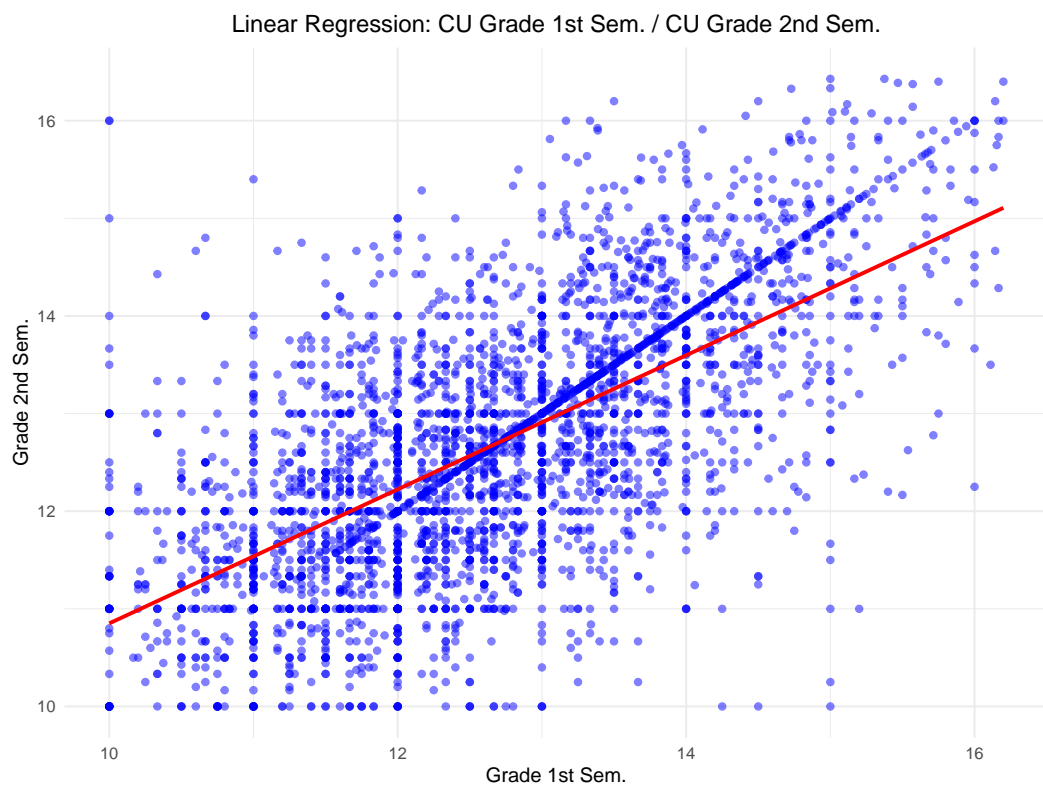


Figura 5.1: Regressione Lineare: 1st Sem. Grade / 2nd Sem. Grade

Per tentare di migliorare la precisione dei valori inputati, sono state testate anche diverse alternative alla regressione lineare semplice, tra cui:

- **Regressione Lineare Multipla:** includendo variabili indipendenti aggiuntive, come il numero di esami sostenuti e superati.
- **Gaussian Process Regression (GPR):** per modellare eventuali relazioni non lineari tra le variabili.

Tuttavia, i miglioramenti ottenuti erano trascurabili rispetto alla regressione lineare semplice e l'aumento della complessità del modello non giustificava il guadagno marginale in accuratezza. Per questo motivo, seguendo il principio del **Rasoio di Occam** (principio di parsimonia), il quale indica di scegliere la soluzione più semplice tra più soluzioni ugualmente valide di un problema, si è deciso di optare per la **Regressione Lineare Semplice** mostrata in precedenza, garantendo il giusto equilibrio tra semplicità e precisione.

5.2.3 Riduzione della Dimensionalità

Durante l'analisi bivariata è stato osservato che molte delle variabili relative al primo anno accademico presentano un'elevata correlazione tra loro. Questo fenomeno è prevedibile, in quanto variabili come il numero di esami sostenuti, superati e la media dei voti sono intrinsecamente legate tra loro: ad esempio, uno studente che risulta iscritto ad un numero elevato di unità curriculari tenderà ad avere chiaramente un numero di valutazioni maggiore e numero di unità curriculari approvate più elevato.

Sebbene la multicollinearità non sia un problema grave come in modelli di regressione o di classificazione, l'elevata correlazione tra le feature può comportare diversi problemi anche nel contesto del clustering:

- **Ridondanza nei Dati:** Variabili collineari potrebbero comportare una ripetizione dell'informazione, senza aggiungere valore al clustering, e questo può rendere i gruppi meno distinti e meno interpretabili;

- **Effetto sulla Distanza:** Se due o più variabili sono fortemente collineari, possono dominare il calcolo delle distanze, spingendo gli algoritmi a dare eccessiva importanza a certe dimensioni rispetto ad altre;
- **Curse of Dimensionality:** avere molte feature (magari ridondanti), può rendere più difficile identificare gruppi distinti e ben separati nei dati.

Per affrontare questi problemi, è stata applicata un'Analisi delle Componenti Principali (**Principal Component Analysis - PCA**), una tecnica di riduzione della dimensionalità che consente di proiettare i dati in un nuovo spazio ottenuto attraverso la combinazione lineare delle variabili originali. Le componenti principali (**Principal Components - PC**) sono costruite in modo da massimizzare la varianza dei dati, con la prima componente che cattura la massima varianza possibile, la seconda componente che cattura la massima varianza rimanente, e così via. Formalmente, la trasformazione delle variabili originali X in componenti principali Z è definita come:

$$Z = XW$$

dove:

- X è la matrice dei dati standardizzati,
- W è la matrice degli autovettori (eigenvectors), che definisce la direzione delle componenti principali,
- Z è la nuova rappresentazione dei dati nello spazio delle componenti principali.

Per determinare il numero ottimale di componenti principali da mantenere, è stato analizzato lo Scree Plot (Figura 5.2), che mostra la percentuale di varianza spiegata da ciascuna componente principale. Dall'analisi del grafico, si osserva che le prime tre componenti principali spiegano complessivamente l'85.4% della varianza totale dei dati, con la prima componente che ne cattura il 51.6%, la seconda il 26% e la terza il 7.8%. Dopo la terza componente, il contributo di ciascuna Componente Principale diminuisce drasticamente, suggerendo che il guadagno informativo derivante dall'inclusione di ulteriori componenti è limitato. Sulla base di questi risultati, per il clustering sono state selezionate le prime **3 componenti principali**, che offrono un

buon compromesso tra riduzione della dimensionalità e preservazione delle informazioni essenziali.

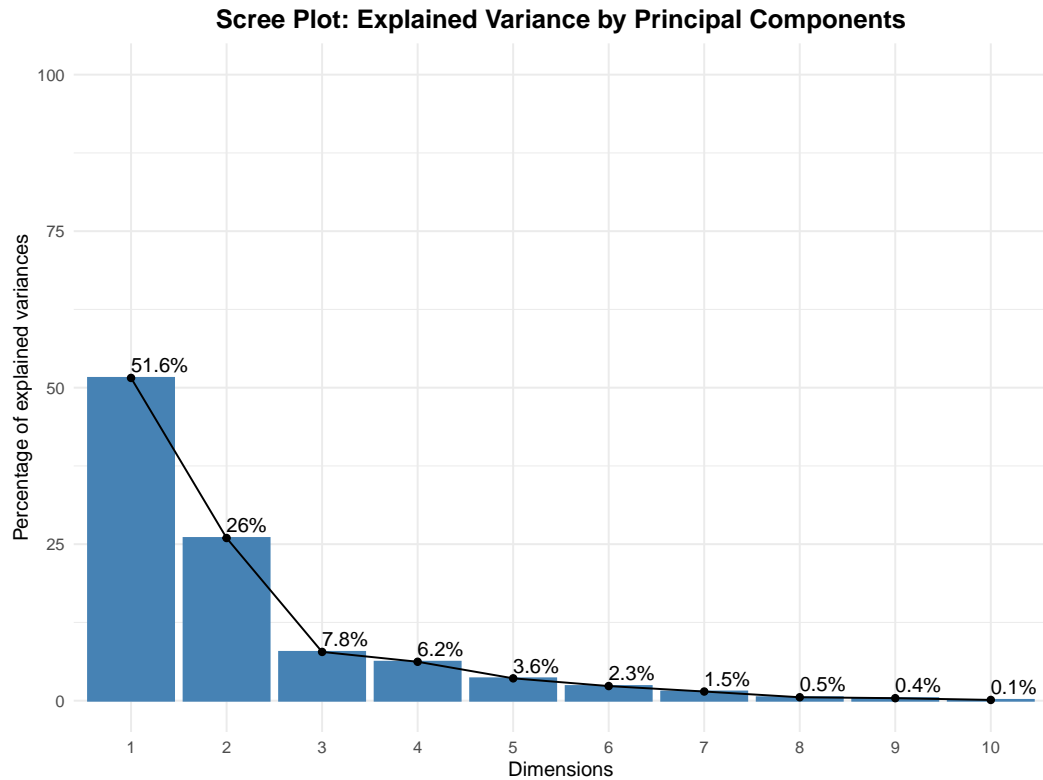


Figura 5.2: Screeplot PCA

5.3 Scelta dell'Algoritmo di Clustering

Per individuare gruppi distinti di studenti sulla base di comportamenti e performance accademiche nel primo anno di studi, sono stati valutati due approcci principali: il **clustering gerarchico** e l'**algoritmo K-means**. La scelta dell'algoritmo più adatto è stata effettuata considerando la natura del dataset e gli obiettivi dell'analisi, analizzando i vantaggi e le limitazioni di entrambi i metodi.

Il **clustering gerarchico** permette di costruire una struttura gerarchica dei dati, senza richiedere di specificare a priori il numero di cluster. Questo approccio è particolarmente utile per esplorare le relazioni tra gruppi e può risultare efficace per dataset di dimensioni ridotte. Tuttavia, presenta alcune limitazioni significative:

- Ha una complessità computazionale elevata ($O(n^2)$), rendendolo poco efficiente per dataset di grandi dimensioni.
- È sensibile alla scelta della metrica di distanza e del criterio di linkage, il che può influenzare i risultati.
- Una volta effettuata un'unione o una divisione, i cluster non possono essere riassegnati, riducendo la flessibilità dell'algoritmo.

Dall'altro lato, l'**algoritmo K-means** è un metodo partizionale che suddivide le osservazioni in **K cluster**, minimizzando la varianza intra-cluster. È computazionalmente più efficiente, particolarmente adatto a dataset di grandi dimensioni, e consente una riassegnazione dinamica delle osservazioni, garantendo maggiore flessibilità rispetto al clustering gerarchico. Tuttavia, presenta alcune criticità:

- Richiede di specificare a priori il numero di cluster K .
- È sensibile alla scelta iniziale dei centroidi, il che può portare a soluzioni subottimali.
- Assume che i cluster siano di forma sferica, ipotesi che potrebbe non essere sempre valida nei dati reali.

Dopo un'attenta valutazione, è stato deciso di adottare il **K-means clustering**, in quanto più adatto alla dimensione del dataset e agli obiettivi dell'analisi. Sebbene richieda la definizione del numero di cluster, questo aspetto può essere gestito attraverso tecniche, come l'Elbow Method e il Silhouette Score.

5.4 K-means: Scelta del Numero Ottimale di Cluster

La determinazione del numero ottimale di cluster K è un passaggio cruciale nel clustering K-means, poiché influisce direttamente sulla qualità della segmentazione e sull'interpretabilità dei risultati. Per stabilire il valore più adeguato di K , sono stati utilizzati i seguenti criteri di valutazione:

- **Indice di Silhouette:** misura la coesione intra-cluster e la separabilità inter-cluster, indicando quanto bene ogni osservazione è assegnata al proprio cluster rispetto ai cluster vicini.
- **Elbow Method:** analizza la somma delle distanze intra-cluster (*Within-Cluster Sum of Squares*, WCSS) e identifica il punto in cui la riduzione della varianza intra-cluster diventa meno significativa, segnalando un possibile valore ottimale di K .

Dopo aver eseguito il clustering con valori di K compresi tra 2 e 10, sono stati analizzati i risultati ottenuti. L'Indice di Silhouette (Figura 5.4) suggerisce che il valore ottimale sia $K = 2$, evidenziando la miglior separazione tra i cluster. Tuttavia, il metodo del gomito (Figura 5.3) mostra una riduzione significativa della WCSS per $K = 3$ o $K = 4$, suggerendo che tali valori rappresentino un buon compromesso tra coesione intra-cluster e separabilità inter-cluster.

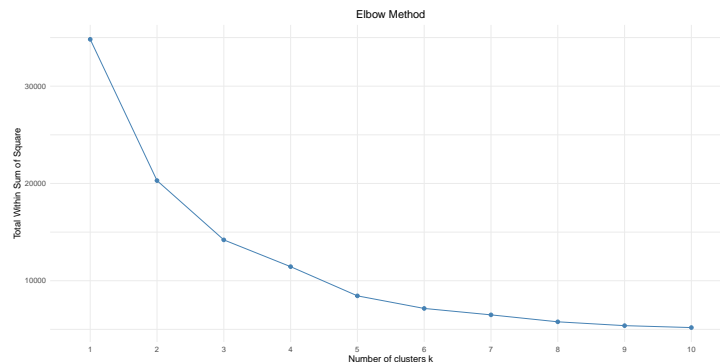


Figura 5.3: Elbow Method: WCSS per $K = 1, \dots, 10$

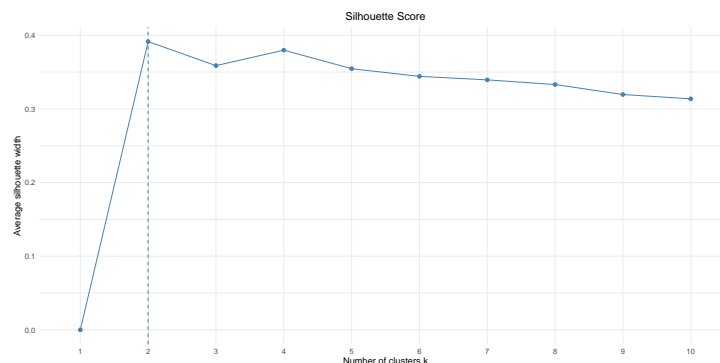


Figura 5.4: Silhouette Score medio per $K = 1, \dots, 10$

Dopo un'analisi dettagliata delle diverse configurazioni proposte, è stato deciso di adottare $K = 4$ per i seguenti motivi:

- La configurazione con $K = 4$ presentava un valore più elevato dell'**Indice di Calinski-Harabasz**, il quale valuta il rapporto tra la varianza inter-cluster e quella intra-cluster. Un valore maggiore di questo indice indica cluster compatti e ben separati.
- Pur non avendo il punteggio di Silhouette più alto in assoluto, la soluzione con $K = 4$ risultava essere la seconda migliore, con un valore solo leggermente inferiore rispetto a $K = 2$.
- L'adozione di $K = 4$ consente un'analisi più dettagliata della segmentazione degli studenti rispetto a $K = 2$, permettendo di identificare gruppi più specifici e di sviluppare strategie di supporto più mirate per ciascuna categoria.

Inoltre, per ridurre la sensibilità del modello alla scelta casuale dei centroidi iniziali e garantire una maggiore stabilità nei risultati, grazie alla libreria R utilizzata, il clustering è stato eseguito più volte con diverse configurazioni iniziali. In particolare, il modello è stato iterato 25 volte con differenti assegnazioni iniziali dei centroidi, selezionando la configurazione che minimizzava la varianza intra-cluster. Questo approccio ha permesso di mitigare il rischio di convergenza a soluzioni subottimali e migliorare l'affidabilità della segmentazione.

5.5 Analisi dei Risultati

Dopo aver eseguito il clustering con K-Means ($K=4$), abbiamo utilizzato le prime due Componenti Principali (PC1 e PC2) per rappresentare i risultati in uno spazio bidimensionale. Questa visualizzazione consente di ottenere un primo riscontro sulla separabilità dei cluster individuati e sulle loro principali caratteristiche. Dalla Figura 5.5, si osserva che i quattro cluster tendono a formare gruppi distinti, ma con alcune aree di sovrapposizione, in particolare tra il cluster viola (4) e il cluster verde (2). Questo suggerisce che alcuni studenti in questi due gruppi condividano caratteristiche simili, rendendo più difficile una netta separazione.

Un altro aspetto da considerare è la distribuzione densa e lineare di punti nel cluster verde (2), che può essere attribuita a diversi fattori interconnessi. La PCA, combinando più variabili nelle componenti principali, ha probabilmente proiettato gruppi di osservazioni con valori simili lungo direzioni specifiche, generando strutture allungate. Inoltre, la presenza di variabili numeriche discrete, con un numero limitato di valori distinti, ha portato molte osservazioni ad avere esattamente gli stessi valori su più dimensioni, facilitando la formazione di raggruppamenti naturali. A questo si aggiunge l'effetto della Winsorization, applicata per attenuare l'influenza degli outlier, che ha ulteriormente ridotto la variabilità nei valori estremi, contribuendo alla concentrazione di punti in determinate aree dello spazio trasformato. Data la combinazione di questi elementi, non è stato attribuito eccessivo peso a questa struttura. Inoltre, pattern simili erano già presenti nei dati originali, in particolare nella distribuzione dei voti, confermando che questo comportamento non è un artefatto del clustering, ma una caratteristica intrinseca del dataset.

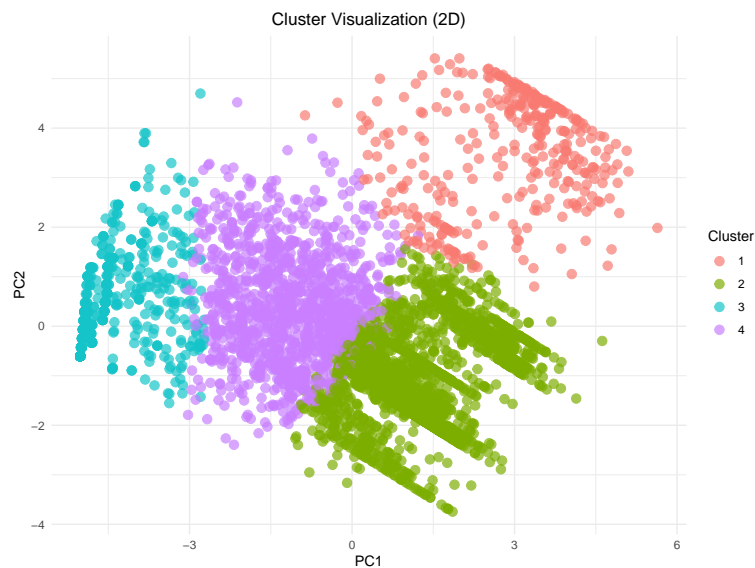


Figura 5.5: Visualizzazione 2D dei Cluster

Questa rappresentazione offre solo una panoramica iniziale della segmentazione ottenuta. Nelle Figure 5.6 e 5.7 è possibile osservare in dettaglio come le diverse variabili siano distribuite all'interno dei cluster. Nelle sezioni successive, verranno analizzati più approfonditamente i singoli cluster per comprendere meglio le loro caratteristiche distintive.

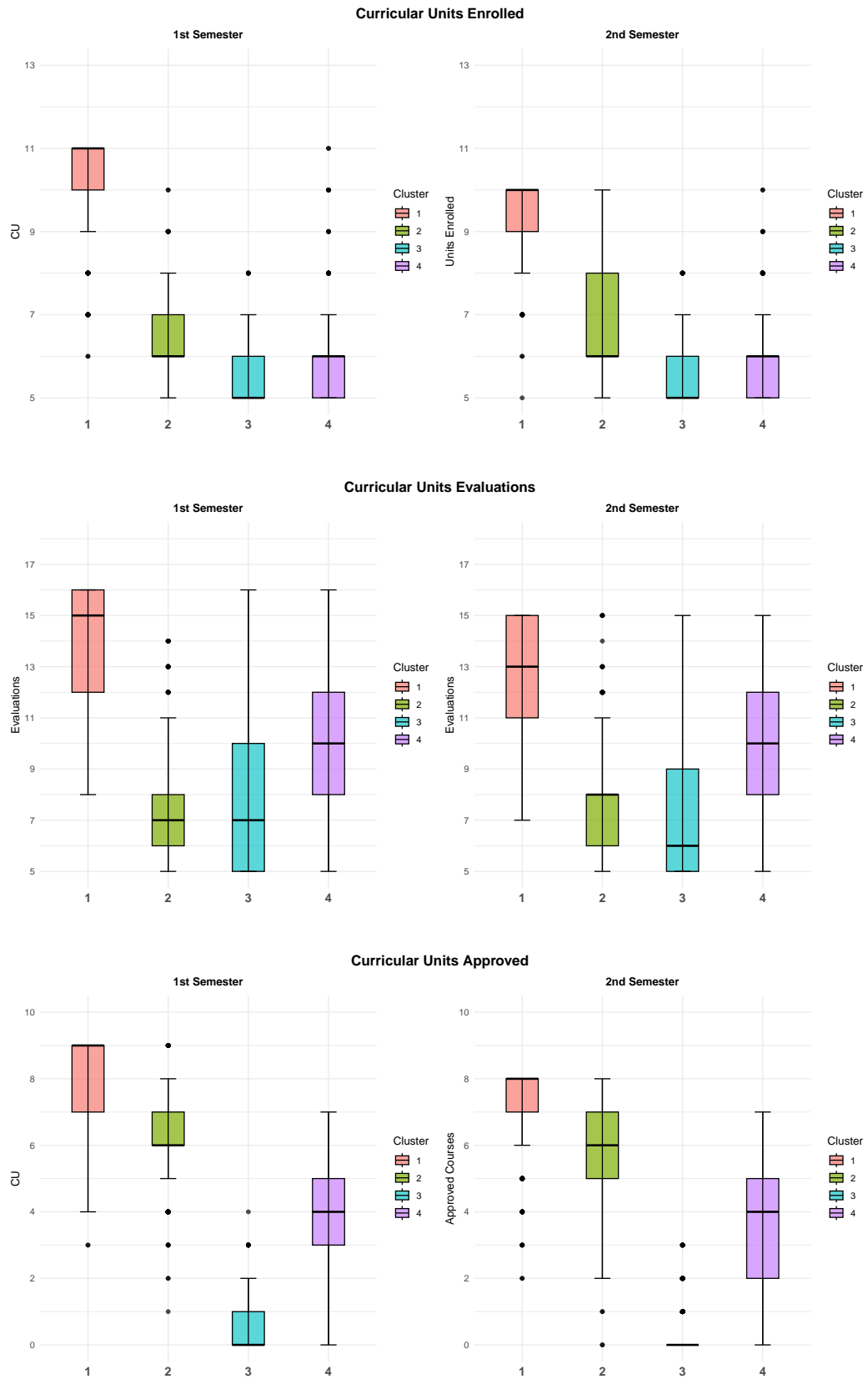


Figura 5.6: Clustering: Distribuzione delle features nei Cluster (Parte 1)

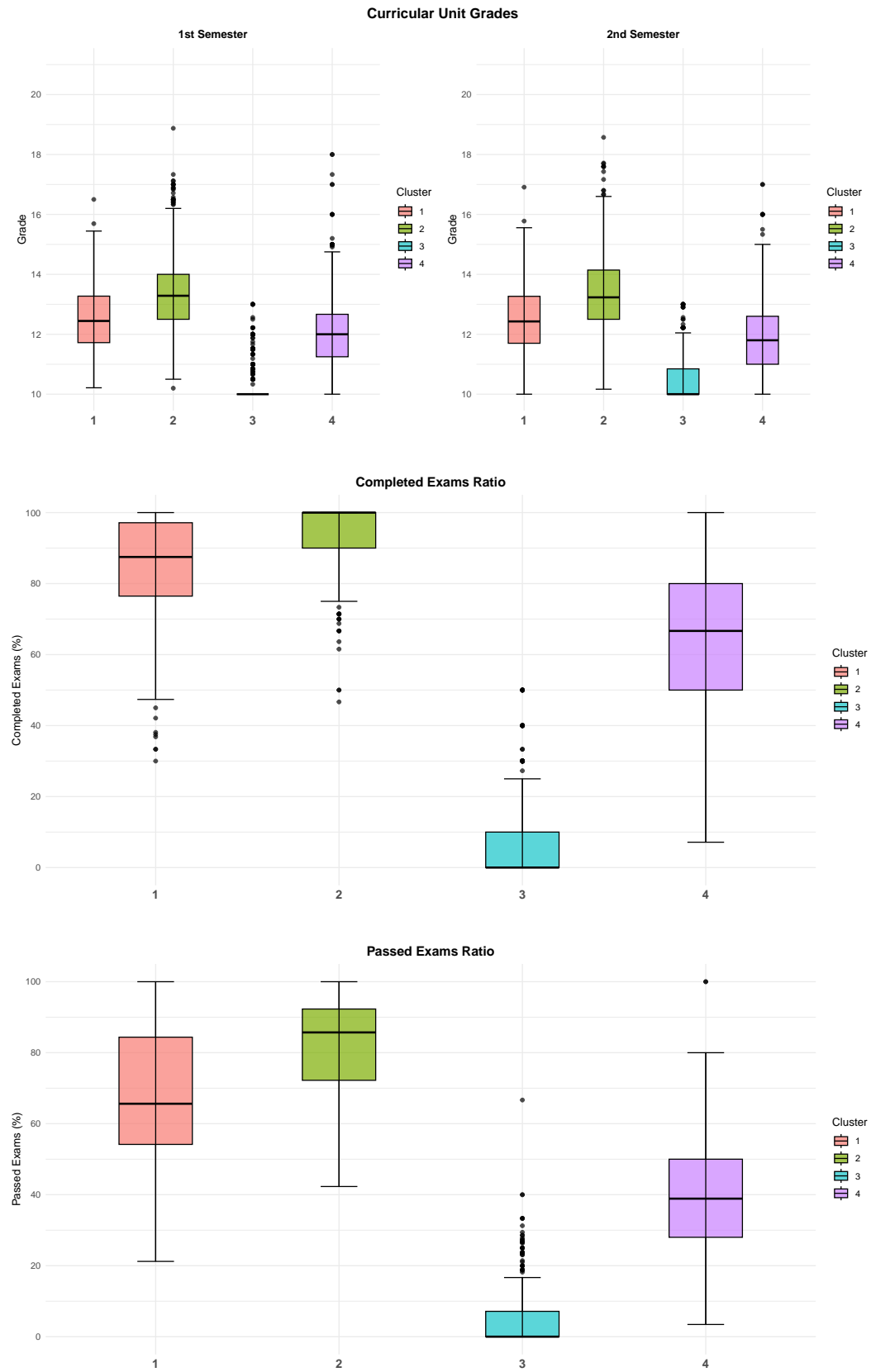


Figura 5.7: Clustering: Distribuzione delle features nei Cluster (Parte 2)

5.5.1 Cluster 2 (Verde)

Il Cluster 2 si distingue per le sue **ottime performance accademiche** rispetto agli altri gruppi. Gli studenti appartenenti a questo cluster mostrano un elevato impegno e una forte capacità di completare con successo gli esami previsti:

- Il **Completed Exams Ratio** indica che la maggior parte degli studenti ha completato quasi il 100% delle unità curriculari previste. Analogamente, il **Passed Exams Ratio**, anch'esso molto elevato, suggerisce che questi studenti non solo completano gli esami, ma li superano con successo nella maggior parte dei casi.
- Le **medie dei voti** risultano le più alte rispetto agli altri cluster, con una dispersione relativamente contenuta. Questo evidenzia un rendimento accademico eccellente e costante tra gli studenti appartenenti a questo gruppo.
- Gli studenti di questo cluster tendono a iscriversi a un numero **moderato** di unità curriculari per semestre. Inoltre, il numero di esami superati e il numero di valutazioni ricevute (esami tentati) sono coerenti con le unità curriculari frequentate, confermando che questi studenti riescono a completare la quasi totalità delle CU previste e generalmente al primo tentativo di esame.

5.5.2 Cluster 1 (Rosso)

Il **Cluster 1** rappresenta un gruppo di studenti anche essi molto performanti, con risultati accademici positivi, sebbene leggermente inferiori rispetto al **Cluster 2**:

- Il **Completed Exams Ratio** e il **Passed Exams Ratio** sono entrambi elevati, seppur leggermente inferiori rispetto a quelli del Cluster 2. Questo suggerisce che anche questi studenti riescono a completare con successo la maggior parte degli esami previsti, sebbene con una performance lievemente più variabile.
- Il **numero di unità curriculari frequentate (CU Enrolled)** risulta significativamente più alto rispetto agli altri cluster. Questo potrebbe essere dovuto alla struttura dei corsi frequentati dagli studenti appartenenti a questo gruppo, i quali potrebbero prevedere un piano di studi con un numero maggiore di unità curriculari per semestre.

- In linea con il numero di CU Enrolled, anche il **numero di esami superati (CU Approved)** e il **numero di valutazioni ricevute (Evaluations)** sono più elevati rispetto agli altri gruppi. Questo conferma che questi studenti affrontano un carico accademico maggiore, riuscendo comunque a mantenere un rendimento positivo.

Complessivamente, il Cluster 1 sembra essere molto simile al Cluster 2 in termini di performance accademiche, ma con la differenza sostanziale di frequentare corsi che prevedono un numero maggiore di unità curriculari. Questo aspetto potrebbe influire sul rendimento accademico complessivo degli studenti, rendendo più complesso mantenere livelli di successo pari a quelli del Cluster 2.

5.5.3 Cluster 4 (Viola)

Il Cluster 4 rappresenta un gruppo di studenti con performance accademiche inter-medie rispetto agli altri gruppi:

- Il **Completed Exams Ratio** e il **Passed Exams Ratio** mostrano una maggiore variabilità rispetto ai Cluster 1 e 2, indicando la presenza sia di studenti con buone performance che di studenti che incontrano difficoltà nel completare gli esami.
- Le **medie dei voti** sono inferiori rispetto ai Cluster 1 e 2, suggerendo una prestazione accademica generalmente più debole.
- Il numero di **Curricular Units Enrolled** è simile a quello del Cluster 3, ma gli studenti di questo gruppo mostrano un tasso di **CU Approved** più elevato rispetto a quest'ultimo. Tuttavia, il numero di **Evaluations** è mediamente superiore al numero di CU Enrolled, suggerendo che questi studenti devono tentare più volte un esame prima di riuscire a superarlo.

Questo cluster potrebbe rappresentare un gruppo di studenti con difficoltà moderate: alcuni riescono a progredire nel percorso di studi con successo, mentre altri incontrano ostacoli più significativi, pur evitando di raggiungere livelli critici di insuccesso come quelli osservati nel Cluster 3.

5.5.4 Cluster 3 (Azzurro)

Il Cluster 3 rappresenta il gruppo di studenti con le peggiori performance accademiche tra quelli individuati nel clustering:

- Il **Completed Exams Ratio** è il più basso tra tutti i cluster, indicando che la maggior parte degli studenti in questo gruppo ha completato solo una piccola percentuale o nessuna delle unità curriculari previste. Anche il **Passed Exams Ratio** mostra valori estremamente bassi, suggerendo che questi studenti non solo completano pochi esami, ma incontrano anche grandi difficoltà nel superarli.
- Le **medie dei voti** sono significativamente più basse rispetto agli altri cluster, confermando un rendimento accademico complessivamente debole.
- Il numero di **Curricular Units Enrolled** è simile a quello del Cluster 4, ma il numero di **CU Approved** è nettamente inferiore. Questo suggerisce che molti studenti di questo cluster pur iscrivendosi ai corsi, faticano a superarli.

Questo cluster potrebbe rappresentare un gruppo di studenti particolarmente a rischio di abbandono universitario, a causa delle difficoltà nel superare gli esami e nel completare il percorso accademico.

5.5.5 Considerazioni Finali

L'analisi del clustering ha permesso di individuare quattro gruppi distinti di studenti, caratterizzati da diversi comportamenti e performance accademiche. Per rispondere alla RQ2, è stata sfruttata anche la variabile target (*Graduate, Enrolled, Dropout*) al fine di comprendere meglio la distribuzione degli esiti accademici nei cluster e identificare strategie di supporto mirate per ciascun gruppo. La Figura 5.8 mostra come i cluster mostrino differenze sostanziali anche rispetto agli esiti accademici degli studenti. Di seguito, vengono sintetizzate le caratteristiche principali di ciascun cluster e le possibili strategie di intervento:

- **Cluster 2 (Verde):** Composto prevalentemente da studenti che riescono a laurearsi nei tempi previsti (*Graduate*), questo gruppo si distingue per le performan-

ce accademiche eccellenti. Sebbene non necessitino di interventi di supporto, potrebbero beneficiare di programmi di eccellenza e percorsi di approfondimento per incentivare ulteriormente il loro rendimento.

- **Cluster 1 (Rosso):** Anche qui la maggior parte degli studenti completa con successo gli studi, ma con una quota di *Dropout* ed *Enrolled* leggermente più elevata rispetto al Cluster 2. Questo cluster potrebbe beneficiare degli stessi percorsi di approfondimento del Cluster 2. Tuttavia, il carico di studio risulta tendenzialmente più elevato in questo gruppo, pertanto una rivalutazione dell'organizzazione dei corsi e l'implementazione di strategie di supporto nella gestione del workload accademico potrebbero contribuire a ridurre ulteriormente il rischio di abbandono.
- **Cluster 4 (Viola):** Caratterizzato da una distribuzione piuttosto bilanciata tra *Graduate*, *Enrolled* e *Dropout*, questo gruppo mostra un rendimento accademico intermedio. L'elevato numero di tentativi d'esame suggerisce difficoltà nel superare le unità curriculari. Questi studenti potrebbero trarre beneficio da un maggiore supporto didattico, tutoraggio e programmi di accompagnamento per rafforzare la loro progressione accademica.
- **Cluster 3 (Azzurro):** Qui si concentra la più alta percentuale di studenti *Dropout*, in linea con le prestazioni accademiche molto basse. Strategie di intervento mirate, come tutoraggi personalizzati, percorsi di recupero ed indagini più approfondite su altri fattori, potrebbero essere fondamentali per ridurre il rischio di abbandono universitario e migliorare il coinvolgimento accademico.

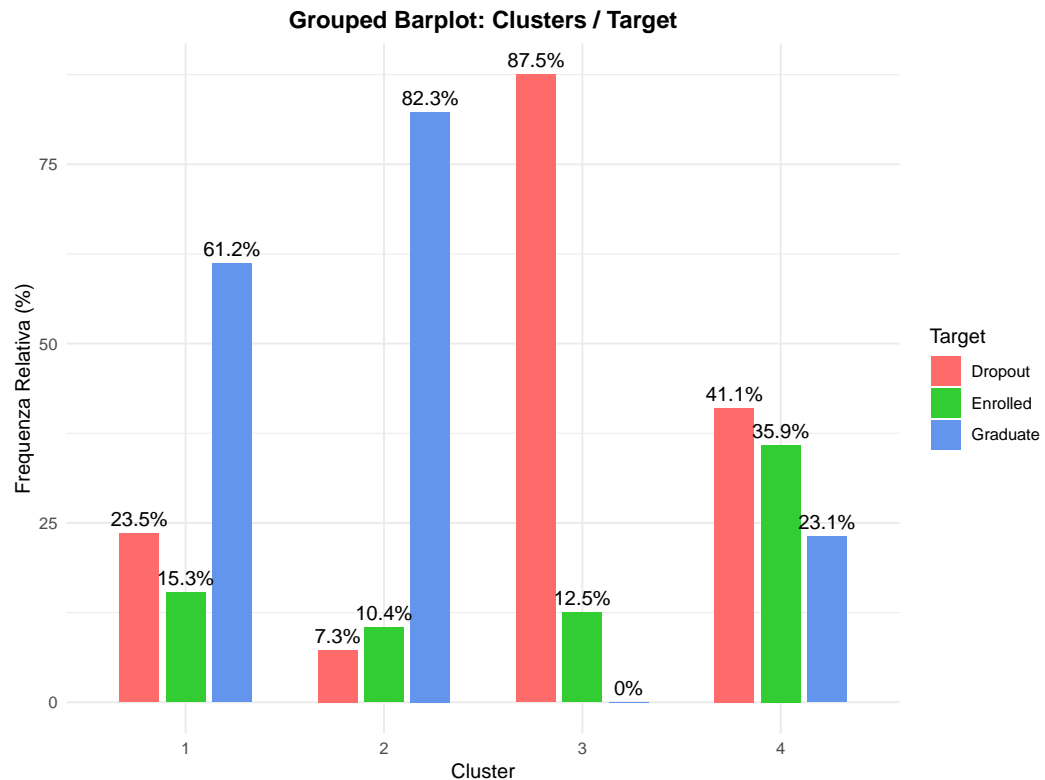


Figura 5.8: Clustering: Distribuzione delle Variabile Target nei Cluster

🔍 Finding RQ2. L'analisi dei cluster ha permesso di individuare 4 gruppi di studenti con caratteristiche e necessità differenti, fornendo una base solida per l'adozione di strategie educative mirate. L'implementazione di interventi personalizzati potrebbe non solo migliorare il rendimento accademico, ma anche ridurre significativamente il rischio di *Dropout*, favorendo una maggiore continuità e successo nel percorso di studi.

Stima delle Medie dei Voti e Differenze di Genere

In questo capitolo verranno presentate le operazioni effettuate per rispondere alla RQ 3. Verrà quindi illustrato il processo di stima intervallare per la media dei voti ottenuti dagli studenti e per la differenza tra la media dei voti degli studenti maschi e quella delle studentesse femmine. L'obiettivo è quantificare l'incertezza associata alle medie campionarie attraverso la costruzione di intervalli di confidenza al 99%, che forniscono un range di valori entro cui ci si aspetta che cada la media reale della popolazione con un alto livello di confidenza. Tra le feature di cui verranno calcolate le medie abbiamo:

- **Curricular Units 1st Sem. Grade e Curricular Units 2nd Sem. Grade:** Medie dei voti ottenuti dagli studenti durante il primo e il secondo semestre. Per semplicità verranno accorpati in una votazione unica per l'intero anno accademico;
- **Admission Grade:** Voto ottenuto dagli studenti durante il test di ammissione;
- **Previous Qualification Grade:** Voto ottenuto dagli studenti per la qualifica precedente.

Nei paragrafi successivi verranno descritti nel dettaglio i metodi utilizzati per la costruzione degli intervalli di confidenza e i risultati ottenuti.

6.1 Stima Intervallare della Media dei Voti

Per stimare la media dei voti ottenuti dagli studenti, è stato costruito un **intervallo di confidenza al 99%**. Tuttavia, poiché la distribuzione della popolazione dei voti non è nota, è stato necessario fare affidamento sul **Teorema del Limite Centrale (TLC)** per garantire che la distribuzione della media campionaria segua un andamento normale.

Il **Teorema del Limite Centrale** afferma che, indipendentemente dalla distribuzione della popolazione di partenza, la media campionaria segue approssimativamente una distribuzione normale per campioni sufficientemente grandi. In particolare, se X_1, X_2, \dots, X_n sono variabili casuali indipendenti e identicamente distribuite con media μ e varianza finita σ^2 , allora la media campionaria:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

per n sufficientemente grande segue una distribuzione normale:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Questo significa che, standardizzando la media campionaria, possiamo approssimare la sua distribuzione con la **normale standard Z**:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Nel nostro caso, il numero di osservazioni è pari a circa 4000, un valore sufficientemente grande affinché il TLC sia applicabile. Pertanto, possiamo costruire un intervallo di confidenza utilizzando il metodo pivotale. Poiché la varianza della popolazione non è nota, la deviazione standard campionaria (s) è stata utilizzata come stima di σ . L'intervallo di confidenza al livello di confidenza $1 - \alpha$ è dato dalla formula:

$$IC = \left[\bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \quad \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

dove:

- \bar{X} è la media campionaria dei voti,

- s è la deviazione standard campionaria,
- n è la numerosità campionaria,
- $z_{\alpha/2}$ è il quantile della distribuzione normale standard associato al livello di confidenza desiderato (nel nostro caso $\alpha = 0.01$).

Questo intervallo rappresenta il range entro cui ci si aspetta che cada la media reale della popolazione con una probabilità del 99%.

6.2 Stima Intervallare della Differenza tra le Medie

Per verificare l'esistenza di una differenza statisticamente significativa tra le medie dei voti ottenuti dagli studenti maschi e femmine, è stato calcolato un intervallo di confidenza al 99% per tale differenza.

Siano X_M e X_F le variabili casuali che rappresentano rispettivamente i voti degli studenti maschi e delle studentesse femmine. Essendo le due popolazioni indipendenti e disponendo di campioni sufficientemente ampi, possiamo sfruttare il Teorema del Limite Centrale per approssimare la distribuzione della differenza tra le medie campionarie con una normale standard. L'intervallo di confidenza al 99% è dunque calcolato tramite il metodo pivotale, considerando che le varianze delle popolazioni non sono note e quindi stimate dalle deviazioni standard campionarie:

$$IC = \left[(\bar{X}_M - \bar{X}_F) - z_{\frac{\alpha}{2}} \cdot SE_{diff}, \quad (\bar{X}_M - \bar{X}_F) + z_{\frac{\alpha}{2}} \cdot SE_{diff} \right]$$

dove:

- \bar{X}_M e \bar{X}_F rappresentano le medie campionarie dei voti rispettivamente degli studenti maschi e femmine;
- s_M e s_F sono le deviazioni standard campionarie per ciascun gruppo;
- n_M e n_F indicano le numerosità dei rispettivi campioni;

- SE_{diff} è l'errore standard della differenza tra le medie, calcolato come:

$$SE_{diff} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}}$$

- $z_{\alpha/2}$ è il quantile della distribuzione normale standard associato al livello di confidenza desiderato (nel nostro caso $\alpha = 0.01$).

L'intervallo ottenuto indica l'intervallo di valori plausibili per la vera differenza tra le medie dei due gruppi con un livello di confidenza del 99%. Se l'intervallo non contiene il valore 0, è possibile concludere che esiste una differenza statisticamente significativa tra le medie dei due gruppi analizzati.

6.3 1st Year Grade

In questa sezione verranno mostrati i risultati ottenuti nella stima intervallare delle medie dei voti degli studenti durante il primo anno accademico e della stima nella differenza delle medie di tali voti tra Maschi e Femmine.

6.3.1 Stima Intervallare della Media

L'intervallo di confidenza per la media dei voti del primo anno accademico è stato calcolato come segue:

- **Media campionaria:**

$$\bar{X} = 12.318$$

- **Deviazione standard campionaria:**

$$s = 2.053$$

- **Numero di osservazioni nel campione:**

$$n = 3748$$

- **Valore critico della normale standard con $\alpha = 0.01$:**

$$z_{\alpha/2} = 2.576$$

- **Errore standard della media campionaria:**

$$SE = \frac{s}{\sqrt{n}} = \frac{2.053}{\sqrt{3748}} = 0.034$$

- **Intervallo di confidenza al 99%:**

$$IC = [\bar{X} - z_{\alpha/2} \cdot SE, \quad \bar{X} + z_{\alpha/2} \cdot SE]$$

- **Risultato:**

$$IC = [12.231, \quad 12.404]$$

🔍 Finding RQ3. Con una probabilità del 99%, la vera media dei voti degli studenti durante il primo anno accademico si trova nell'intervallo [12.231, 12.404]

6.3.2 Confronto Maschi e Femmine

L'intervallo di confidenza al 99% per la differenza tra la media dei voti del primo anno accademico degli studenti maschi e quella delle studentesse femmine è stato calcolato come segue:

- **Media campionaria maschi e femmine:**

$$\bar{X}_M = 11.889, \quad \bar{X}_F = 12.517$$

- **Deviazione standard campionaria maschi e femmine:**

$$s_M = 2.377, \quad s_F = 1.850$$

- **Numero di osservazioni nei due campioni:**

$$n_M = 1192, \quad n_F = 2556$$

- **Valore critico della normale standard con $\alpha = 0.01$:**

$$z_{\alpha/2} = 2.576$$

- **Errore standard della differenza tra le medie:**

$$SE_{\Delta} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{2.377^2}{1192} + \frac{1.850^2}{2556}} = 0.077$$

- Intervallo di confidenza al 99% per la differenza tra le medie:

$$IC = [(\bar{X}_M - \bar{X}_F) - (z_{\alpha/2} \cdot SE_{\Delta}), (\bar{X}_M - \bar{X}_F) + (z_{\alpha/2} \cdot SE_{\Delta})]$$

- Risultato:

$$IC = [-0.828, -0.427]$$

🔍 Finding RQ3. L'intervallo è completamente negativo, quindi con un livello di confidenza del 99%, la media dei voti del primo anno accademico delle studentesse femmine è statisticamente superiore a quella degli studenti maschi

6.4 Admission Grade

In questa sezione verranno mostrati i risultati ottenuti nella stima intervallare delle medie dei voti di ammissione degli studenti e della stima nella differenza delle medie di tali voti tra Maschi e Femmine.

6.4.1 Stima Intervallare della Media

L'intervallo di confidenza per la media dei voti di ammissione è stato calcolato come segue:

- Media campionaria:

$$\bar{X} = 126.978$$

- Deviazione standard campionaria:

$$s = 14.482$$

- Numero di osservazioni nel campione:

$$n = 4424$$

- Valore critico della normale standard con $\alpha = 0.01$:

$$z_{\alpha/2} = 2.576$$

- **Errore standard della media campionaria:**


$$SE = \frac{s}{\sqrt{n}} = \frac{14.482}{\sqrt{4424}} = 0.217$$

- **Intervallo di confidenza al 99%:**

$$IC = [\bar{X} - z_{\alpha/2} \cdot SE, \quad \bar{X} + z_{\alpha/2} \cdot SE]$$

- **Risultato:**

$$IC = [126.417, \quad 127.539]$$

 **Finding RQ3.** Con una probabilità del 99%, la vera media dei voti di ammissione degli studenti si trova nell'intervallo [126.417, 127.539]

6.4.2 Confronto Maschi e Femmine

L'intervallo di confidenza al 99% per la differenza tra la media dei voti di ammissione degli studenti maschi e quella delle studentesse femmine è stato calcolato come segue:

- **Media campionaria maschi e femmine:**

$$\bar{X}_M = 127.141, \quad \bar{X}_F = 126.889$$

- **Deviazione standard campionaria maschi e femmine:**

$$s_M = 15.161, \quad s_F = 14.101$$

- **Numero di osservazioni nei due campioni:**

$$n_M = 1556, \quad n_F = 2868$$

- **Valore critico della normale standard con $\alpha = 0.01$:**

$$z_{\alpha/2} = 2.576$$

- **Errore standard della differenza tra le medie:**

$$SE_{\Delta} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{15.161^2}{1556} + \frac{14.101^2}{2868}} = 0.465$$

- **Intervallo di confidenza al 99% per la differenza tra le medie:**

$$IC = [(\bar{X}_M - \bar{X}_F) - (z_{\alpha/2} \cdot SE_{\Delta}), (\bar{X}_M - \bar{X}_F) + (z_{\alpha/2} \cdot SE_{\Delta})]$$

- **Risultato:**

$$IC = [-0.948, 1.452]$$

🔍 Finding RQ3. L'intervallo comprende lo 0, quindi non abbiamo prove sufficienti per affermare che esista una differenza tra le medie dei voti di ammissione degli studenti maschi e delle studentesse femmine

6.5 Previous Qualification Grade

In questa sezione verranno mostrati i risultati ottenuti nella stima intervallare delle medie dei voti per la qualifica precedente degli studenti e della stima nella differenza delle medie di tali voti tra Maschi e Femmine.

6.5.1 Stima Intervallare della Media

L'intervallo di confidenza per la media dei voti della qualifica precedente è stato calcolato come segue:

- **Media campionaria:**

$$\bar{X} = 132.613$$

- **Deviazione standard campionaria:**

$$s = 13.188$$

- **Numero di osservazioni nel campione:**

$$n = 4424$$

- **Valore critico della normale standard con $\alpha = 0.01$:**

$$z_{\alpha/2} = 2.576$$

- **Errore standard della media campionaria:**


$$SE = \frac{s}{\sqrt{n}} = \frac{13.188}{\sqrt{4424}} = 0.198$$

- **Intervallo di confidenza al 99%:**

$$IC = [\bar{X} - z_{\alpha/2} \cdot SE, \quad \bar{X} + z_{\alpha/2} \cdot SE]$$

- **Risultato:**

$$IC = [132.103, \quad 133.124]$$

 **Finding RQ3.** Con una probabilità del 99%, la vera media dei voti della qualifica precedente degli studenti si trova nell'intervallo [132.103, 133.124]

6.5.2 Confronto Maschi e Femmine

L'intervallo di confidenza al 99% per la differenza tra la media dei voti per la qualifica precedente degli studenti maschi e quella delle studentesse femmine è stato calcolato come segue:

- **Media campionaria maschi e femmine:**

$$\bar{X}_M = 131.756, \quad \bar{X}_F = 133.078$$

- **Deviazione standard campionaria maschi e femmine:**

$$s_M = 13.404, \quad s_F = 13.048$$

- **Numero di osservazioni nei due campioni:**

$$n_M = 1556, \quad n_F = 2868$$

- **Valore critico della normale standard con $\alpha = 0.01$:**

$$z_{\alpha/2} = 2.576$$

- **Errore standard della differenza tra le medie:**

$$SE_{\Delta} = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{13.404^2}{1556} + \frac{13.048^2}{2868}} = 0.418$$

- **Intervallo di confidenza al 99% per la differenza tra le medie:**

$$IC = [(\bar{X}_M - \bar{X}_F) - (z_{\alpha/2} \cdot SE_{\Delta}), (\bar{X}_M - \bar{X}_F) + (z_{\alpha/2} \cdot SE_{\Delta})]$$

- **Risultato:**

$$IC = [-2.399, -0.245]$$

🔍 Finding RQ3. L'intervallo è completamente negativo, quindi con un livello di confidenza del 99% la media dei voti per la qualifica precedente delle studentesse femmine è statisticamente superiore a quella degli studenti maschi.

CAPITOLO 7

Analisi Dataset Sintetico

In questo capitolo verrà analizzato il processo di creazione di un dataset sintetico simile a quello analizzato nei capitoli precedenti utilizzando un LLM (*ChatGPT 4o*). Dopo una panoramica sulle tecniche impiegate e sui criteri adottati per ottimizzare la generazione dei dati, verranno presentati i risultati ottenuti, evidenziando le principali differenze tra il dataset sintetico e quello reale. In particolare, saranno discussi i limiti e le criticità riscontrate nella capacità del LLM di generare dati realistici e coerenti con quelli originali. È importante sottolineare che, nell'analisi delle differenze tra i due dataset, non sono state illustrate le discrepanze per ogni singola variabile, bensì sono stati selezionati alcuni esempi rappresentativi. Questi permettono comunque di generalizzare la maggior parte delle osservazioni anche alle altre feature presenti nel dataset.

7.1 Strategia di Prompting

La generazione del dataset sintetico è stata realizzata tramite un **Large Language Model (LLM)**, nello specifico *ChatGPT 4o*, attraverso un prompt strutturato e dettagliato. Data la complessità del dataset originale e l'elevato numero di feature presenti, è stata adottata una strategia basata su un file **JSON**, che fornisce una descrizione

completa della struttura del dataset e delle caratteristiche di ciascuna variabile. L'utilizzo di questa metodologia ha permesso di trasmettere al modello vincoli precisi sulla generazione dei dati, garantendo maggiore coerenza con il dataset originale. La struttura del file JSON verrà analizzata più nel dettaglio successivamente.

7.1.1 Struttura del Prompt

Il prompt è stato suddiviso nelle seguenti sezioni principali:

- **Context:** sezione introduttiva che descrive il contesto e gli obiettivi della generazione del dataset sintetico.
- **Provided Input:** descrive gli input forniti al modello, in particolare il file **JSON**, che specifica la struttura del dataset, e il **Few-Shot Learning Sample**, un insieme di righe di esempio estratte dal dataset reale, utili per garantire coerenza nello stile e nella distribuzione dei dati generati.
- **Rules for Data Interpretation and Generation:** definisce le regole per la generazione dei dati, tra cui il rispetto della struttura e dei vincoli specificati nel JSON, il formato dei dati (numerico continuo/discreto, categorico), le distribuzioni da adottare in base alle specifiche fornite e l'introduzione controllata di outlier per rendere il dataset più realistico.
- **Expected Output:** specifica i requisiti formali del dataset generato, come il formato in output (CSV senza testo esplicativo aggiuntivo), la dimensione esatta del dataset (4000 righe) e il rispetto della sequenza delle colonne secondo l'ordine specificato nel JSON.
- **Generation Strategy:** fornisce indicazioni sulla procedura che il modello deve seguire per generare il dataset, specificando l'ordine dei passaggi, dall'analisi del JSON e del Few-Shot Learning fino alla generazione e all'esportazione del dataset nel formato richiesto.
- **Important Notes:** contiene informazioni aggiuntive per garantire che il modello segua con precisione le istruzioni, evitando deviazioni dalle specifiche fornite.

Il prompt è stato redatto in lingua inglese per garantire una maggiore compatibilità con il modello di linguaggio utilizzato. Infatti, gli LLM come *ChatGPT 4o* tendono a fornire risposte più coerenti e aderenti alle istruzioni quando i comandi sono formulati in inglese, data la predominanza di testi in questa lingua nei dati di addestramento.

7.1.2 Struttura del File JSON

Il file JSON utilizzato per la generazione del dataset sintetico è stato progettato per fornire una descrizione dettagliata e strutturata delle feature presenti nel dataset originale. La sua organizzazione segue una logica gerarchica, suddivisa in due sezioni principali: una descrizione generale del dataset e un elenco dettagliato delle feature.

Descrizione Generale del Dataset

La sezione `dataset_description` fornisce informazioni contestuali sul dataset, includendo il `name`, ovvero il nome del dataset, e la `description`, che ne descrive l'obiettivo principale, ovvero la predizione dell'abbandono o del successo accademico come problema di classificazione a tre categorie.

Elenco delle Feature

La sezione `features` costituisce il nucleo del file JSON e definisce le caratteristiche di ciascuna variabile. Ogni feature è descritta dai seguenti elementi fondamentali:

- `name`: il nome della feature, coerente con la nomenclatura originale.
- `type`: la tipologia della variabile, classificata come *categorical* (valori discreti predefiniti), *quantitative discrete* (valori numerici interi) o *quantitative continuous* (valori numerici continui).
- `description`: una breve spiegazione del significato della variabile.
- `value_map` (per variabili categoriche): associazione tra valori numerici e categorie testuali.
- `range` (per variabili quantitative): intervallo di validità dei valori generati.

- `unit` (se applicabile): unità di misura della variabile.
- `observations`: osservazioni sulla distribuzione della variabile, su eventuali vincoli, correlazioni e associazioni con altre features.

L'adozione di questa struttura ha permesso di definire in modo rigoroso i vincoli sui dati sintetici, garantendo maggiore coerenza con il dataset reale.

Esempio: Mother's Qualification

La feature `Mother's Qualification` è una variabile categorica con sei modalità, definite nel campo `value_map`. Le osservazioni (`observations`) forniscono informazioni aggiuntive sulla distribuzione della variabile, indicando che i livelli di istruzione più frequenti sono quelli di base o secondari. Inoltre, viene segnalato che la categoria `No Qualification` (6) è associata a un tasso di abbandono (*Dropout Rate*) più elevato rispetto alle altre categorie:

```
{
  "name": "Mothers_qualification",
  "type": "categorical",
  "description": "Mother's educational qualification",
  "value_map": {
    "1": "Higher Education",
    "2": "Secondary Education",
    "3": "Basic Education (3° ciclo)",
    "4": "Basic Education (2° ciclo)",
    "5": "Basic Education (1° ciclo)",
    "6": "No Qualification"
  },
  "observations": [
    "The most frequent qualifications are basic/secondary education levels.",
    "The category 6-No Qualification has an higher Dropout Rate than the others."
  ]
}
```

Esempio: Admission Grade

La feature `Admission Grade` è una variabile numerica continua, con valori compresi tra 95 e 200, come specificato nel campo `range`. Nelle osservazioni (`observations`) è indicata una correlazione positiva moderata con la variabile `Previous Qualification (Grade)`. Tuttavia, non è stata fornita alcuna indicazione esplicita sulla distribuzione

da seguire. Pertanto, come stabilito nel prompt, il LLM genererà i dati basandosi sul significato della variabile e sugli esempi forniti tramite il *Few-Shot Learning*:

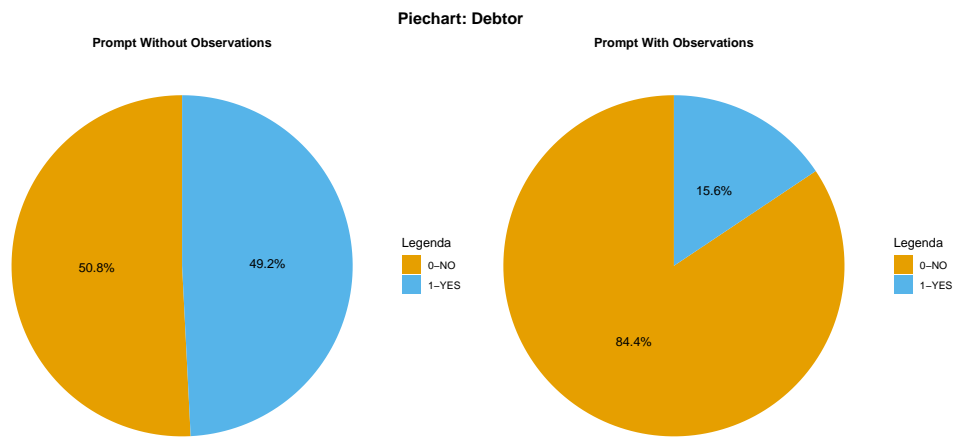
```
{
  "name": "Admission_grade",
  "type": "quantitative continuous",
  "description": "Admission grade to the course",
  "range": [
    { "min": 95, "max": 200 }
  ],
  "observations": [
    "Admission_grade has a moderate positive correlation with Previous_qualification_(grade).",
  ]
}
```

7.2 Distribuzione delle Feature Categoriche

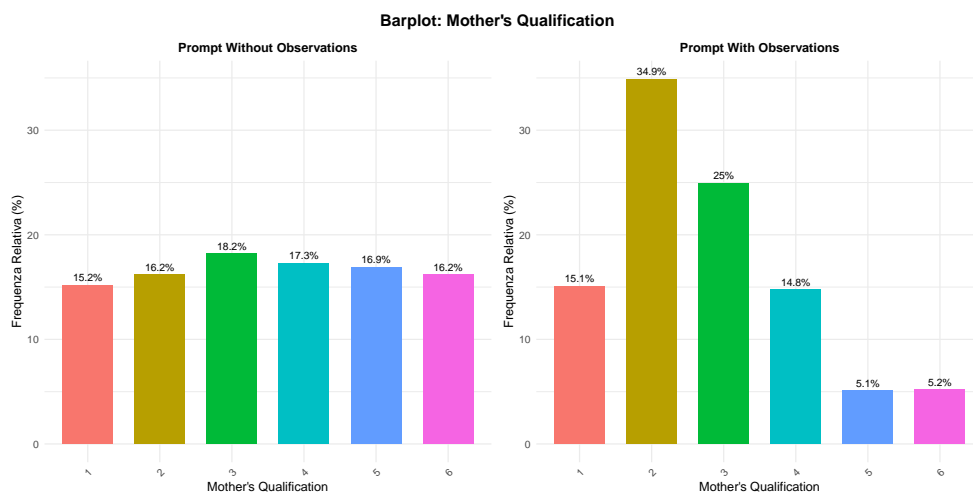
Nella generazione delle variabili categoriche, il LLM ha mostrato comportamenti diversi a seconda delle informazioni fornite nel prompt. In assenza di indicazioni esplicite sulla distribuzione attesa, il modello tende a generare le categorie in modo approssimativamente bilanciato. Tuttavia, specificando nel prompt la distribuzione reale e fornendo esempi di righe dal dataset originale tramite *few-shot learning*, è stato possibile ottenere risultati più aderenti alla realtà.

Esempio: Debtor

La Figura 7.1 mostra la distribuzione della variabile binaria **Debtor**, che indica la presenza di debiti finanziari tra gli studenti. Nel dataset sintetico generato senza informazioni aggiuntive, la distribuzione risulta bilanciata (~50%-50%). Tuttavia, fornendo indicazioni esplicite, il modello ha generato una distribuzione più coerente con quella reale (~84% No, ~16% Yes).

**Figura 7.1:** Dataset Sintetico: analisi "Debtor"**Esempio: Mother's Qualification**

La Figura 7.2 mostra la distribuzione della variabile **Mother's Qualification**, caratterizzata da un numero maggiore di categorie. Nel dataset sintetico generato senza specifiche aggiuntive, le categorie risultano distribuite in modo uniforme. Tuttavia, anche in questo caso, includendo informazioni aggiuntive, il modello ha generato una distribuzione più coerente con quella reale, con una maggiore concentrazione nelle categorie di istruzione di base e secondaria.

**Figura 7.2:** Dataset Sintetico: analisi "Mother's Qualification"

Considerazioni Finali

Gli esempi riportati sono rappresentativi di un comportamento generale osservato per altre variabili categoriche nel dataset. In sintesi, fornendo un contesto chiaro e dettagliato, il LLM è in grado di generare distribuzioni più coerenti con la realtà. Tuttavia, la gestione di variabili con molte modalità può risultare complessa, aumentando il rischio di generare sbilanciamenti poco naturali. Questo aspetto è particolarmente critico in dataset con numerose categorie, dove fornire indicazioni dettagliate può non essere sufficiente a ottenere una distribuzione realmente fedele ai dati originali.

🔍 Finding RQ4. Per le variabili categoriche il LLM tende a generare distribuzioni più realistiche e fedeli ai dati originali quando riceve un contesto chiaro e dettagliato

7.3 Distribuzione delle Feature Numeriche

A differenza delle variabili categoriche, per le feature numeriche è stata lasciata piena libertà al LLM nella generazione dei dati, consentendogli di distribuirli autonomamente in base al significato della variabile, agli esempi forniti nel prompt e ad eventuali vincoli specificati nel file JSON. Questo approccio ha permesso di valutare il realismo delle distribuzioni generate, la loro coerenza con i dati reali e la loro riconducibilità a distribuzioni statistiche note.

Esempio: Age at Enrollment

La distribuzione dell'età di iscrizione, mostrata in Figura 7.3, presenta un andamento simile a quello reale, con una maggiore concentrazione di studenti più giovani e una coda lunga verso destra fino ai 70 anni. Tuttavia, emergono alcune discrepanze. Nel dataset sintetico, il valore più frequente è il limite inferiore di 17 anni, mentre nei dati reali la moda era 18 anni, con pochissime osservazioni a 17. Inoltre, la diminuzione della frequenza con l'aumentare dell'età è meno brusca rispetto alla realtà, con frequenze più elevate nella fascia 25-30 anni. Complessivamente, sebbene la distribuzione sia plausibile, queste differenze rendono il dataset sintetico leggermente meno realistico.

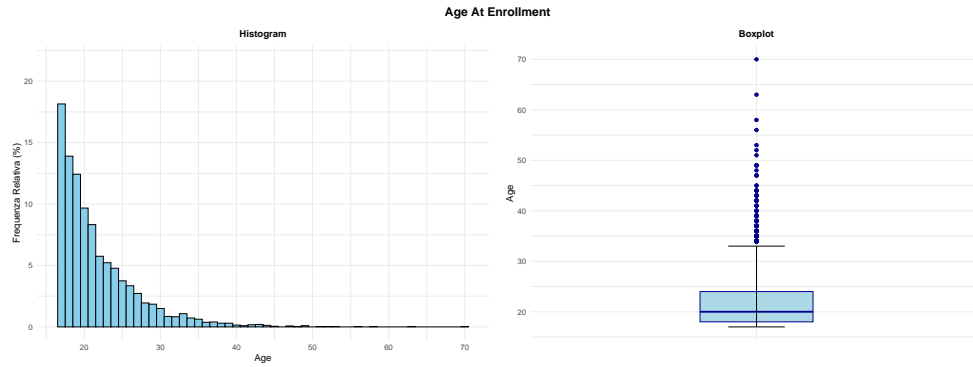


Figura 7.3: Dataset Sintetico: analisi "Age At Enrollment"

Esempio: Curricular Units Enrolled

La distribuzione del numero di unità curriculari a cui gli studenti risultano iscritti, illustrata in Figura 7.4, mostra un andamento coerente con il range tipico di questa variabile. Nel primo semestre, il modello ha prodotto una distribuzione uniforme discreta, mentre nel secondo semestre ha seguito una distribuzione simmetrica intorno a una moda centrale, suggerendo una tendenza binomiale. Purtroppo, rispetto ai dati reali, il dataset sintetico manca di outliers, ossia studenti con un numero significativamente superiore o inferiore di unità curriculari. Inoltre, la variabilità risulta ridotta rispetto ai dati reali, nei quali il numero di unità curriculari frequentate può dipendere da molteplici fattori, come il corso di laurea o la strategia individuale degli studenti. Questo evidenzia come il modello abbia prodotto una distribuzione plausibile, ma semplificata rispetto alla complessità reale.

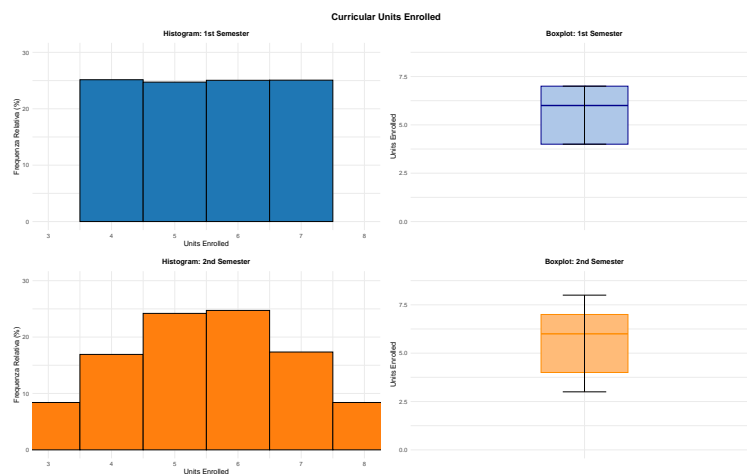


Figura 7.4: Dataset Sintetico: analisi "Curricular Units Enrolled"

Esempio: Admission Grade

La Figura 7.5 mostra la distribuzione dei voti di ammissione. Il modello ha generato dati seguendo una distribuzione apparentemente normale con una coda sinistra troncata, probabilmente per rispettare i vincoli definiti nel JSON. Sebbene la struttura generale sia coerente con i dati reali, emergono alcune differenze. Lo sbilanciamento verso votazioni inferiori, tipico del dataset reale, è meno marcato nel dataset sintetico. Inoltre, la distribuzione appare più regolare e priva delle irregolarità osservate nei dati originali, probabilmente dovute agli arrotondamenti nei sistemi di valutazione. Anche la presenza di outliers risulta ridotta rispetto alla realtà, rendendo la distribuzione sintetica più pulita ma meno rappresentativa della realtà.

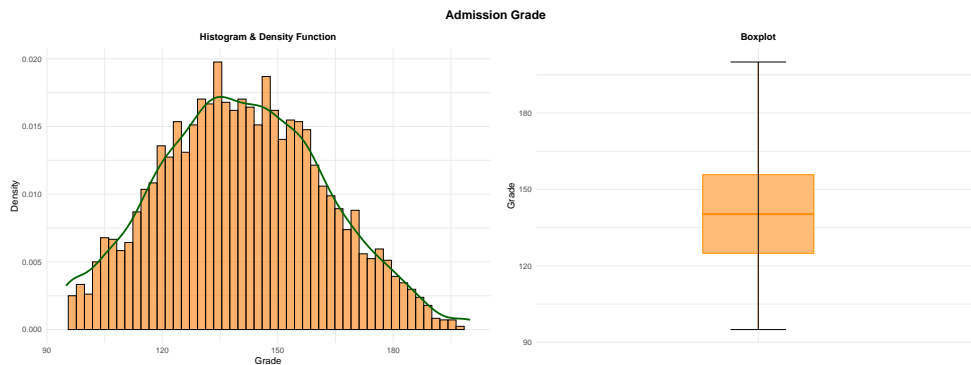


Figura 7.5: Dataset Sintetico: analisi "Admission Grade"

Esempio: Inflation Rate

Per il tasso di inflazione, illustrato in Figura 7.6, così come per altri fattori macroeconomici, nel JSON non sono stati imposti vincoli specifici in termini di range validi, lasciando al modello totale libertà nella generazione dei dati. Il LLM ha prodotto valori realistici, ma non ha riconosciuto autonomamente che, nei dati reali, questa variabile assumeva un numero discreto di valori, poiché i dati sono stati raccolti in un periodo temporale limitato. Di conseguenza, il dataset sintetico contiene un numero molto più elevato di valori distinti rispetto ai dati originali. Inoltre, data l'assenza di limiti nei valori, i dati sintetici dei fattori macro-economici sembrano seguire una distribuzione normale ancora più perfetta rispetto ad altre variabili, con Skewness = 0.01 e Curtosi = 3.05, confermando la tendenza del modello a produrre distribuzioni estremamente regolari e prive di variazioni naturali.

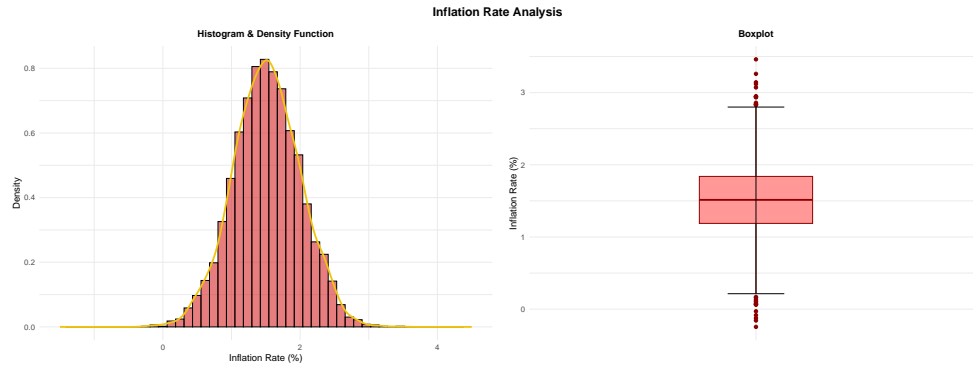


Figura 7.6: Dataset Sintetico: analisi "Inflation Rate"

Considerazioni Finali

Gli esempi analizzati riflettono il comportamento generale delle feature numeriche nel dataset sintetico. Sebbene il modello abbia generato dati coerenti con i vincoli imposti e con il significato delle variabili, emergono alcune tendenze ricorrenti:

- Le distribuzioni sembrano essere riconducibili a modelli statistici noti, come la Normale, la Binomiale o l'Uniforme, confermando l'approccio strutturato del modello nella generazione dei dati.
- Il LLM ha rispettato i vincoli sui range imposti nel JSON, ma in alcuni casi non ha riconosciuto proprietà strutturali interne delle variabili.
- La distribuzione dei dati è molto più regolare rispetto a quella reale, con una riduzione significativa della presenza di outliers e anomalie, rendendo il dataset sintetico più pulito ma meno realistico.
- Il modello non è riuscito a catturare le imperfezioni presenti nei dati reali, dovute a fenomeni come l'arrotondamento dei voti o la variabilità del comportamento degli studenti.

🔍 Finding RQ4. Il LLM sembra riprodurre le caratteristiche generali delle variabili numeriche, seguendo modelli statistici noti e rispettando i vincoli imposti. Tuttavia, manca della complessità e del rumore dei dati reali, risultando più regolare, con meno outliers e senza imperfezioni strutturali

7.4 Associazioni con la Variabile Target

Uno degli aspetti più rilevanti del dataset reale era la relazione tra le diverse feature e la variabile target. Per valutare la capacità del LLM di generare un dataset sintetico che mantenesse queste relazioni, sono state testate diverse strategie di *prompt engineering*, specificando in dettaglio le associazioni sia nel prompt che nel file JSON. Tuttavia, il modello non è mai riuscito a generare dati coerenti con le specifiche fornite, evidenziando una chiara limitazione nella conservazione delle relazioni tra le variabili.

Esempio: Mother's Qualification VS Target

La Figura 7.7 mostra la distribuzione della variabile **Mother's Qualification** rispetto alla variabile target. Nel dataset reale, le diverse categorie di questa feature presentavano una distribuzione simile, con l'eccezione della modalità *6-No Qualification*, che mostrava un tasso di *Dropout* più elevato. Nonostante questa informazione fosse stata esplicitata nel JSON, il modello ha mantenuto la distribuzione globale della variabile target all'interno di ciascun gruppo, senza differenziare le categorie.

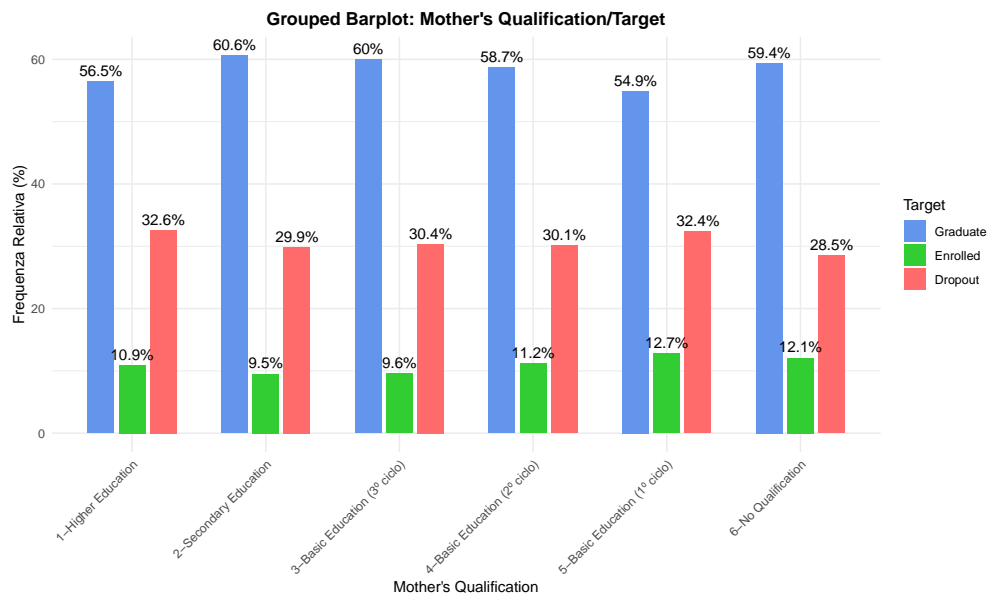


Figura 7.7: Dataset Sintetico: Mother's Qualification VS Target

Esempio: Age at Enrollment VS Target

La Figura 7.8 mostra la distribuzione della variabile **Age at Enrollment** rispetto alla variabile target. Nei dati reali, era evidente una relazione tra queste variabili: gli studenti più giovani avevano una maggiore probabilità di laurearsi nei tempi previsti, mentre quelli con un'età più avanzata tendevano a rimanere iscritti più a lungo o ad abbandonare gli studi. Nonostante questa relazione fosse stata esplicitamente indicata nel JSON, il modello non è stato in grado di riprodurla, generando una distribuzione dell'età identica per tutti i gruppi della variabile target, ignorando completamente l'effetto che l'età di immatricolazione può avere sul successo accademico.

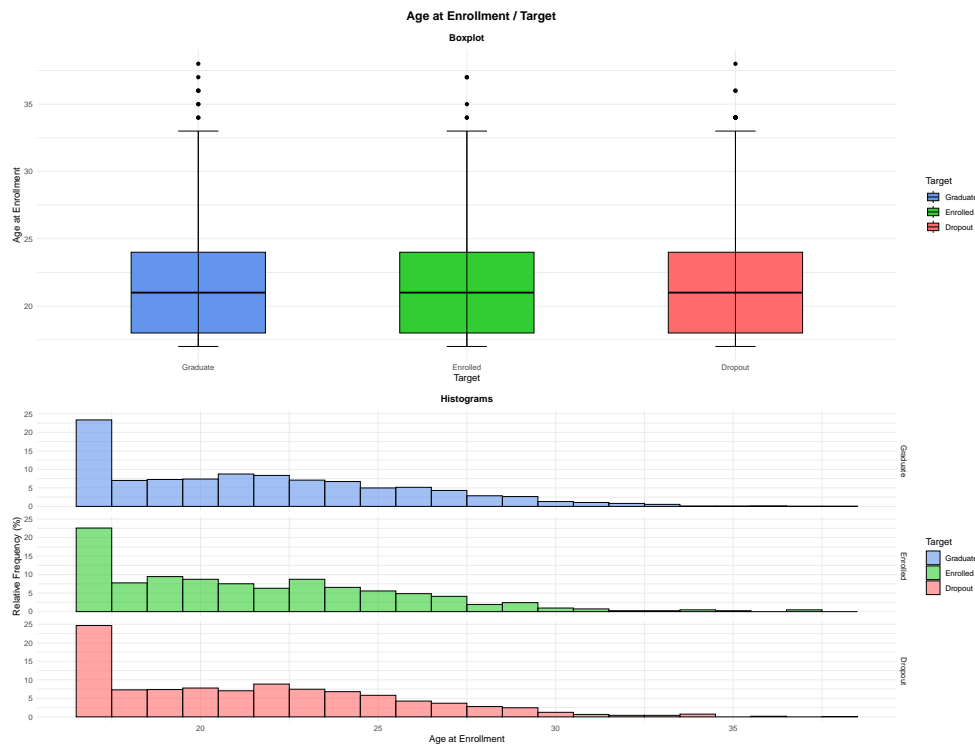


Figura 7.8: Dataset Sintetico: Age At Enrollment VS Target

Esempio: Admission Grade VS Target

La Figura 7.9 mostra la distribuzione della variabile **Admission Grade** rispetto alla variabile target. Nei dati reali, questa variabile non presentava un'associazione particolarmente forte con il successo accademico, ma si osservava comunque una leggera differenza: gli studenti con votazioni più basse tendevano ad avere probabilità di abbandono leggermente più elevata. Nel dataset sintetico, invece, il modello ha gene-

rato una distribuzione dell'*Admission Grade* praticamente identica, con una leggera differenza solo per gli studenti Enrolled (probabilmente dovuta solo al caso), tra tutti i gruppi della variabile target, annullando qualsiasi differenza osservabile.

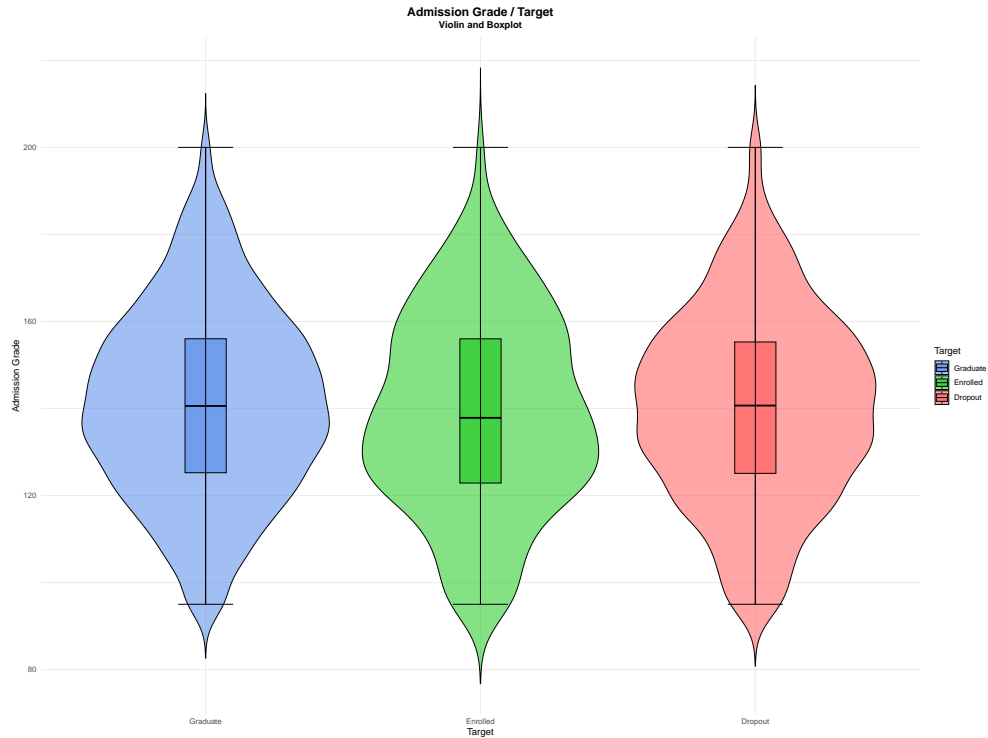


Figura 7.9: Dataset Sintetico: Admission Grade VS Target

Considerazioni Finali

L'analisi delle relazioni tra le feature e la variabile target ha evidenziato una significativa limitazione del modello nella generazione di dati coerenti con la realtà. Nonostante le dettagliate specifiche fornite nel prompt e nel file JSON, il LLM non è stato in grado di preservare le relazioni osservate nei dati reali. In particolare:

- Per le variabili numeriche, le distribuzioni risultano indipendenti dalla variabile target, eliminando qualsiasi possibilità di analisi sulle relazioni tra i dati.
- Per le variabili categoriche, il modello ha mantenuto la distribuzione globale della variabile target in ogni categoria, ignorando completamente le differenze osservate nei dati reali.

Queste limitazioni riducono notevolmente l'utilità del dataset sintetico per analisi che dipendono dalle interazioni tra le variabili. Nel dataset reale, le relazioni tra le

feature e la variabile target rappresentavano l'aspetto più informativo e utile per comprendere i fattori che influenzano il successo accademico. La loro assenza nei dati sintetici compromette la capacità del dataset generato di essere utilizzato in contesti di analisi predittiva o interpretativa.

🔍 Finding RQ4. Il modello non ha preservato le relazioni tra le feature e la variabile target, generando dati indipendenti e privi di correlazioni reali. Le distribuzioni delle variabili numeriche risultano uguali per tutti i gruppi Target, mentre le categoriche mantengono solo la distribuzione globale

7.5 Correlazioni tra Feature Numeriche

A differenza delle frequenze congiunte tra variabili categoriche e delle distribuzioni numeriche nei gruppi target, specificando nel file JSON la presenza di correlazioni tra variabili numeriche, il modello è riuscito a generare dati che rispettano le specifiche imposte, mantenendo le relazioni indicate. La Figura 7.10 e la Figura 7.11 mostrano due esempi di correlazioni tra variabili numeriche: i voti ottenuti nelle unità curriculari tra il primo e il secondo semestre, e la relazione tra il voto della qualifica precedente e il voto di ammissione. Come si può osservare, le correlazioni sono quasi perfette, con una fitta nuvola di punti disposti lungo una linea retta quasi ideale.

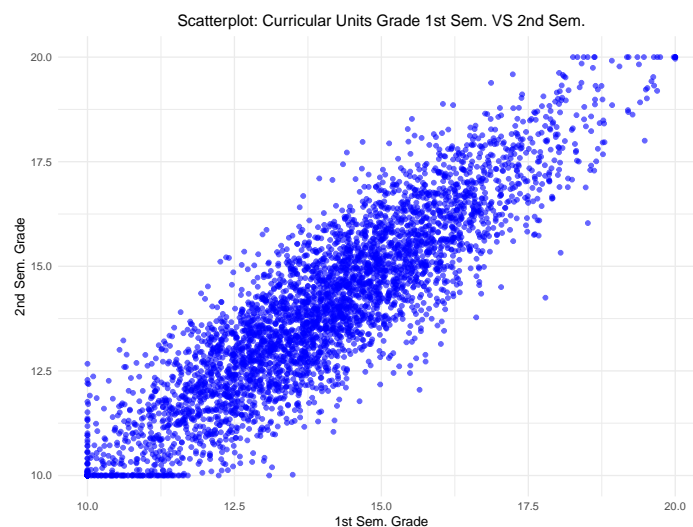


Figura 7.10: Dataset Sintetico: CU Grade 1st Sem. VS 2nd Sem.

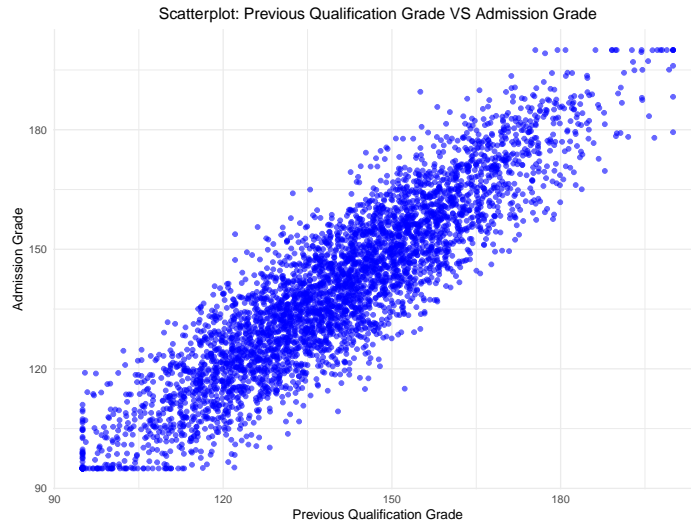


Figura 7.11: Dataset Sintetico: Previous Qualification Grade Vs Admission Grade

Sebbene il rispetto delle correlazioni imposte sia un aspetto positivo, emergono alcune criticità:

- Nel dataset reale, tali correlazioni erano moderate e non perfette, poiché il comportamento umano introduce una componente di variabilità. Fattori esterni, come cambiamenti di rendimento, motivazione e difficoltà soggettive, influenzano le relazioni tra variabili come i voti, determinando una dispersione che nei dati sintetici è completamente assente;
- il modello non è in grado di inferire autonomamente correlazioni tra i dati in base al loro significato o dalle righe di esempio fornite con il prompt. Ogni relazione desiderata deve essere specificata manualmente, rendendo complessa la gestione delle interdipendenze in dataset con molte variabili;
- In alcuni casi, per soddisfare una correlazione tra due variabili, il modello ha generato valori che violavano vincoli logici imposti nel dataset. Ad esempio, sono state trovate osservazioni con *Curricular Units Approved* < 0 , un valore non ammissibile.

🔍 Finding RQ4. Il LLM rispetta le correlazioni imposte, ma in modo rigido e privo della variabilità tipica dei dati reali. Inoltre, non inferisce autonomamente relazioni tra variabili, richiedendo specifiche manuali per ogni interdipendenza.

7.6 Test del Chi-Quadro Bilaterale

Nelle sezioni precedenti abbiamo analizzato alcune variabili numeriche generate dal LLM, osservando che nella maggior parte dei casi esse sembrano seguire delle distribuzioni note. In questa sezione verrà approfondita questa osservazione, verificando con un'analisi statistica se le variabili quantitative continue possono essere effettivamente ricondotte a una distribuzione normale, come osservato precedentemente. A tale scopo, viene eseguito il **test del Chi-Quadro bilaterale** con un livello di significatività del 99%.

7.6.1 Regola di Scott

Un passo fondamentale per eseguire il test del Chi-Quadro consiste nel suddividere i dati in intervalli discreti. Per dati continui, è essenziale scegliere un numero ottimale di intervalli, in modo da garantire un'adeguata rappresentazione della distribuzione senza introdurre distorsioni. A tale scopo è stata utilizzata la **Regola di Scott**, un metodo comunemente adottato per determinare il numero di classi in un istogramma, particolarmente efficace per dataset di grandi dimensioni e distribuzioni simmetriche (come nel nostro caso). La formula per calcolare il numero di intervalli è la seguente:

$$k = \frac{\max(X) - \min(X)}{\frac{3.5 \cdot \sigma}{n^{1/3}}}$$

dove:

- k è il numero ottimale di intervalli;
- $\max(X)$ e $\min(X)$ sono il valore massimo e minimo del dataset;
- σ è la deviazione standard della variabile;
- n è il numero totale di osservazioni.

7.6.2 Metodologia del Test

Per verificare se le variabili numeriche siano normalmente distribuite, è stato applicato il **test del Chi-Quadro Bilaterale**, andando a confrontare la distribuzione empirica dei dati con una distribuzione normale teorica (nel nostro caso specifico).

Le ipotesi del test effettuato nel nostro caso sono le seguenti:

- **Ipotesi nulla** (H_0): la variabile segue una distribuzione normale;
- **Ipotesi alternativa** (H_1): la variabile non segue una distribuzione normale.

Il test è stato condotto seguendo i seguenti passi:

- **Suddivisione dei dati in intervalli**: una volta determinato il numero ottimale di intervalli k per ogni variabile, i dati sono stati suddivisi utilizzando i quantili della distribuzione normale $N(\bar{X}, s)$. Poiché i quantili della distribuzione normale suddividono i dati in intervalli con la stessa probabilità teorica, la frequenza attesa in ciascun intervallo è pari a:

$$E_i = n * \frac{1}{k}.$$

- **Calcolo della statistica del Chi-Quadro**: per ogni intervallo, è stata calcolata la frequenza osservata O_i , e la statistica del Chi-Quadro è stata calcolata come:

$$\chi^2 = \sum_{i=1}^k \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2.$$

- **Determinazione dell'intervallo critico**: il valore della statistica è stato confrontato con i quantili della distribuzione del Chi-Quadro con $k - p - 1$ gradi di libertà, dove $p = 2$ rappresenta i parametri stimati (media e deviazione standard). L'ipotesi nulla viene accettata se:

$$\chi_{\alpha/2, df}^2 \leq \chi^2 \leq \chi_{1-\alpha/2, df}^2$$

con $\alpha = 0.01$ e $df = k - 3$.

7.6.3 Risultati del Test

La Tabella 7.1 riporta i risultati ottenuti per ciascuna variabile analizzata. I risultati ottenuti confermano che tutte le variabili quantitative continue analizzate possono essere considerate distribuite normalmente con un livello di confidenza del 99%.

Questo è coerente con quanto osservato nei grafici delle sezioni precedenti, dove le distribuzioni presentano la tipica forma gaussiana.

Variabile	Intervalli	Valore χ^2	Intervallo critico
GDP	32	27.536	[16.047, 45.722]
Inflation Rate	35	24.650	[18.291, 49.480]
Unemployment Rate	33	41.807	[16.791, 46.979]
Admission Grade	22	24.768	[8.907, 32.852]
Previous Qualification Grade	24	12.164	[10.283, 35.479]
Curricular Units 1st Sem. Grade	24	21.884	[10.283, 35.479]
Curricular Units 2nd Sem. Grade	22	22.392	[8.907, 32.852]

Tabella 7.1: Risultati del test del Chi-Quadro bilaterale

🔗 Finding RQ4. Con un livello di confidenza del 99%, le variabili quantitative continue generate dal LLM sono tutte riconducibili ad una distribuzione normale

7.7 Considerazioni Finali

L'analisi del dataset sintetico generato dall'LLM ha evidenziato qualche punto di forza e limitazioni significative nella capacità del modello di replicare le caratteristiche del dataset reale:

- Per quanto riguarda le **feature categoriche**, il modello ha inizialmente prodotto dati con una distribuzione delle categorie bilanciata. Tuttavia, fornendo informazioni più specifiche sulla distribuzione attesa, è stato possibile ottenere dati con sbilanciamenti più realistici, in linea con il dataset originale. Nonostante questo miglioramento, per variabili con un numero elevato di categorie, il controllo della distribuzione rimane complesso e l'LLM tende a generare sbilanciamenti poco naturali.
- Sul fronte delle **feature numeriche**, il modello ha generato distribuzioni generalmente coerenti con il significato delle variabili. Tuttavia, queste risultano

spesso riconducibili a distribuzioni note (es. Normale, Binomiale, Uniforme), con dati fin troppo regolari rispetto ai dati reali. In particolare, mancano elementi di casualità come outlier, errori di misurazione o arrotondamenti discreti, rendendo il dataset sintetico meno realistico.

- Una delle principali limitazioni si è manifestata nella generazione delle **distribuzioni congiunte** tra variabili, specialmente nella relazione tra feature e variabile target. Nonostante tentativi di specificare nel prompt relazioni più complesse, il LLM non è stato in grado di riprodurre fedelmente le frequenze congiunte osservate nei dati reali. Questo aspetto rappresenta un limite significativo, poiché le relazioni tra variabili sono spesso l'elemento più interessante nei dataset come questo in analisi.
- Il modello si è dimostrato capace di generare **correlazioni tra feature numeriche**, rispettando le specifiche fornite. Tuttavia, le correlazioni risultano eccessivamente perfette, prive della variabilità tipica dei dati reali. Inoltre, la gestione delle correlazioni incrociate tra più variabili è risultata problematica, con il modello che in alcuni casi ha generato dati incoerenti per soddisfare vincoli di correlazione.

In sintesi, sebbene l'LLM abbia dimostrato buone capacità nel generare distribuzioni realistiche per singole feature e nel rispettare correlazioni numeriche, la sua efficacia nella riproduzione delle relazioni tra variabili rimane limitata. Questo aspetto riduce significativamente la validità di un dataset sintetico generato con questa metodologia, specialmente per analisi che dipendono dalle interazioni tra le variabili. Per comprendere queste difficoltà del LLM, basta ragionare sulla loro natura: essi generano dati token per token sulla base di probabilità condizionate locali, senza una visione d'insieme delle dipendenze tra variabili. Questo li rende efficaci nel rispettare vincoli isolati, ma inadatti a garantire coerenza nelle distribuzioni congiunte su larga scala.

CAPITOLO 8

Conclusioni

Nel corso di questo studio, è stata analizzata la relazione tra i fattori demografici, socio-economici e accademici degli studenti e il loro esito formativo, con l'obiettivo di comprendere quali elementi influenzano il successo o l'abbandono universitario. Inoltre, è stata condotta un'analisi di clustering per identificare gruppi di studenti con caratteristiche/comportamenti simili e sono stati valutati intervalli di confidenza sui voti. Infine, è stata verificata la capacità di un LLM di generare un dataset sintetico che replicasse le proprietà statistiche del dataset originale. Nel seguente capitolo, vengono riportate le risposte alle Research Questions (RQ), basate sui risultati ottenuti.

8.1 Research Question 1

L'analisi bivariata condotta nel Capitolo 3 e i test statistici eseguiti nel Capitolo 4 hanno permesso di verificare come alcune variabili siano significativamente associate agli esiti accademici degli studenti. In particolare, le variabili che mostrano le associazioni più rilevanti sono:

- **Fattori Economici:** variabili economiche come la regolarità nel pagamento delle tasse universitarie, la presenza di debiti economici e il beneficio di una borsa

di studio risultano fortemente associate agli esiti accademici, influenzandoli in modo positivo o negativo a seconda del contesto.

- **Performance nel primo anno:** gli indicatori di rendimento durante il primo anno accademico rappresentano alcuni dei fattori maggiormente associati all'esito degli studenti. In particolare, la percentuale di esami superati, il tasso di successo negli esami tentati e la media dei voti nei due semestri sono strettamente legati alla probabilità di completare il percorso o di abbandonarlo.

Anche altre variabili, come il sesso degli studenti, il corso frequentato e l'età al momento dell'immatricolazione, mostrano un'associazione significativa con l'esito accademico. Mentre fattori come il titolo di studio e l'occupazione dei genitori risultano anch'essi correlati, seppur con un impatto meno marcato rispetto alle variabili precedenti e più concentrato in alcune categorie specifiche.

8.2 Research Question 2

Il clustering condotto nel Capitolo 5 ha permesso di individuare **quattro gruppi distinti** di studenti, caratterizzati da esigenze differenti. Questi risultati offrono una base utile per l'implementazione di strategie di supporto mirate, volte a migliorare il successo accademico e ridurre i tassi di abbandono:

- **Cluster 1 - Alto Rendimento e Alto Carico di Studio:** composto da studenti con un rendimento elevato, ma con un carico di studio particolarmente intenso. Questi studenti potrebbero beneficiare di percorsi di approfondimento, oltre che di programmi per la gestione del carico accademico, al fine di ridurre i tassi di dropout. Inoltre, sarebbe opportuno valutare un miglioramento nell'organizzazione delle unità curriculari dei corsi frequentati da questi studenti, per garantire un percorso di studi più sostenibile.
- **Cluster 2 - Altissimo Rendimento:** include studenti che ottengono costantemente risultati accademici eccellenti. Per loro potrebbero essere previsti programmi di eccellenza o percorsi di approfondimento avanzati, con l'obiettivo di stimolare ulteriormente il loro rendimento e incentivarne la crescita accademica.

- **Cluster 3 - Performance Molto Scarse:** comprende studenti con prestazioni accademiche estremamente basse, che non riescono a completare la maggior parte delle unità curriculari previste. Per questi studenti sarebbe opportuno attivare programmi di tutorato e supporto personalizzati. Inoltre, potrebbe essere utile condurre indagini più approfondite su fattori extra-accademici che potrebbero influenzare negativamente la loro performance, con l'obiettivo di ridurre il rischio di dropout.
- **Cluster 4 - Media Performance e Progressione Lenta:** composto da studenti con rendimento accademico intermedio, che non ottengono risultati eccellenti e avanzano con maggiore difficoltà nel percorso di studi. Per questi studenti sarebbe utile implementare programmi di tutorato, finalizzati a favorire il completamento del percorso nei tempi previsti ed evitare l'abbandono prematuro.

8.3 Research Question 3

L'analisi condotta nel Capitolo 6 ha permesso di ottenere le seguenti conclusioni con un livello di significatività del 99%:

- **1st Year Grade:** la media dei voti ottenuti durante il primo anno accademico per l'intera popolazione è compresa nell'intervallo di confidenza $[12.231, 12.404]$. Inoltre, le studentesse presentano una media significativamente superiore rispetto agli studenti maschi, con un intervallo di confidenza per la differenza tra le medie pari a $[-0.828, -0.427]$.
- **Admission Grade:** la media dei voti di ammissione per l'intera popolazione rientra nell'intervallo di confidenza $[126.417, 127.539]$. Non emergono differenze statisticamente significative tra i voti di ammissione degli studenti maschi e delle studentesse femmine.
- **Previous Qualification Grade:** la media dei voti relativi alla qualifica precedente per l'intera popolazione è compresa nell'intervallo di confidenza $[132.103, 133.124]$. Le studentesse mostrano una media significativamente superiore rispetto agli

studenti maschi, con un intervallo di confidenza per la differenza tra le medie pari a $[-2.399, -0.245]$.

8.4 Research Question 4

L'analisi del dataset sintetico generato con *ChatGPT-4o* condotta nel Capitolo 7 ha evidenziato risultati contrastanti. Il modello è stato in grado di produrre dati con proprietà statistiche plausibili, ma ha mostrato limiti significativi nella riproduzione della complessità e delle relazioni strutturali osservate nel dataset reale:

- **Distribuzione delle singole feature:** Il modello ha generato dati coerenti con le specifiche fornite, rispettando i vincoli definiti nel JSON e le distribuzioni attese. Per le variabili categoriche, in particolare, fornire informazioni dettagliate nel prompt ha permesso di ottenere distribuzioni più realistiche, con sbilanciamenti coerenti con il dataset originale.
- **Distribuzione delle variabili numeriche:** Sebbene i dati numerici generati risultino plausibili, essi sono quasi sempre riconducibili a distribuzioni note (es. Normale, Binomiale, Uniforme). Tuttavia, rispetto ai dati reali, il dataset sintetico appare eccessivamente regolare, privo della variabilità naturale, della casualità e degli outlier tipici dei dati raccolti nel mondo reale. Questo limita la capacità del modello di rappresentare la complessità della popolazione osservata.
- **Assenza di relazioni tra feature e variabile target:** Uno degli aspetti più critici riguarda la mancata riproduzione delle relazioni congiunte tra le feature e la variabile target (*Graduate*, *Enrolled*, *Dropout*). Nonostante specifiche dettagliate nel prompt e nel JSON, il modello ha generato distribuzioni indipendenti, ignorando completamente i pattern presenti nel dataset reale. Questo compromette l'utilità del dataset per analisi predittive e interpretative.
- **Correlazioni numeriche riprodotte in modo eccessivamente rigido:** Quando specificate esplicitamente nel JSON, le correlazioni tra variabili numeriche sono state rispettate, ma con una rigidità eccessiva. Le relazioni tra le variabili

risultano quasi perfette, senza la naturale dispersione osservata nei dati reali. Fattori come il comportamento umano, l'errore di misurazione e le dinamiche individuali tendono a introdurre variabilità nelle correlazioni, un elemento che il modello non è stato in grado di replicare.

In conclusione, sebbene il modello sia in grado di produrre dati con distribuzioni plausibili per singole feature e di rispettare correlazioni numeriche quando specificate, la mancanza di relazioni congiunte tra le variabili e la rigidità delle correlazioni imposte ne limitano l'utilità. Per compiti che richiedono una maggiore fedeltà alle strutture dei dati reali, potrebbe essere più efficace adottare strumenti specifici per la generazione di dataset sintetici, che offrono una maggiore capacità di modellare relazioni complesse tra le variabili.

Bibliografia

- [1] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, 2022. [Online]. Available: <https://www.mdpi.com/2306-5729/7/11/146> (Citato alle pagine 1, 2 e 3)
- [2] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early prediction of student's performance in higher education: A case study," in *Trends and Applications in Information Systems and Technologies*, Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, and A. M. Ramalho Correia, Eds. Cham: Springer International Publishing, 2021, pp. 166–175. (Citato alle pagine 1, 2 e 3)