F9101Q005M - Machine Learning Project

Sara Fabbro (821317)[1], Lucia Gerbi (817537)[1],
Davide Porcellini (816586)[1], Alessandro Vaccarino (811751)[1]
[1] *Università di Milano Bicocca, Piazza dell'Ateneo Nuovo 1, Milano, Italia*
s.fabbro@campus.unimib.it, l.gerbi@campus.unimib.it, d.porcellini2@campus.unimib.it, a.vaccarino@campus.unimib.it

Keywords: Portugal, scholarship, Machine Learning, KNIME, F9101Q005M.

Abstract: Per la costruzione di una società intelligente e libera occorre investire nell'istruzione, in quanto l'accesso al sapere crea cittadini in grado di approcciarsi alla realtà quotidiana in modo critico e consapevole. È importante che ci siano delle condizioni favorevoli a tale sviluppo, sostenute da politiche attente alla crescita della persona. Ma come è possibile sostenere tali politiche? Quale contributo può essere dato dalla moderna disponibilità informativa? Nel presente report si presenta un approccio data driven alla valutazione delle performance scolastiche di un gruppo di studenti delle scuole superiori Portoghesi, al fine di identificare quali siano i fattori che maggiormente possono influenzare il successo nello studio.

1 INTRODUZIONE

Al giorno d'oggi, il mercato del lavoro presenta dinamiche sempre più frenetiche e sfidanti. La natura delle figure professionali e delle competenze richieste cambia ad una velocità fino a pochi anni fa difficilmente immaginabile. Secondo uno studio del World Economic Forum (Schwab and Samans, 2016) il 65% dei bambini che oggi frequentano le scuole elementari svolgeranno in età adulta un lavoro che oggi non esiste. Quindi iventa strategico, in un simile contesto, investire in modo oculato nella formazione degli studenti, fin dalla prima scoltarità. Ad oggi in Italia (AICA, 2019) e in Europa (Cedefop, 2019a) (Cedefop, 2019b) si registrano differenti iniziative orientate all'analisi dell'evoluzione del mercato del lavoro, tra i cui fini c'è il supporto ad una più puntuale pianificazione dell'offerta formativa.

D'altro canto, è necessario supportare gli studenti fin dai primi anni di formazione, affinché essi possano esprimere al meglio le proprie potenzialtà. Esistono numerose iniziative a tal riguardo, che contemplano:

- Stimolo agli studenti più profittevoli mediante borse di studio, esperienze Erasmus ed attività integrative.
- Sostegno agli studenti con maggiori diffcoltà, con attività di assistenza allo studio.

Ma come è possibile massimizzare l'efficienza di tali leve? Quali sono i fattori che possono maggiormente incidere sulle capacità di apprendimento di uno studente?
In questo studio è stato considerato il contesto scolastico Portoghese, caratterizzato da un'importante crescita in termini di performance (PISA, 2003) (PISA, 2018) e dalla disponibilità di una banca dati adeguata a cui è stato possibile accedere.

Lo studio proposto si articola secondo le seguenti sezioni:

- Nel capitolo *CONTESTO* verrà brevemente approfondito il contesto scoltastico portoghese, oggetto dello studio.
- Il successivo capitolo *DATI* presenterà la fonte dati utilizzata, descrivendone il contenuto e le caratteristiche informative fondamentali.
- Il capitolo *METODOLOGIA* riepilogherà le metodologie utilizzate per lo studio, presentando i modelli e le tecniche utilizzate.
- Il capitolo *ANALISI* presenterà le analisi effettuate sui dati in possesso, descrivendone modalità, criticità riscontrate e risultati ottenuti
- Il capitolo *RISULTATI E CONCLUSIONI* riassumerà i risultati ottenuti nello studio, l'attinenza con quanto ipotizzato originariamente ed una conclusione relativa allo studio svolto

2 CONTESTO

Negli ultimi anni il Portogallo ha assistito ad un importante crescita a livello scolastico, come eviden-

ziato dai risultati ottenuti dal Programma per la Valutazione Internazionale degli Studenti (*PISA*) (PISA, 2003) (PISA, 2018). Tale iniziativa rappresenta una ricerca internazionale a cadenza triennale supportata dall'*OCSE* (Organizzazione per la Cooperazione e lo Sviluppo Economico) al fine di ottenere valutazioni circa il livello di istruzione dei ragazzi in età adolescenziale nei paesi industrializzati.
L'indagine si pone l'obiettivo di valutare le capacità nella lettura, nella matematica e in generale in ambito scientifico, fornendo un punteggio medio che sintetizza il livello di tali conoscenze per ogni paese della comunità europea. (OECD, 2020).

Per quanto concerne la situazione portoghese, l'ultima indagine svolta (PISA, 2018) evidenzia un progressivo miglioramento avvenuto rispetto alle prime indagini svolte nel 2000. Tale miglioramento ha portato il Portogallo ad ottenere punteggi superiori alla media OCSE (rif. Figura 1).

Concentrandosi sul caso scolastico portoghese appena presentato, lo studio si pone l'obiettivo di analizzare quali sono le variabili di contesto che hanno maggiormente influenzato le performance degli studenti di scuola secondaria.

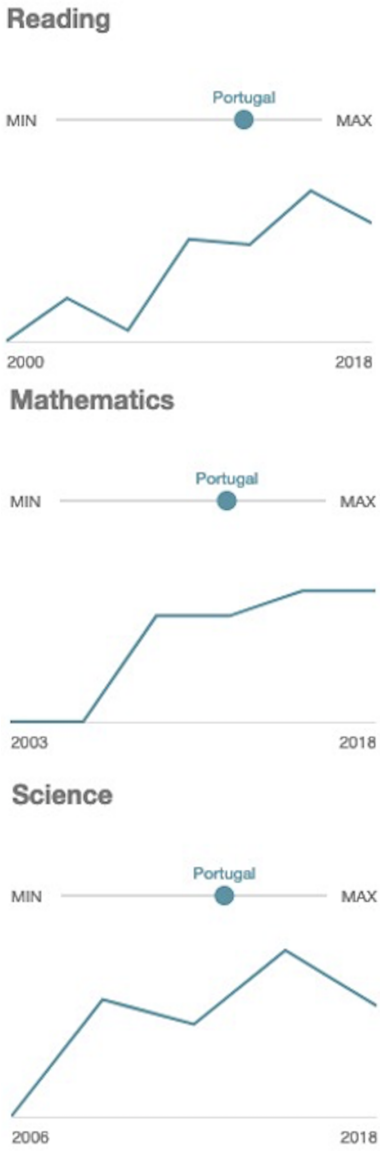

Figure 1: Portugal performance in PISA 2018

3 DATI

Al fine di raggiungere l'obiettivo di ricerca presentato, sono stati utilizzati dati riguardanti studenti di due scuole secondarie portoghesi, "*Gabriel Pereira*" e "*Mousinho da Silveira*", ottenuti tramite questionari. Le risposte sono state raccolte e rese disponibili nel dataset "Student Alcohol Consumption" (Kaggle, 2017). Nel presente studio è stato considerato solo il dataset riguardante il corso di portoghese. I fattori indagati sono:

- le quantità di alcohol assunto
- il background familiare
- il background scolastico
- il background sociale

L'obiettivo è **quantificare l'impatto che tali fattori possono avere nel voto finale che viene attribuito a ciascun studente al termine del corso di studi**.
Con voto finale si intende un punteggio compreso tra 0 e 20, stratificato dal ministero portoghese (Ministerio da ciencia, 2005) secondo la metodologia riportata in tabella 1.
La tabella 2 presenta invece le variabili che compongono il dataset.

Punteggio	Valutazione
<10	Insufficiente
10-13	Sufficiente
14-15	Buono
16-17	Molto Buono
18-20	Eccellente

Table 1: Criteri di assegnamento valutazione

Colonna	Descrizione	Valori
school	Istituto	'GP' (Gabriel Pereira), 'MS' (Mousinho da Silveira)
sex	Sesso	'F' (Donna), 'M' (Uomo)
age	Età	Numerica, da 15 a 22
address	Tipologia di residenza dello studente	'U' (Urbana), 'R' (Rurale)
famsize	Dimensione della famiglia	'LE3' ($\leq$3), 'GT3' (>3)
Pstatus	Stato di convivenza dei genitori	'T' (Vivono insieme), 'A' (Vivono separati)
Medu	Livello educativo della madre	0 (Nessuno), 1 (Scuola primaria), 2 (Scuola secondaria inferiore), 3 (Scuola secondaria superiore), 4 (Laurea o superiore)
Fedu	Livello educativo del padre	0 (Nessuno), 1 (Scuola primaria), 2 (Scuola secondaria inferiore), 3 (Scuola secondaria superiore), 4 (Laurea o superiore)
Mjob	Lavoro della madre	'teacher', 'health', civil 'services', 'at_home', 'other'
Fjob	Lavoro del padre	'teacher', 'health', civil 'services', 'at_home', 'other'
reason	Ragione della scelta della scuola	'home' (vicinanza a casa), 'reputation' (reputazione della scuola), 'course' (preferenza del corso), 'other' (altro)
guardian	Tutore dello studente	'mother', 'father', 'other'
traveltime	Tempo di percorrenza per la scuola	1 (<15'), 2 (15-30'), 3 (30'-1h), 4 (>1h)
studytime	Tempo settimanalmente dedicato allo studio	1 (<2h), 2 (2-5h), 3 (5-10h), 4 (>10h)
failures	Numero di esami non passati durante la carriera	n se 1$\leq$n<3, altrimenti 4
schoolsup	Supporto extra-scolastico	True/False
famsup	Supporto familiare nello studio	True/False
paid	Accesso a corsi integrativi (a pagamento) di matematica o lingua	True/False
activities	Attività extra-curriculari	True/False
nursery	Frequenza della scuola materna	True/False
higher	Volontà di frequentare le scuole superiori	True/False
internet	Accesso ad internet da casa	True/False
romantic	Impegnato in una relazione	True/False
famrel	Qualità delle relazioni in famiglia	Da 1 (pessimo) a 5 (eccellente)
freetime	Tempo libero dopo scuola	Da 1 (poco) a 5 (molto)
goout	Tempo dedicato alla frequentazione di amici	Da 1 (poco) a 5 (molto)
Dalc	Consumo di alcolici durante la settimana	Da 1 (poco) a 5 (molto)
Walc	Consumo di alcolici durante la il fine settimana	Da 1 (poco) a 5 (molto)
health	Attuale quadro sanitaria	Da 1 (pessimo) a 5 (eccellente)
absences	Numero di assenze da scuola	Da 0 a 93
G1	Votazione conseguita nel primo semestre	Da 0 a 20
G2	Votazione conseguita nel secondo semestre	Da 0 a 20
G3	Votazione conseguita al temrine dell'anno (voto finale)	Da 0 a 20

Table 2: Variabili contenute nel dataset

Sono state rilevate 649 osservazioni su 33 variabili. Tra queste, la variabile *G3* indica il voto finale ed è stata selezionata come target dell'analisi; essa assume la seguente distribuzione:

Figure 2: Istogramma *G3*

Il range di valori assunti da *G3* varia da 0 a 19, perciò nessuno studente ha conseguito il voto massimo.

Sono state escluse dallo studio le covariate *G1* e *G2* dopo aver osservato forti correlazioni con *G3*, pari rispettivamente a 0.826 e 0.918. È un risultato coerente con le aspettative: è ragionevole pensare che gli studenti mantengano un livello di risultati costante durante il corso dell'anno e quindi, osservando le valutazioni intermedie, è possibile stimare la valutazione finale con una precisione elevata. La scelta di esclusione è stata compiuta al fine di evidenziare effetti differenti dalle valutazioni del primo e del secondo trimestre che potrebbero non emergere sotto la capacità esplicativa delle variabili citate. In figura 3 si riporta la matrice delle correlazioni ottenuta escludendo le valutazioni *G1*, *G2* e *G3*:

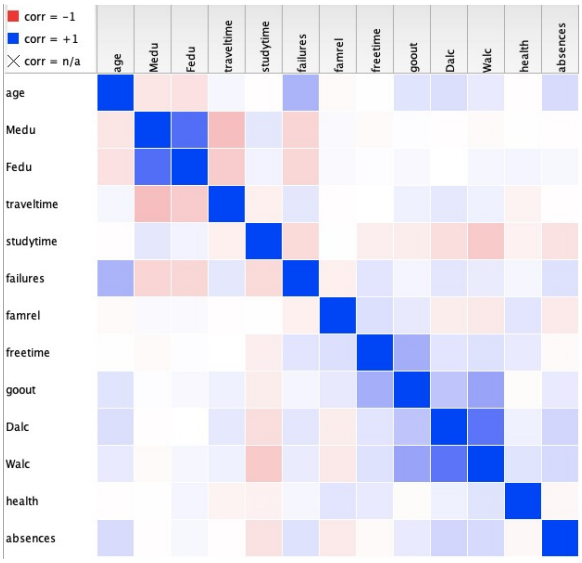

Figure 3: Correlazioni escludendo G1, G2 e G3

Non sono presenti coppie di variabili con correlazione superiore a 0.7.

Le variabili categoriali a due livelli sono state binarizzate al fine di rendere più facile l'interpretazione, per esempio per le dicotomiche a valori *Yes* e *no* si è scelto di codificare il primo con 1 e il secondo con 0.

L'analisi delle distribuzioni non ha fatto emergere variabili con valori considerabili come anomali, ad eccezione della variabile *Absences*:

Figure 4: Istogramma *Absences*

Come si può notare dalla distribuzione di *Absences* riportata in figura 4, la variabile ha un valore mediano pari a 2 mentre il massimo è pari a 32, associato al questionario di una studentessa della scuola *Gabriel Pereira* che ha conseguito con voto finale *G3* pari a 14. Indagando i valori assunti dalle altre variabili per il caso in analisi si è distinta *Health*, a cui è associato il livello 1, indicante *pessime condizioni di salute*. Si è quindi ipotizzato che la ragazza in questione abbia avuto problemi fisici o psichici che non le abbiamo permesso di frequentare le lezioni per molti giorni.

4 METODOLOGIA

Con l'obiettivo di stimare la variabile target sono state applicati differenti modelli classificativi presentati di seguito.

4.1 Modelli euristici

Permettono di risolvere un problema di classificazione con tecniche poco complesse in grado di ottenere risultati ammissibili e approssimativamente buoni, ma di cui non è assicurata l'ottimalità. Inoltre, sono algoritmi che potrebbero "fallire" nel fornire una soluzione, anche se questo non comporta il fatto che non esistano soluzioni al problema ma solo che il modello non è in grado di individuarle. Tra i modelli euristici più conosciuti e usati, nell'analisi qui esposta sono presenti: *Random Forest*, *Decision Tree semplice* e *J48*. Un *Decision Tree* o *albero decisionale* è una tecnica di apprendimento automatizzato che ha il vantaggio di poter essere applicato per tutti i tipi di attributi e che sfrutta un test univariato per classificare ciascun'osservazione, utilizzando tecniche statistiche per analizzare la relazione tra una variabile dipendente e altre indipendenti. Consiste in un insieme di

nodi e collegamenti e ripartisce il dataset in gruppi mutuamente esclusivi composti da soggetti il più possibile omogenei tra di loro. La classe con maggior proporzione all'interno del nodo decide la classificazione di tutte le osservazioni presenti in esso; nel nostro caso è stato utilizzato come criterio di divisione l'indice di eterogeneità di Gini.
La *Random Forest* è un'aggregazione di alberi decisionali, combinando le previsioni dello stesso modello su diversi dataset di training.

4.2 Modelli probabilistici

Usano la teoria Bayesiana e computano la probabilità a posteriori per classificare i record. Di questi modelli sono state applicate le versioni *Naive Bayes*, *Naive Bayes Tree* e *Bayes Network Classifier* (con metrica *tan-Tree Augmented Naive Bayes* e metrica *K2*). Hanno il vantaggio di essere semplici da implementare e permettono di ottenere buoni risultati nella maggior parte dei casi. Di contro assumono l'indipendenza condizionata alle classi e sono influenzati da input ridondanti.

4.3 Modelli di regressione

In particolare, è stato usato un *modello logistico robusto*, il quale si costruisce attraverso operazioni di missing imputation, analisi della collinearità, model selection, eliminazione di valori influenti e degli attributi con varianza quasi nulla. Essi hanno il vantaggio di essere applicabili a tutte le variabili, indipendentemente dal tipo (qualitativo o quantitativo). Nel caso indagato non è stato necessario agire sui valori mancanti in quanto assenti e sono stati utilizzati il modello logistico e quello logistico semplice.

4.4 Modelli di separazione

Sono state applicate due tecniche di apprendimento, *Support Vector Machines* e *Multi-layer Perceptron*, al fine di classificare le osservazioni. La prima consiste in un algoritmo di apprendimento supervisionato che ha lo scopo di ottenere la classificazione ottima mappando le osservazioni in uno spazio e cercando il miglior metodo di separazione. Le versioni eseguite sono: *pegasus*, *polykernel* e *puk*. La seconda tecnica è un tipo di *Neural Network* e consiste in un insieme di nodi, chiamati neuroni, che associano gli attributi classificativi agli attributi di classi. È complesso nelle sue tecniche di apprendimento e con tante variabili indipendenti fa fatica a convergere, come è successo nel caso in esame.

Al fine di ridurre il rischio di overfitting e underfitting nella valutazione del modello effettuata solo tramite accuracy, sono stati utilizzati due approcci che permettono di verificare la reale efficienza dei classificatori.

- Nell'approccio **Holdout** si partizionano i dati iniziali in due gruppi: uno a cui appartiene 1/3 delle osservazioni (33,333%) usato come validation, i restanti 2/3 delle osservazioni (il 66,667%) vengono ulteriormente divisi in 1/3 e 2/3, che rappresentano rispettivamente test set e training set[1]. È un metodo di splitting molto semplice, ma nella sua semplicità ha il difetto di essere sensibile alla scelta del dataset di test. In particolare, con un numero limitato di osservazioni nel dataset di partenza potrebbe convenire applicare metodologie più robuste, quale ad esempio la Cross Validation.
- Nell'approccio **Cross Validation** il dataset di partenza è diviso in j partizioni disgiunte e per ogni sottogruppo viene effettuato il test usando come training gli altri *j-1* gruppi, ripetendo questo procedimento per ogni partizione. È un miglioramento del metodo holdout: le stime di accuratezza calcolate in ogni passo *j* vengono sintetizzate in un unico valore tramite media e questa risulta essere una misura più veritiera dell'adattamento del modello. In particolar modo, essendo il campione a disposizione composto da poche osservazioni, le metriche per valutare la bontà dei modelli risultano molto sensibili al variare delle partizioni utilizzate per addestrare i modelli, quindi si ritiene più corretto utilizzare quest'approccio invece dell'Holdout.

Per valutare la bontà di un modello, cioè la sua capacità di prevedere e adattarsi ai dati, sono state utilizzate due metriche:

- **Accuracy**
  Questa misura si ottiene come:

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP}$$

  Dove *TN* e *TP* sono rispettivamente *True Negative* e *True Positive*, ossia le osservazioni correttamente classificate negativamente o positivamente; mentre *FN* e *FP* indicano *False Negative* e *False Positive*, cioè nel primo caso le osservazioni classificate come positive ma in realtà negative e nel secondo caso l'opposto. Rappresenta la probabilità che, utilizzando il classificatore su nuovi

[1] Le proporzioni riportate non sono le uniche possibili, ma sono quelle che sono state utilizzate per il presente studio.

record, essi vengano assegnati alla classe corretta. Ciò significa che valori elevati di Accuracy indicano una buona performance del modello e di conseguenza si sceglie come classificatore il modello a cui corrisponde il valore più alto.

È possibile costruire un intervallo di confidenza per l'Accuracy considerando la previsione corretta di nuove osservazioni distribuita come una *Bin(n,θ)*, dove *n* è il numero di osservazioni del dataset e θ l'accuracy vera ignota. Se il dataset ha un numero di record sufficientemente elevato è possibile usare il Teorema Centrale del Limite approssimando la distribuzione ad una Normale e costruendo così il seguente intervallo di confidenza:

$$\left(\frac{\left(acc + \frac{Z^2_{1-\frac{\alpha}{2}}}{2N} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{acc}{N} - \frac{acc^2}{N} + \frac{Z^2_{1-\frac{\alpha}{2}}}{4N^2}} \right)}{1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{N}}, \frac{\left(acc + \frac{Z^2_{1-\frac{\alpha}{2}}}{2N} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{acc}{N} - \frac{acc^2}{N} + \frac{Z^2_{1-\frac{\alpha}{2}}}{4N^2}} \right)}{1 + \frac{Z^2_{1-\frac{\alpha}{2}}}{N}} \right)$$

Figure 5: Intervallo di confidenza per l'Accuracy

- **Curva ROC**

  È uno strumento grafico che è in grado di mostrare contemporaneamente i valori di *Sensitivity* e *Specificity* assunti da ciascun modello. Queste quantità sono due misure di performance individuate come:

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN+TP}{TN+FP}$$

  Il grafico restituito presenta sull'asse delle ascisse il tasso di falsi positivi *(1-Specificity)* e sulle ordinate il tasso di veri positivi *(Sensitivity)*. La performance viene valutata sulla base dell'area sottostante la curva ROC, definita *AUC (Area Under Curve)*, la quale traccia la probabilità di un risultato vero positivo in funzione della probabilità di un risultato falso positivo per tutte le possibili soglie. La curva ideale si avvicina all'angolo in alto a sinistra, cioè ha area tendente a 1, mentre il caso peggiore si ha quando la curva coincide con la diagonale.

5 ANALISI

Sono state effettuate due diverse analisi a seconda dello scopo perseguito.

5.1 Prima domanda di ricerca

Una prima analisi è stata dedicata allo studio delle variabili che risultano più influenti sulle performance scolastiche. La variabile target G3 è stata resa dicotomica dividendo le valutazioni maggiori di 13 (con 13 compreso) e minori di 13. È stata scelta questa suddivisione al fine di concentrarsi sugli studenti che non sono solo sufficienti, ma hanno ottenuto dei voti considerati buoni nel sistema Portoghese. Quindi l'obiettivo non è solo evincere quali sono i fattori che condizionano l'essere promossi, ma quali incidono sul superare il corso con un buon voto. Sono stati implementati diversi modelli con lo scopo di classificare la variabile target ed è stata valutata l'Accuracy tramite Holdout. Al fine di confrontare le stime ottenute con tale approccio su test e validation si è deciso di affiancare l'intervallo di confidenza sull'Accuracy sulla prima partizione alla stima puntuale risultante dalla seconda partizione (Figura 6).
Si noti che il range di Accuracy del Validation set varia da 0.61 (*Decision Tree*) a 0.72 (*SLog*).

Per ogni famiglia di modelli presentata nel capitolo precedente si è deciso di analizzare il migliore in termini di performance, escludendo la categoria dei modelli probabilistici che in generale non ha ottenuto buoni risultati. Pertanto i tre modelli approfonditi sono: *Simple Logistic*, *Random Forest* e *Support Vector Machines Poly*.

Per il primo modello, il **Simple Logistic**, è stato necessario effettuare inizialmente delle tecniche di preprocessing. L'analisi delle correlazioni ha messo in evidenza in particolare due coppie di variabili che hanno una correlazione media: *Medu-Fedu* con correlazione 0.647, ossia il livello di educazione dei genitori, e *Walc-Dalc* con correlazione 0.617, ossia l'assunzione di alcohol durante la settimana e nel weekend. Si è scelto di non escludere nessuno dei quattro attributi.
Per la selezione degli attributi è stato utilizzato un *Decision Tree J48 pruned*, il quale ad ogni impostazione del seme selezionava variabili differenti, a causa del basso numero di osservazioni rispetto alla numerosità delle covariate. Per questo motivo si è deciso di mantenere tutte le variabili esplicative.

Il modello **Random Forest** in letteratura è conosciuto come uno tra i più performanti, non necessita di pre-processing ed è facile da implementare. La particolarità di questo classificatore è che per effettuare gli split non si considerano tutte le variabili, ma solo un campione casuale di *k* attributi fissato. Questa strate-

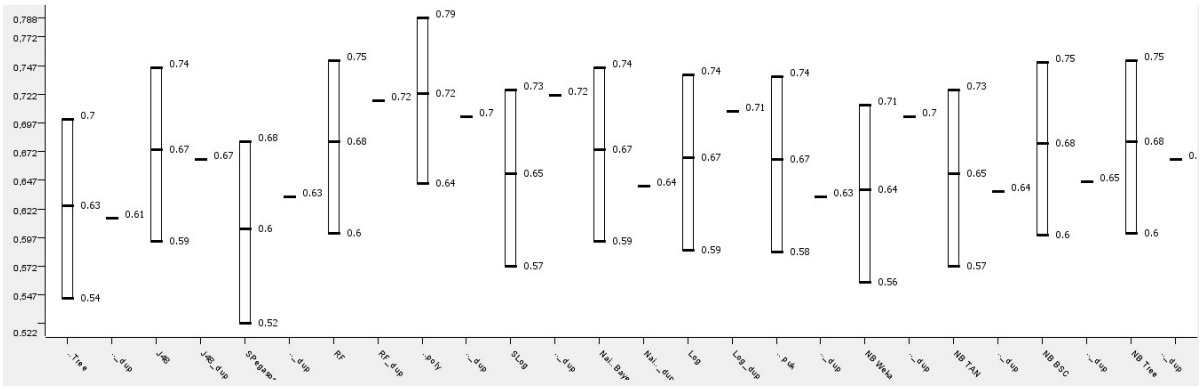

Figure 6: Confronto IC (test) vs puntuale (validation)

gia permette al modello di essere robusto rispetto alla collinearità e di ottenere stime più stabili nella valutazione dell'Accuracy, come si era notato precedentemente. Sono inoltre robuste al rumore, tutte le osservazioni hanno la stessa possibilità di influenzare ciascun albero prodotto, perciò i valori outlier hanno un effetto minore sui modelli singoli rispetto alla previsione generale. Calcolando l'importanza delle variabili nei 3 split eseguiti dal modello si osserva che: nel primo le variabili con importanza più elevata sono *failures*, *Medu* e *school*; nel secondo split la variabile con importanza più elevata si conferma essere *failures*, seguita da *Fjob* e *higher*; nel terzo split per ordine di importanza gli attributi sono *Fjob*, *age* e *reason* a pari merito con *absences*.

Il modello **Support Vector Machines** (SVM) si basa sul concetto di separazione, avendo lo scopo di trovare il limite di decisione lineare ottimale per separare le classi del target. Il metodo *SVM* non funziona esclusivamente con problemi di separazione lineare ma è adatta anche nel caso non lineare: si applica una funzione al target in modo da trasformare il problema non-lineare in lineare e identificare il limite di decisione lineare ottimale al fine di separare le classi del target.

Poichè il metodo Holdout è sensibile alla partizione iniziale e il campione di dati a dispozione ha bassa numerosità si è valutata anche l'adozione dell'approccio Cross Validation su tutti i modelli.

I tre modelli citati sono tra i più performanti rispetto all'Accuracy, come evidenziato dalla figura seguente (figura 7), che considera solo i classificatori che hanno ottenuto i risultati migliori. Sono riportati i livelli di accuracy raggiunti sul Validation test con Holdout e sull'intero campione con Cross Validation.

Dal grafico è possibile notare che sia con l'approccio

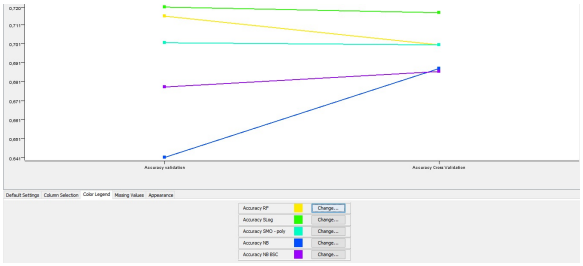

Figure 7: Confronto valori di Accuracy

Holdout che con la Cross Validation, il modello migliore è il Simple Logistic, che mantiene il livello di Accuracy intorno a 0.72.

Si può osservare che il secondo modello per qualità di performance è Random Forest con un valore di Accuracy superiore allo 0.70 sia sul Validation set sia con Cross Validation, coerentemente rispetto a quanto detto nel paragrafo dedicato. La Random Forest valutata con Cross Validation ha lo stesso rendimento del modello Support Vector Machines nella versione Polynomial Kernel (Poly).

Dei modelli più performanti si riportano anche le curve ROC, tra le quali non si distingue una curva uniformemente migliore delle altre.

5.2 Seconda domanda di ricerca

La seconda analisi effettuata si è sviluppata supponendo di voler implementare un metodo di assegnazione di borse di studio per studenti meritevoli. Non si considerano più studenti con votazioni medio-elevate (maggiori di 13/20) e con valutazioni basse o insufficienti, ma si è diviso il target in tre livelli:

- Livello *"basso"* con valutazioni finali da 0 a 11 (corrispondente al 46.4% nel campione);

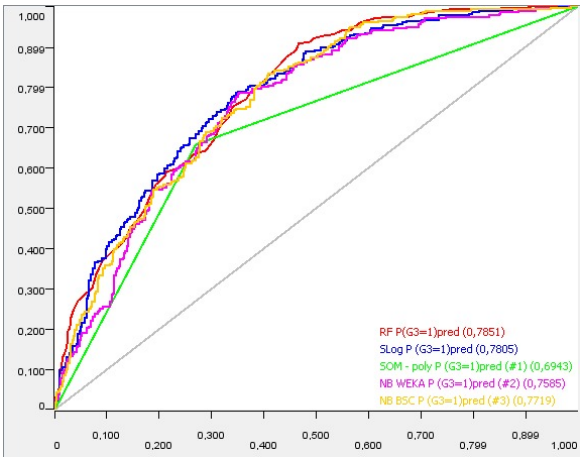

Figure 8: Curve ROC

- Livello *"intermedio"* con valutazioni finali da 12 a 15 (corrispondente al 41%);
- Livello *"alto"* con valutazioni finali da 16 a 20 (corrispondente al 12.6%).

Si è deciso di impostare la soglia dell'eccelenza a 16, nonostante siano evidenti i problemi di sbilanciamento del campione a dispozione. Infatti, aumentando la soglia a 17, più rappresentativa per l'obiettivo posto, la percentuale di studenti che hanno un voto oltre la soglia sarebbe ridotta al 7%. D'altro canto, una soglia inferiore per la determinazione di studenti meritevoli avrebbe ridotto la valenza di dominio della domanda di ricerca, includendo studenti non specificatamente meritevoli.

Inizialmente, è stato utilizzato un approccio simile a quanto adottato per la *Domanda di Ricerca 1*, opportunamente adattato per rispondere all'esigenza di una classificazione multiclasse. In particolare, poiché in questo nuovo scenario la variabile target *G3* non è più dicotomica, non è possibile implementare il modello logistico. Tutti gli altri classificatori presentati sono invece stati applicati anche in questa seconda domanda di ricerca.
In particolare, ci si è avvalsi dei seguenti modelli:

- Decision Tree
- J48
- RandomForest
- SMO Poly
- SMO Puc
- MLP

Le performance ottenute dai modelli applicati sono riportate di seguito. In particolare, in figura 9 è riportato il confronto tra le curve ROC ottenute con il metodo Cross Validation, che mostra 2 fattori rilevanti:

1. La Random Forest risulta essere il modello con la *AUC (Area Under Curve)* maggiore, pari a 0.7301.
2. Le perfromance generali ottenute non sono particolarmente rilevanti, a causa presumibilmente del ridotto numero di osservazioni a disposizione.

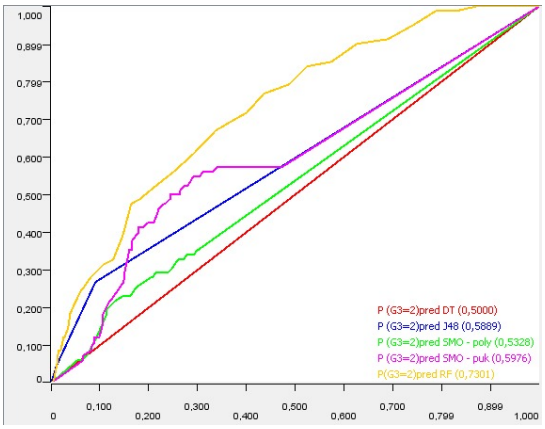

Figure 9: Curve ROC con Cross Validation

L'analisi dei risultati ottenuti ha portato a prendere in considerazione l'adozione di una matrice dei costi al fine di indirizzare il sistema, penalizzando maggiormente classificazioni particolarmente errate (classificare come *"basso"* uno studente che andrebbe classificato come *"alto"* e viceversa.
Differenti elaborazioni, reiterate modificando i valori dei costi, hanno portato alla definizione della matrice riportata in tabella 3.
Da un punto di vista di attenzione al dominio si è ipotizzato il caso di un ente disposto a erogare borse di studio a studenti di alto e medio livello, a supporto del loro percorso formativo. In tale scenario, si è arrivati alla definizione dei seguenti profili di costo:

- Costo negativo (-25) in caso di corretta classificazione, in quanto il sistema sta supportando l'ente in una classificazione che premia la meritocrazia, con conseguente beneficio (o *costo negativo*) sociale
- Costo pari a 20 in caso di misclassificazioni di "lieve entità" (tra classe 0 e 1 e tra classe 1 e 2). In questo caso, il costo può essere inteso come:
  - Un effettivo costo economico, nel caso in cui venga erogata una borsa di studio ad uno studente non adeguatamente meritevole
  - Un costo sociale, nel caso in cui non venga erogata (o venga erogata parzialmente) una

		Livello Stimato		
		Basso	Medio	Alto
Livello	Basso	-25	20	50
Effettivo	Medio	20	-25	20
	Alto	50	20	-25

Table 3: Matrice dei costi ipotizzata

borsa di studio ad uno studente maggiormente meritevole

- Costo pari a 50 in caso di misclassificazioni di "rilevante entità" (tra classe 0 e 2). Il criterio dietro tale scelta è il medesimo di quanto presentato per il costo pari a 20, ma è stato adeguatamente riproporzionato sulla base dell'entità della misclassificazione:
  - Assegnazione di una borsa di studio di rilevante entità ad uno studente assolutamente non meritevole, con conseguente perdita di capitale
  - Mancata assegnazione di una borsa di studio ad uno studente altamente meritevole, con conseguente costo sociale

In questo scenario, il modello Logistic mostra le performance migliori, con un'accuracy che si attesta al 64,7% (rif: 10).

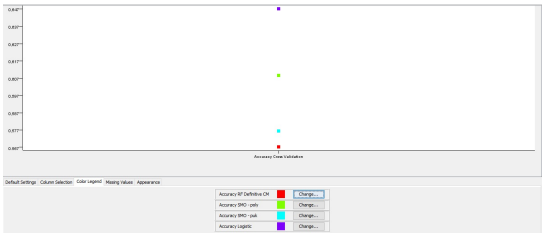

Figure 10: Accuracy valutata su Cross Validation con Matrice dei Costi

L'analisi delle curve ROC per i modelli valutati conferma quanto ipotizzato nello scenario precedente, senza la valutazione della Matrice dei Costi: la *AUC* mantiene valori massimi piuttosto ridotti, che si attestato poco sotto il 60%, nonostante la valutazione sia stata fatta applicando il metodo CrossValidation.

6 RISULTATI E CONCLUSIONI

Per quanto riguarda la prima domanda di ricerca, ovvero cercare di capire quali varibili contribuissero per uno studento ad ottenere una valutazione medio/alta, piuttosto che medio/bassa, il modello che ha ottenuto complessivamente dei punteggi migliori per quanto riguarda la capacità di predire il target,

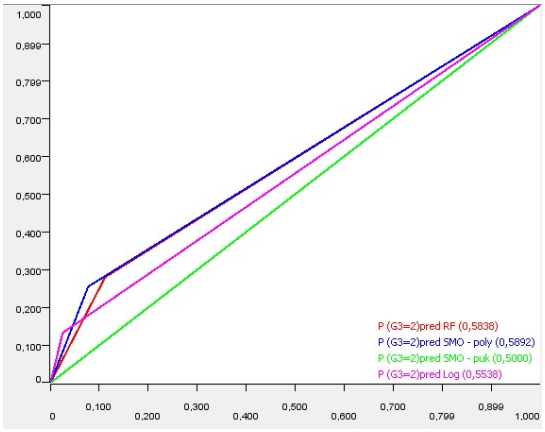

Figure 11: Curve ROC con Cross Validation con Matrice dei Costi

sia a livello di Accuracy, sia di area sotto la curva ROC è stato il modello Simple Logistic, il quale evidenziava come più determinanti le seguenti variabili: *failures, Medu, school, schoolsup, studytime, higher, Dalc e absences*. Tenendo in considerazione il contesto di analisi, si può ritenere coerente quanto identificato dal modello poiché sono tutte variabili che possono verosimilmente influenzare i rendimento scolastico di un alunno, in quanto: *Medu* contribuisce alla definizione del background culturale dello studente, che può essere determinante nelle opportunità e nell'educazione dello stesso; *failures*, *absences*, *higher* e *studytime* possono essere interpretate come indicatore sia dell'impegno dello studente, sia della sua volontà e consapevolezza di aver intrapreso un certo percorso di studi.

Per quanto riguarda la seconda domanda di ricerca, ovvero cercare di implementare un metodo di assegnazione di borse di studio per studenti meritevoli, la variabile target è stata divisa in 3 livelli in base alle performance ottenute dagli studenti: basso, intermedio e alto. Per trovare il modello migliore sono stati seguiti due approcci: il primo che consisteva nel cercare il modello che prevedesse in generale il più elevato numero di osservazioni corrette rispetto al target senza tener conto di quali stesse sbagliando ed il secondo che invece, grazie ad una matrice di costi, tenesse in considerazione di quali osservazioni stesse classificando sbagliate e quale peso avesse questo errore. Rispetto al primo approccio, il modello che ha ottenuto complessivamente una migliore performance nelle metriche considerate è il modello Random Forest. In relazione al secondo approccio in generale le performance ottenute dai modelli sono più o meno equivalenti e nessun modello presenta delle performance nettamente migliori rispetto agli altri. Il modello che per entrambi gli approcci

sembra restituire dei risultati migliori è il modello Random Forest che evidenzia come variabili più importanti: *failures, Medu, school, Walc e absences*. Anche in questo caso queste variabili sembrano poter essere direttamente collegate con il rendimento degli studenti. Notiamo questa volta la presenza della variabile *Walc* che rappresenta il consumo di alcolici durante il fine settimana.

In generale le performance ottenute dai vari modelli in entrambe le domande di ricerca non sono state molto elevate: si può supporre che ciò sia dovuto alla numerosità ridotta del dataset che non permette ai modelli di sfruttare al massimo le loro potenzialità. In futuro, oltre a raccogliere una maggiore quantità di osservazioni, si potrebbe pensare anche, rispetto alla seconda domanda di ricerca, di modificare la matrice di costi per renderla più fedele alla rappresentazione del fenomeno di assegnazione delle borse di studio.

REFERENCES

AICA, Aintec-Assinform, A. A. (2019). Osservatorio delle competenze digitali.

Cedefop (2019a). Real-time labour market information and skill requirements: Setting up the infrastructure for eu system.

Cedefop (2019b). Skills-ovate: Skills online vacancy analysis tool for europe.

Kaggle (2017). Student alcohol consumption - social, gender and study data from secondary school students.

Ministerio da ciencia, i. e. e. s. (2005). Decreto-lei n.o 42/2005.

OECD (2020). Pisa - programme for international student assessment. program presentation.

PISA (2003). Pisa (programme for international student assessment) 2003 - first outcomes.

PISA (2018). Pisa (programme for international student assessment) 2018 results - shaphot of students' performance in reading, mathematics and science.

Schwab, K. and Samans, R. (2016). Chapter 1: The Future of Jobs and Skills. *The Future of Jobs Report*.