Università Politecnica delle Marche

Dipartimento di Ingegneria dell'Informazione

Facoltà di Ingegneria Informatica e dell'Automazione

RELAZIONE BIG DATA ANALYTICS E MACHINE LEARNING

Progetto N° 62

Relatore:
Prof. Potena Domenico

Studenti:
Bernovschi Denis
Licci Giacomo

ANNO ACCADEMICO 2020/2021

Indice

1 Introduzione

L'analisi si è svolta sulla base di dati proveniente dall'Università Politecnica delle Marche, inerente alle nuove iscrizioni nei percorsi triennali. In particolare l'obbiettivo è analizzare il percorso di studi precedente e la correlazione con i percorsi universitari intrapresi. Infine si pone in analisi i risultati dei percorsi scelti.

1.1 Dataset Utilizzato

Dati i due dataset di partenza, in particolar modo: *libretti* e *iscritti*, sono stati verificati e analizzati, al fine di costruire un unico dataset che comprendesse solo valori corretti e consistenti. Altresì è fondamentale riportare alcune operazioni svolte per uniformare il dataset, come la normalizzazione del voto di maturità, la cancellazione di valori duplicati, cancellazione di record contenenti valori nulli o non veritieri. All'interno del progetto vedremo nel dettaglio tutte le operazioni implementate. Va osservato come le scelte adottate hanno l'obbiettivo di preservare molti più record possibili.

	PERS_ID	MATRICOLA	AA_ISCR_ID	CDS_ID_x	SESSO	AA_CONSEG_TITOLO	VOTO	LODE	BASE_VOTO	MIUR_SCUOLE_COD	Tipo_EDUSCOPIO	CDS_ID_y	AVG(VOTO)	sum(PESO_AD)	AA_SUP_ID	min(DATA_SUP)
0	29AE4E616F531F66A98F9323FE63DB77	00027EDDD55C51F668558C1BBC93852D	2020	10362	M	2020	74.0	0	100.0	PSPS01000G	Scientifico	10362	20.000000	12.0	2020	2021-04-12 00:00:00
1	2FD123EE6A3914B846CF2CC091016E84	0004D7F73A2E5D764603CD68DCDEC115	2019	10362	F	2016	61.0	0	100.0	PSPS020006	Scientifico Scienze applicate	10362	27.000000	16.0	2019	2020-01-30 00:00:00
2	C4E48C39CCF876F937A41C48923FCD7F	0010B15A6FE7B318A144E2B5C8A87CC5	2017	10204	M	2017	67.0	0	100.0	CHPS00301T	Scientifico	10204	22.000000	27.0	2017	2018-01-23 00:00:00
3	DB6CB9B6DBA3C630AFE95AADEC2B4C11	00122E576F0F42D3C800EB8F9BB6D35F	2016	10040	M	2016	62.0	0	100.0	APPS030005	Scientifico	10040	22.714286	60.0	2016	2017-02-13 00:00:00
4	0756DC8F726266FB88782DF74A80B516	00152A8901E4C527DF527DBFFAF73B42	2015	10179	F	2014	63.0	0	100.0	PDIS01300X	Classico	10179	27.384615	54.0	2015	2016-02-08 00:00:00

Figura 1: Dataset Filtrato e Unito

1.2 Analisi Base di Dati

Le due basi di dati proveniente dall'università politecnica delle marche, le quali riportano: la prima i risultati conseguiti tramite l'istruzione superiore secondaria, la seconda invece riporta i risultati universitari del primo anno accademico. Siamo andati ad analizzare in particolare la prima base di dati, ponendo particolare attenzione alla numerosità di studenti iscritti presso ciascuna tipologia di scuola superiore [1] (Fig. 2).

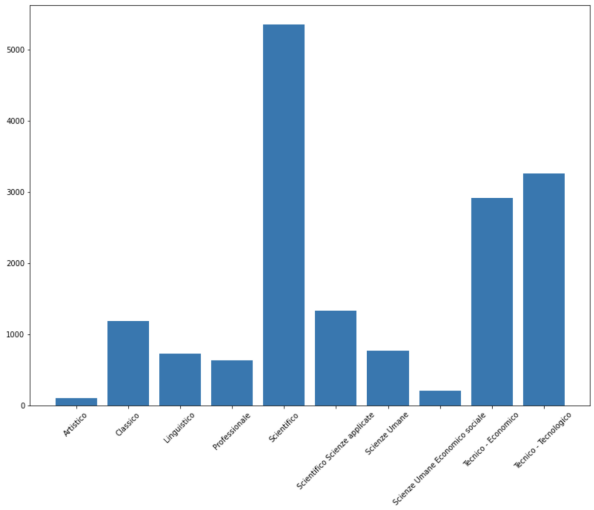


Figura 2: Analisi 1

[1] Per tipologia di scuola, facciamo riferimento **EDUSCOPIO**

2 Analisi

2.1 Analisi percorso intrapreso

Un'ulteriore analisi è stata svolta con la mission di trovare quelli che sono i percorsi universitari scelti sulla base della istruzione superiore secondaria conseguita. Siamo andati quindi ad analizzare per ciascuna tipologia di istituto il percorso universitario maggiormente scelto (Fig. 3). Qui riportiamo solo il caso dell'istituto tecnico-tecnologico, ma l'analisi è stata svolta su tutte le tipologie dell'elenco EDUSCOPIO

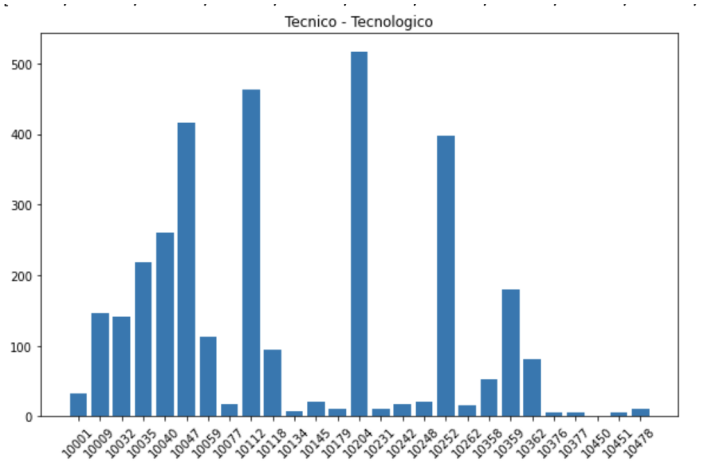


Figura 3: Analisi percorso intrapreso

2.2 Analisi Risultati Percorso Intrapreso

Al fine di evidenziare se il percorso intrapreso alla fine del conseguimento del titolo risulti più o meno indicato in base alla provenienza. Abbiamo analizzato dato il percorso intrapreso, il risultato di quest'ultimo per ciascuna facoltà e per ciascuna provenienza. Nell'analisi effettuata abbiamo analizzato sia la media dei CFU[2] conseguiti che la votazione media (Fig. 4).

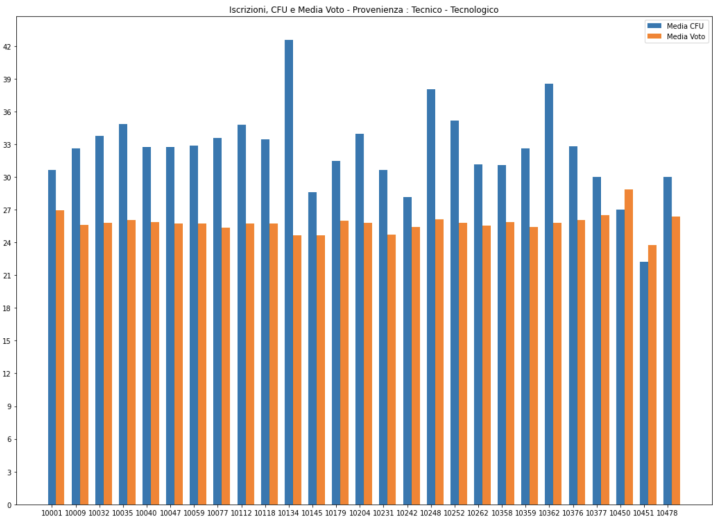


Figura 4: Analisi risultati percorso intrapreso

[2]CFU ⊨ Credito Formativo Universitario

Risultati tempistiche Sempre sulla falsa riga dell'analisi precedente siamo andati a valutare il percorso in termini di tempo per sostenere gli esami, in particolare abbiamo analizzato quello che è il tempo medio (espresso in mesi) per sostenere il primo esame (Fig. 5). Valori negativi e/o nulli sono dovuti al offset

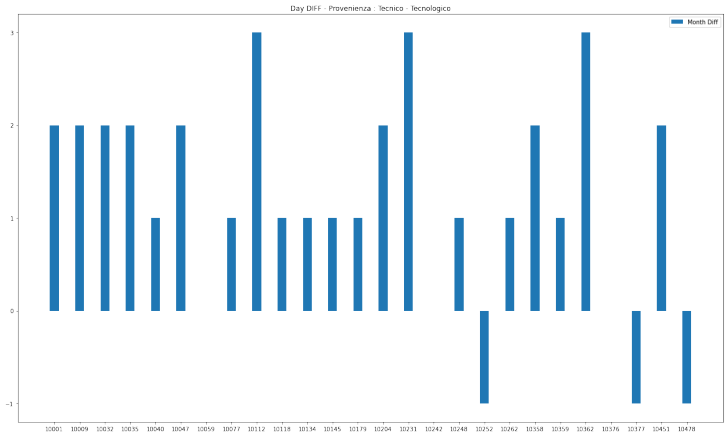


Figura 5: Analisi risultati percorso intrapreso pt.2

considerato, in particolare la data di riferimento utilizzata è quella del primo giorno della sessione invernale di esami: $01/08/A.A$

2.3 Analisi Voto e la possibile relazione con Carriera Universitaria

Data l'analisi delle tempistiche e sopratutto quella del valutazione del percorso universitario intrapreso, abbiamo analizzato se vi è una possibile relazione tra quello che è il risultato conseguito tramite l'istruzione superiore secondaria e i risultati universitari del percorso intrapreso. Abbiamo quindi valutato dato il voto del diploma di maturità i risultati degli studenti sia in termini di CFU che di Voto (Fig. 6). Come osservabile, non vi si può sottolineare una netta relazione tra il risultato conseguito con il diploma di maturità e il relativo rendimento universitario.

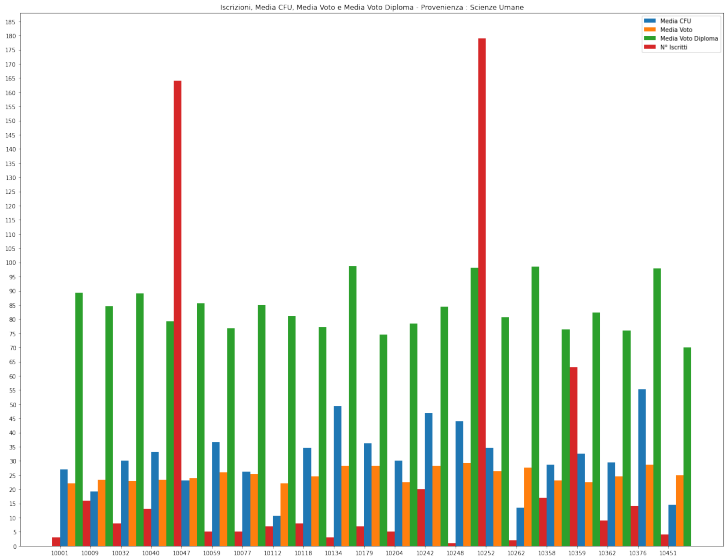


Figura 6: Analisi Voto e relazione con carriera universitaria

2.4 Analisi tempistiche conseguimento titolo e inizio Iscrizione

Abbiamo inoltre analizzato quanto tempo intercorre tra il conseguimento del titolo di Diploma di Maturità e l'inizio del percorso universitario. Come si può osservare, la maggioranza inizia il percorso universitario subito dopo il conseguimento del titolo, va però osservato come vi sono anche alcune persone che intraprendo questo percorso, a molti anni dal conseguimento del diploma (Fig. 7).

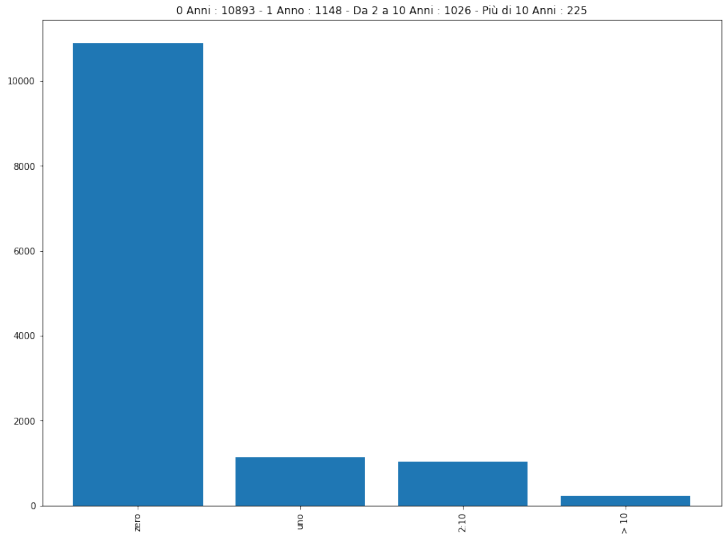


Figura 7: Analisi tempistiche conseguimento titolo e inizio Iscrizione

2.5 Regole di Associazione su Provenienza e Iscrizione

Inoltre abbiamo analizzato quelle che potrebbero essere le regole di associazione che ci permettono di identificare: data la provenienza la possibile iscrizione universitaria, e viceversa. Abbiamo quindi rappresentato il tutto tramite delle matrici di confusione, e le matrici riportanti la confidenza delle possibili regole (Fig. 8), (Fig. 9). La ricerca delle regole di associazione è stata fatta calcolando direttamente la confidenza senza tenere conto del supporto, dato che si tratta di un analisi verso specifiche facoltà e scuole di provenienza.

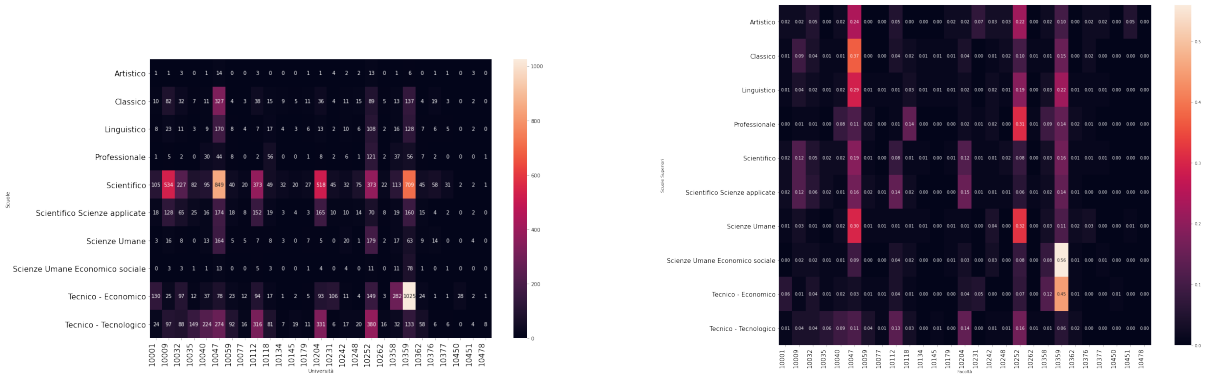
Nella prima analisi si mette in evidenza il percorso di istruzione superiore secondaria di ciascuna facoltà, nella maggior parte delle facoltà sono presenti studenti che hanno frequentato il Liceo Scientifico, a seguire l'istituto Tecnico - Economico e Tecnico - Tecnologico.

(a) Facoltà - Scuola di provenienza

(b) Confidenza

Figura 8: Data la facoltà scelta andiamo a capire quale è la provenienza scolastica

Per quanto riguarda la regola Scuola di Provenienza → Facoltà si possono notare le tendenze di scelta dei studenti alle facoltà sulla base del diploma conseguito. Abbiamo inoltre considerato al fine di ottenere un risultato più veritiero possibile, la numerosità degli studenti iscritti alle diverse tipologie di scuole superiori. E.g. Scuole come Artistico, Classico e Linguistico avrebbe causato valori non rappresentabili rispetto a scuole con elevata numerosità di studenti, come Scientifico e/o Tecnico-Tecnologico.



(a) Scuola di provenienza - Facoltà

(b) Confidenza

Figura 9: Data la provenienza scolastica andiamo a capire quale è la facoltà scelta

2.6 Analisi Cambiamenti di Percorso

Come ulteriore analisi fondamentale per analizzare a pieno la base di dati e fornire informazioni utili su quello che è la valutazione preventiva del percorso intrapreso, e quindi dimostrare la consapevolezza dei studenti nel scegliere il proprio percorso universitario, abbiamo svolto che quella che è l'analisi del comportamento degli studenti in difficoltà e/o di studenti non soddisfatti dalla scelta iniziale(Fig. 10). Analizzando il grafico è osservabile come la maggior parte dei studenti in difficoltà hanno cambiato corso,

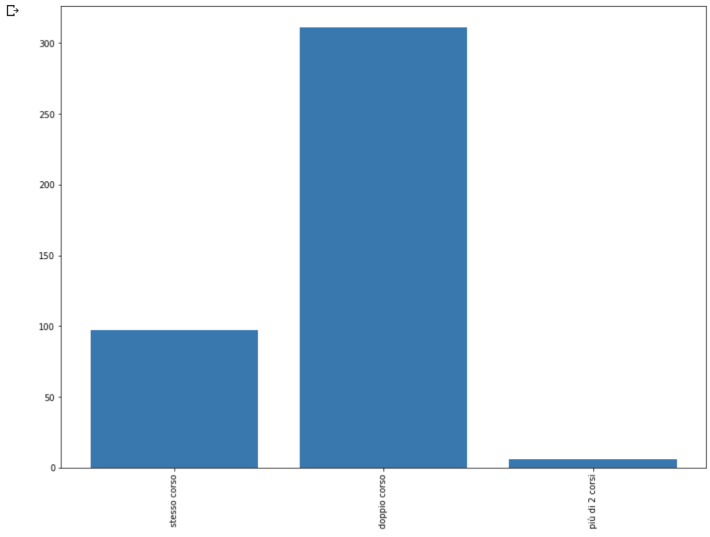


Figura 10: Analisi Cambiamenti di Percorso

è però altrettanto significativo la presenza di studenti che hanno deciso di iniziare nuovamente il medesimo percorso universitario intrapreso inizialmente.

3 Modello predittivo

Come ultimo step siamo andati ad analizzare quello che è un possibile modello predittivo in grado di predire data la scuola di provenienza, la fascia di voto o CFU acquisiti in cui si troverà lo studente al primo anno

3.1 Dataset Utilizzato

Riportiamo anche in questo caso, il dataset utilizzato già codificato con le etichette che è possibile visionare nella sezione 6.

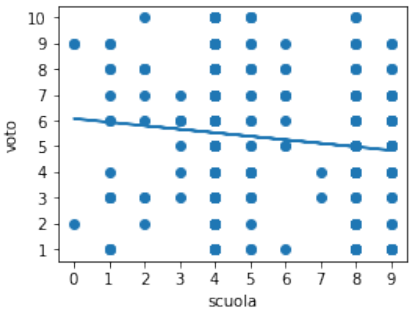
	matricola	scuola	cfu	voto	maturita
0	00027EDDD55C51F668558C1BBC93852D	4	10.0	1.0	74.0
1	0004D7F73A2E5D764603CD68DCDEC115	5	10.0	8.0	61.0
2	0010B15A6FE7B318A144E2B5C8A87CC5	4	20.0	3.0	67.0
3	00122E576F0F42D3C800EB8F9BB6D35F	4	60.0	3.0	62.0
4	00152A8901E4C527DF527DBFFAF73B42	1	50.0	8.0	63.0

Figura 11: Dataset utilizzato per i modelli predittivi

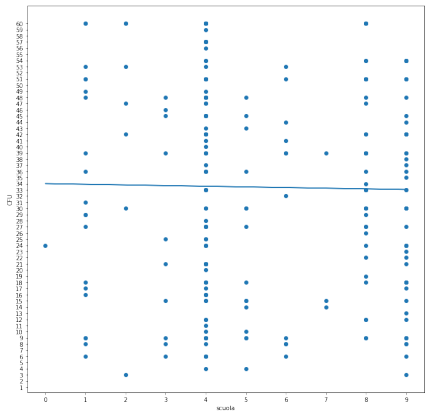
3.2 Regressione Lineare

Tramite la regressione lineare, data una relazione di dipendenza tra due variabili, si può cercare di prevedere il valore di una variabile in funzione del valore assunto dall'altra. Questo ha significato in senso stretto quando si ipotizza una relazione di causalità tra la variabile indipendente, su cui si agisce, e quella dipendente, su cui si vuole produrre un effetto. In particolare nel nostro caso abbiamo analizzato data la scuola il possibile voto o i possibili CFU conseguiti. Tramite il parametro Intercept abbiamo analizzato sia in termini di CFU che in termini di voto.

MODELLO	VOTO	CFU
REGRESSIONE LINEARE	25/26	34



(a) Regressione Lineare Voto-Scuola



(b) Regressione Lineare CFU-Scuola

Figura 12: Data la provenienza scolastica andiamo a capire quale è la facoltà scelta

3.3 Regressione Logistica

Si può considerare la regressione logistica come un metodo di classificazione rientrante nella famiglia degli algoritmi di apprendimento supervisionato. Avvalendosi di metodi statistici, la regressione logistica permette di generare un risultato che, di fatto, rappresenta una probabilità che un dato valore di ingresso appartenga a una determinata classe. Ecco quindi che nel nostro caso siamo andati ad analizzare le probabilità dato il titolo dell'istruzione superiore secondaria, il possibile rendimento del percorso universitario intrapreso. I risultati ottenuti rispecchiano quelli della Regressione Lineare, in particolare abbiamo un'elevata probabilità di ottenere 25/26 come votazione e di acquisire circa 34/35 CFU [3].

3.4 K-NN

Un algoritmo K-Nearest-Neighbor, spesso abbreviato K-NN, è un approccio alla classificazione dei dati che stima la probabilità che un punto (dato) sia membro di un gruppo o dell'altro a seconda del gruppo in cui si trovano i punti (dati) più vicini. Il K-Nearest-Neighbor è un esempio di algoritmo "lazy learner", che significa che non costruisce un modello usando il set di addestramento finché non viene eseguita una query del set di dati.

Implementazione Per quanto concerne l'implementazione abbiamo utilizzato la libreria **sklearn** ed in particolare il KNeighborsClassifier. Abbiamo testato il modello attraverso un fine tuning del parametro **n_neighbors** e diverse tipologie di splitting dei dati.

3.4.1 Risultati "CFU"

Tabella 1: Classification Report CFU

CFU	**precision**	**recall**	**f1-score**	**support**
0.0	0.16	0.12	0.14	34
10.0	0.22	0.06	0.09	34
20.0	0.44	0.10	0.17	39
30.0	0.22	0.46	0.29	50
40.0	0.22	0.19	0.21	42
50.0	0.25	0.42	0.31	38
60.0	0.43	0.15	0.22	20
accuracy			0.23	257
macro avg	0.28	0.21	0.20	257
weighted avg	0.27	0.23	0.21	257

[3] CFU $\models$ Crediti Formativi Universitari

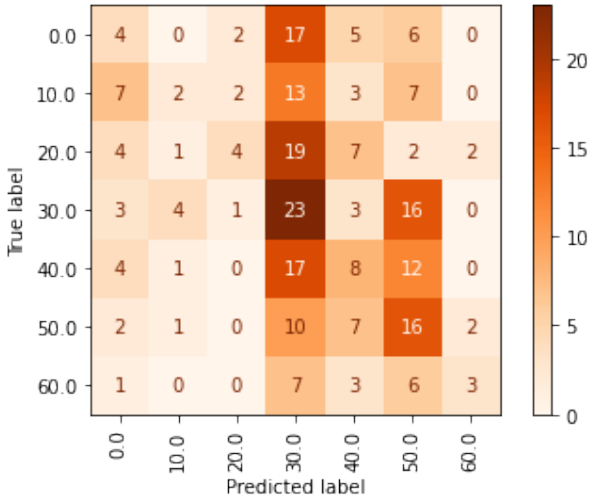


Figura 13: Matrice di Confusione relativa alla previsione dei CFU

3.4.2 Risultati "Voto"

Tabella 2: Classification Report Voto

VOTO	**precision**	**recall**	**f1-score**	**support**
1.0	0.18	0.47	0.26	30
2.0	0.00	0.00	0.00	20
3.0	0.00	0.00	0.00	23
4.0	0.19	0.11	0.14	28
5.0	0.17	0.13	0.15	31
6.0	0.10	0.10	0.10	31
7.0	0.10	0.07	0.08	28
8.0	0.07	0.07	0.07	27
9.0	0.17	0.17	0.17	23
10.0	0.25	0.31	0.28	16
accuracy			0.14	257
macro avg	0.12	0.14	0.12	257
weighted avg	0.12	0.14	0.12	257

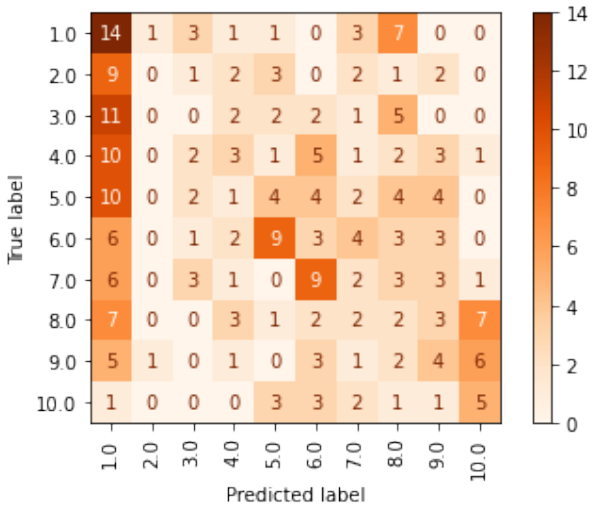


Figura 14: Matrice di Confusione relativa alla previsione dei Voto

3.5 SVM

L'obiettivo dell'algoritmo Support Vector Machine è quello di trovare un iperpiano in uno spazio N-dimensionale (N - il numero di caratteristiche) che classifica distintamente i dati in ingresso. Per separare le due classi di dati, ci sono molti iperpiani possibili che potrebbero essere scelti. Il nostro obiettivo è quello di trovare un piano che abbia il margine massimo, cioè la distanza massima tra i dati di entrambe le classi. Massimizzare la distanza di margine fornisce un certo rinforzo in modo che i dati futuri possano essere classificati con più fiducia. Gli iperpiani sono limiti decisionali che aiutano a classificare i dati.

Implementazione Per l'implementazione abbiamo utilizzato la libreria **sklearn** ed in particolare il pacchetto **SVM**. Come configurazione del modello abbiamo adottato:

```
clf = svm.SVC(kernel='poly', degree=3, C = 1.0, decision_function_shape='ovo')
```

3.5.1 Risultati "Voto"

Tabella 3: Classification Report Voto

Voto	**precision**	**recall**	**f1-score**	**support**
1.0	0.16	0.70	0.26	30
2.0	0.00	0.00	0.00	20
3.0	0.00	0.00	0.00	23
4.0	0.00	0.00	0.00	28
5.0	0.11	0.10	0.10	31
6.0	0.33	0.13	0.19	31
7.0	0.00	0.00	0.00	28
8.0	0.05	0.07	0.06	27
9.0	0.24	0.43	0.31	23
10.0	0.00	0.00	0.00	16
accuracy			0.16	257
macro avg	0.09	0.14	0.09	257
weighted avg	0.10	0.16	0.10	257

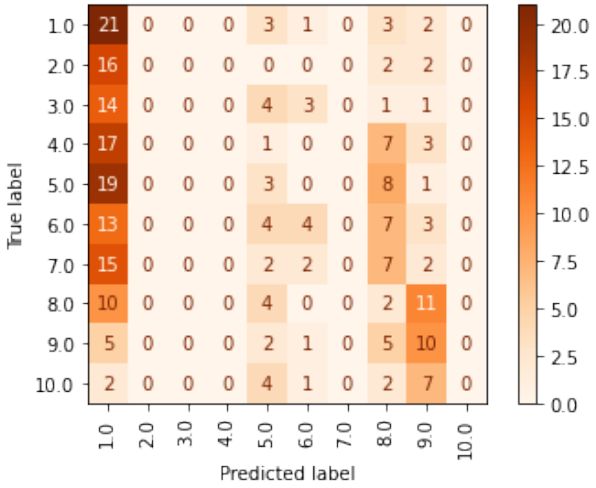


Figura 15: Matrice di Confusione relativa alla previsione dei Voto

3.5.2 Risultati "CFU"

Tabella 4: Classification Report CFU

CFU	**precision**	**recall**	**f1-score**	**support**
0.0	0.00	0.00	0.00	34
10.0	0.00	0.00	0.00	34
20.0	0.25	0.03	0.05	39
30.0	0.19	0.78	0.30	50
40.0	0.00	0.00	0.00	42
50.0	0.19	0.24	0.21	38
60.0	0.00	0.00	0.00	20
accuracy			0.19	257
macro avg	0.09	0.15	0.08	257
weighted avg	0.10	0.19	0.10	257

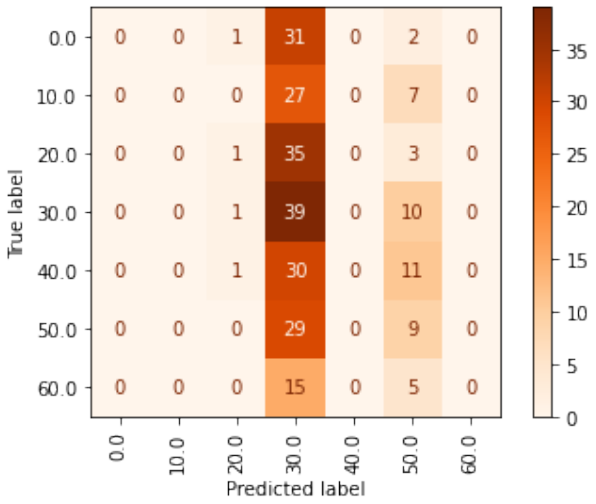


Figura 16: Matrice di Confusione relativa alla previsione dei CFU

3.6 Decision Tree

L'apprendimento del Decision Tree o l'induzione degli alberi di decisione è uno degli approcci di modellizzazione predittiva utilizzati in statistica, data mining e machine learning. Utilizza un albero di decisione (come modello predittivo) per passare dalle osservazioni su un elemento (rappresentato nei rami) alle conclusioni sul valore obiettivo dell'elemento (rappresentato nelle foglie). I modelli ad albero in cui la variabile target può assumere un insieme discreto di valori sono chiamati alberi di classificazione; in queste strutture ad albero, le foglie rappresentano etichette di classe e i rami rappresentano congiunzioni di caratteristiche che portano a quelle etichette di classe.

Implementazione Per l'implementazione abbiamo utilizzato la libreria **sklearn** ed in particolare il pacchetto **tree**. Come configurazione del modello abbiamo adottato:

```
clf = tree.DecisionTreeClassifier(
criterion='gini',
splitter='best',
max_depth=10,
min_samples_split=2,
min_samples_leaf=4,
min_weight_fraction_leaf=0.0,
max_features=1,
random_state=42,
max_leaf_nodes=None,
class_weight=None,
ccp_alpha=0.0)
```

3.6.1 Risultati "CFU"

Tabella 5: Classification Report CFU

CFU	**precision**	**recall**	**f1-score**	**support**
0.0	0.11	0.09	0.10	34
10.0	0.11	0.09	0.10	34
20.0	0.18	0.08	0.11	39
30.0	0.17	0.30	0.22	50
40.0	0.17	0.19	0.18	42
50.0	0.26	0.26	0.26	38
60.0	0.45	0.25	0.32	20
accuracy			0.18	257
macro avg	0.21	0.18	0.18	257
weighted avg	0.19	0.18	0.18	257

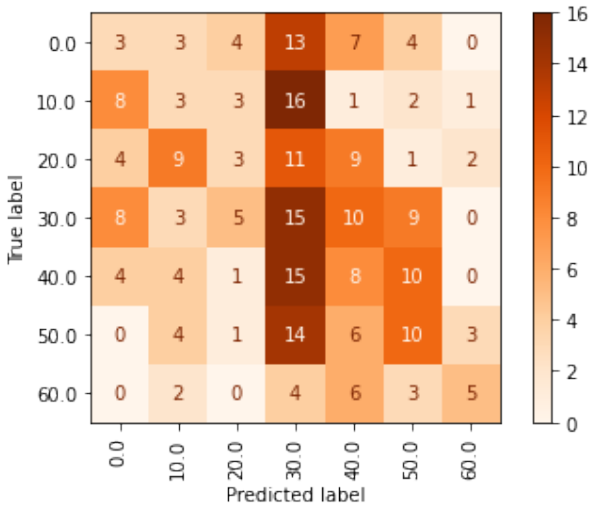


Figura 17: Matrice di Confusione relativa alla previsione dei CFU

3.6.2 Risultati "Voto"

Tabella 6: Classification Report Voto

	precision	**recall**	**f1-score**	**support**
1.0	0.19	0.50	0.27	30
2.0	0.14	0.05	0.07	20
3.0	0.00	0.00	0.00	23
4.0	0.23	0.18	0.20	28
5.0	0.11	0.13	0.12	31
6.0	0.33	0.10	0.15	31
7.0	0.14	0.14	0.14	28
8.0	0.14	0.15	0.15	27
9.0	0.15	0.13	0.14	23
10.0	0.25	0.31	0.28	16
accuracy			0.17	257
macro avg	0.17	0.17	0.15	257
weighted avg	0.17	0.17	0.15	257

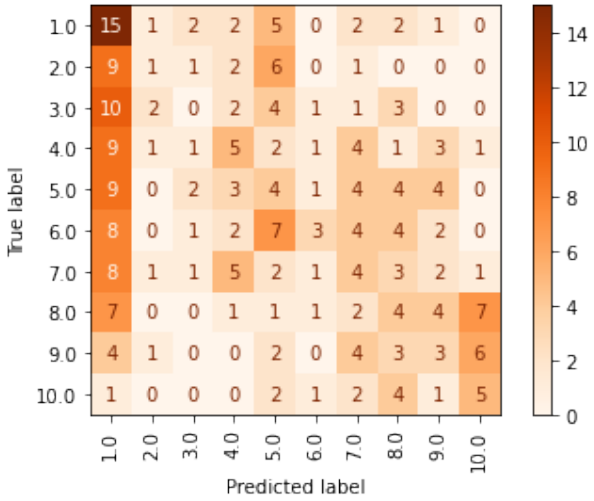


Figura 18: Matrice di Confusione relativa alla previsione dei Voto

3.7 Osservazioni e Conclusioni

Split Utilizzati Riportiamo qui brevemente i split utilizzati nel corso del nostro processo di analisi

```
### -------SPLIT SEMPLICE ------###
#X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state
    =42, shuffle=True, stratify=y)
```

```
### -------SPLIT StratifiedShuffleSplit ------###
#sss = StratifiedShuffleSplit(n_splits=5, test_size=0.5, random_state=0)
#sss.get_n_splits(X, y)
#print(sss)
#StratifiedShuffleSplit(n_splits=5, random_state=0)
#for train_index, test_index in sss.split(X, y):
#     print("TRAIN:", train_index, "TEST:", test_index)
#     X_train, X_test = X[train_index], X[test_index]
#     y_train, y_test = y[train_index], y[test_index]
```

```
### -------SPLIT StratifiedKFold ------###
skf = StratifiedKFold(n_splits=50, random_state=42, shuffle=True)
skf.get_n_splits(X, y)
print(skf)
for train_index, test_index in skf.split(X, y):
     #print("TRAIN:", train_index, "TEST:", test_index)
     X_train, X_test = X[train_index], X[test_index]
     y_train, y_test = y[train_index], y[test_index]
```

```
### -------SPLIT RepeatedStratifiedKFold ------###
skf = RepeatedStratifiedKFold(n_splits=50, random_state=42, n_repeats=10)
skf.get_n_splits(X, y)
```

Conclusioni Come possiamo osservare i modelli citati precedentemente non sono in grado di classificare correttamente e quindi di conseguenza predire con un'accuratezza elevata, il motivo principale riscontrato è dovuto alla elevata sparsità dei dataset di partenza. All'interno dello sviluppo del progetto sono stati testati diversi algoritmi di split al fine di ridurre questa problematica, ed ottenere migliori risultati, nonostante ciò non vi si registrano ulteriori miglioramenti ai risultati già riportati.

4 Modello Predittivo - Rapid Miner

Al fine di verificare ulteriormente i risultati del modello predittivo riportati nella sezione precedente, abbiamo ricreato quelli che sono i modelli citati nel programma **Rapid Miner**, in particolar modo sono stati reimplementati entrambi i modelli predittivi utilizzando K-NN e Alberi decisionali come algoritmi di classificazione. Lo stesso dataset utilizzato precedentemente è stato caricato ed è stata definita la classe di appartenenza di ciascuna tupla all'interno di un sotto processo (Fig. 19).

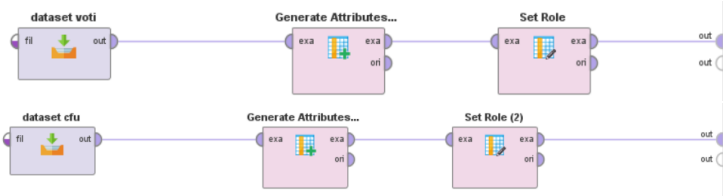


Figura 19: Sotto processo per le due classificazioni

4.1 K-NN

Per trovare il parametro K è stato utilizzato un blocco Validation contenente il modello e due blocchi Performance per misurare l'errore di Training e Test, salvato poi all'interno di un Log. Il blocco è stato poi racchiuso all'interno di un Loop Parameters che esegue un numero predefinito di volte il blocco Validation variando il parametro k dell'algoritmo di classificazione (Fig. 20).

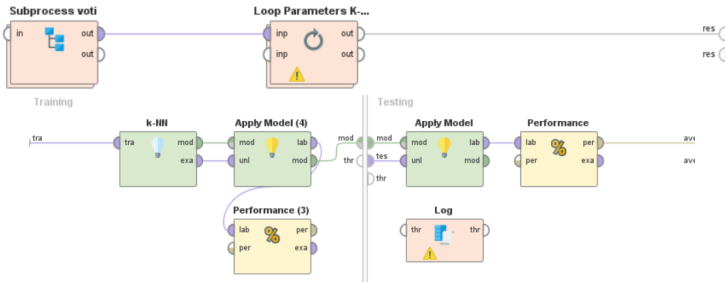


Figura 20: Loop Parameters e blocco Cross Validation utilizzati per trovare il paramentro k migliore

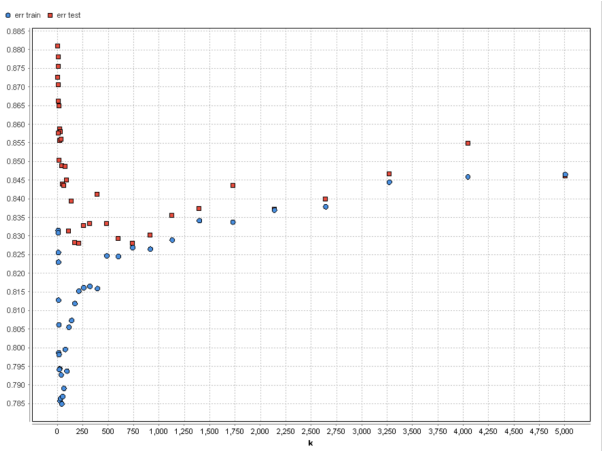


Figura 21: Errore di classificazione di training e test voti

Nonostante l'alto errore di classificazione il migliori parametri k per le due classificazioni sono rispettivamente 2500 per la classificazione con voti (Fig. 21) e 1750 per la classificazione con cfu (Fig. 22).

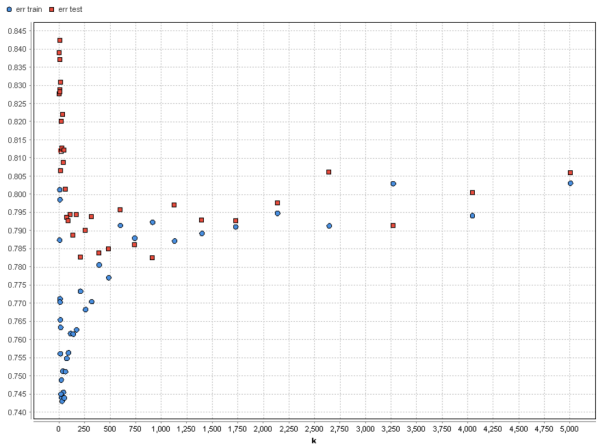


Figura 22: Errore di classificazione di training e test cfu

A questo punto viene eseguito un blocco Cross Validation per ricavare la matrice di confusione e il grafico risultante della classificazione.

4.1.1 Risultati "Voto"

accuracy: 15.98% +/- 0.59% (micro average: 15.98%)

	true 1	true 8	true 3	true 6	true 2	true 7	true 4	true 5	true 9	true 10	class pre...
pred. 1	1028	390	624	637	568	534	636	682	287	140	18.60%
pred. 8	14	23	13	21	10	20	19	35	22	15	11.98%
pred. 3	0	0	0	0	0	0	0	0	0	0	0.00%
pred. 6	53	86	71	95	52	91	86	82	65	26	13.44%
pred. 2	0	0	0	0	0	0	0	0	0	0	0.00%
pred. 7	11	16	8	14	10	16	19	23	20	9	10.96%
pred. 4	0	0	0	0	0	0	0	0	0	0	0.00%
pred. 5	299	342	282	388	237	380	352	410	279	162	13.09%
pred. 9	138	452	172	373	138	384	249	319	486	469	15.28%
pred. 10	0	0	0	0	0	0	0	0	0	0	0.00%
class rec...	66.62%	1.76%	0.00%	6.22%	0.00%	1.12%	0.00%	26.43%	41.93%	0.00%	

Figura 23: Matrice di confusione

Figura 24: Classi predette

4.1.2 Risultati "CFU"

accuracy: 20.79% +/- 0.74% (micro average: 20.79%)

	true 10	true 20	true 60	true 50	true 40	true 30	true 0	class precision
pred. 10	0	0	0	0	0	0	0	0.00%
pred. 20	0	0	0	0	0	0	0	0.00%
pred. 60	0	0	0	0	0	0	0	0.00%
pred. 50	364	462	399	769	602	638	290	21.82%
pred. 40	116	128	76	149	179	192	110	18.84%
pred. 30	973	1140	531	910	1170	1443	997	20.14%
pred. 0	255	212	27	76	139	248	287	23.07%
class recall	0.00%	0.00%	0.00%	40.39%	8.56%	57.24%	17.04%	

Figura 25: Matrice di confusione

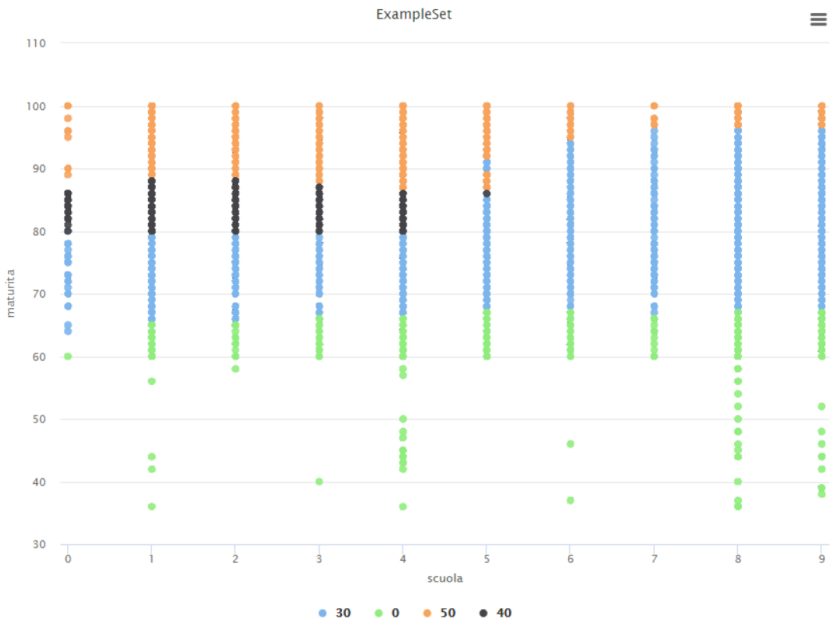


Figura 26: Classi predette

4.2 Decision Tree

Anche per l'algoritmo Decision Tree è stato costruito un processo per trovare la miglior profondità dell'albero con il minor errore di classificazione utilizzando sempre un Loop Parameters dove ogni volta veniva cambiata la profondità dell'albero (Fig. 27).

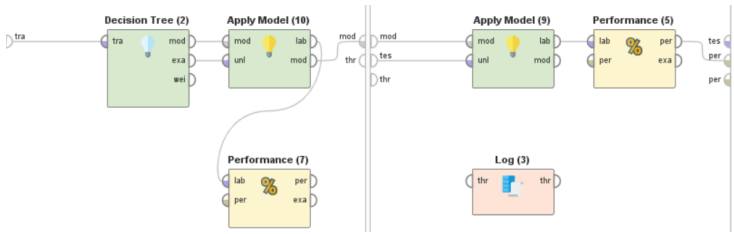


Figura 27: Blocco Cross Validation

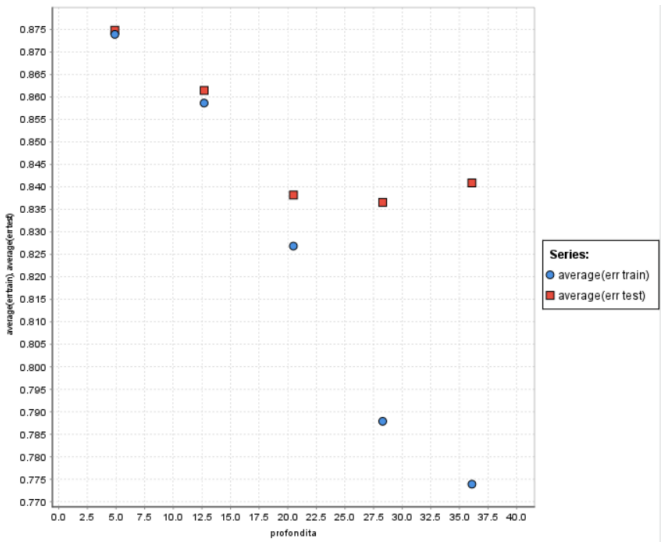


Figura 28: Errore di classificazione di training e test voti

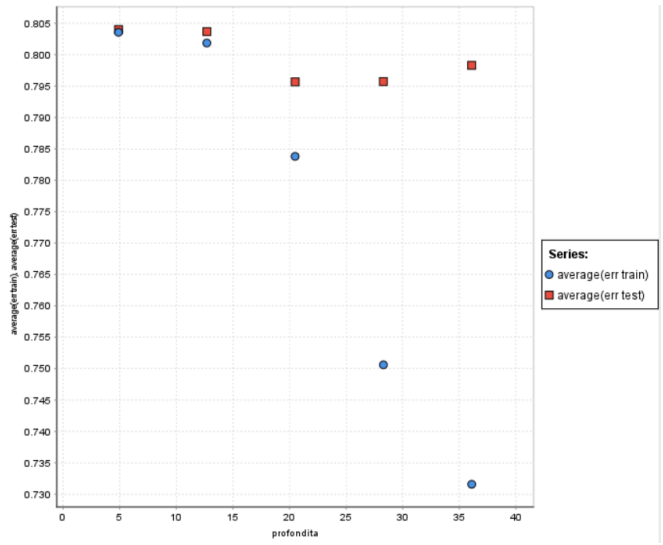


Figura 29: Errore di classificazione di training e test cfu

Sia per la classificazione sui voti, sia per la classificazione dei cfu l'algoritmo Decision Tree con errore più basso è quello con profondità massima pari a 20 (Fig. 28 e 29).

4.2.1 Risultati "Voto"

accuracy: 16.64% +/- 0.94% (micro average: 16.64%)

	true 1	true 8	true 3	true 6	true 2	true 7	true 4	true 5	true 9	true 10	class pre...
pred. 1	1129	507	716	767	674	646	764	823	362	195	17.15%
pred. 8	3	12	5	11	4	21	9	11	15	9	12.00%
pred. 3	0	1	0	0	0	1	0	0	0	0	0.00%
pred. 6	52	50	49	56	42	57	59	66	46	17	11.34%
pred. 2	3	0	1	0	3	2	1	3	1	0	21.43%
pred. 7	20	50	19	48	20	68	45	44	36	40	17.44%
pred. 4	2	5	1	0	0	2	2	2	1	0	13.33%
pred. 5	211	263	227	307	173	292	273	320	214	100	13.45%
pred. 9	112	302	126	242	80	258	177	230	336	243	15.95%
pred. 10	11	119	26	97	19	78	31	52	148	217	27.19%
class rec...	73.17%	0.92%	0.00%	3.66%	0.30%	4.77%	0.15%	20.63%	28.99%	26.43%	

Figura 30: Matrice di confusione

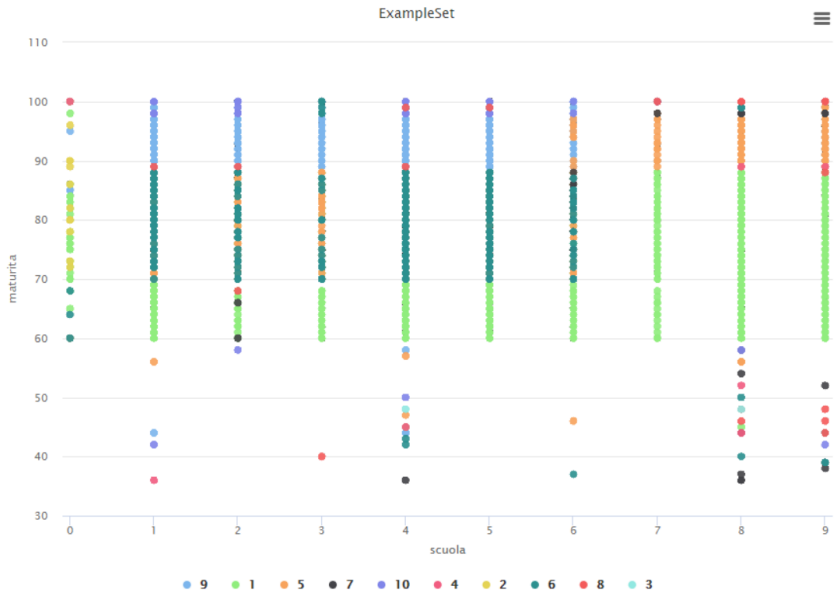


Figura 31: Classi predette

4.2.2 Risultati "CFU"

accuracy: 20.72% +/- 0.93% (micro average: 20.72%)

	true 10	true 20	true 60	true 50	true 40	true 30	true 0	class precision
pred. 10	67	46	4	16	36	56	80	21.97%
pred. 20	43	42	8	21	27	49	45	17.87%
pred. 60	15	17	36	22	33	29	7	22.64%
pred. 50	166	228	200	396	291	296	135	23.13%
pred. 40	4	7	1	11	16	13	6	27.59%
pred. 30	1226	1429	759	1390	1566	1889	1188	20.00%
pred. 0	187	173	25	48	121	189	223	23.08%
class recall	3.92%	2.16%	3.48%	20.80%	0.77%	74.93%	13.24%	

Figura 32: Matrice di confusione

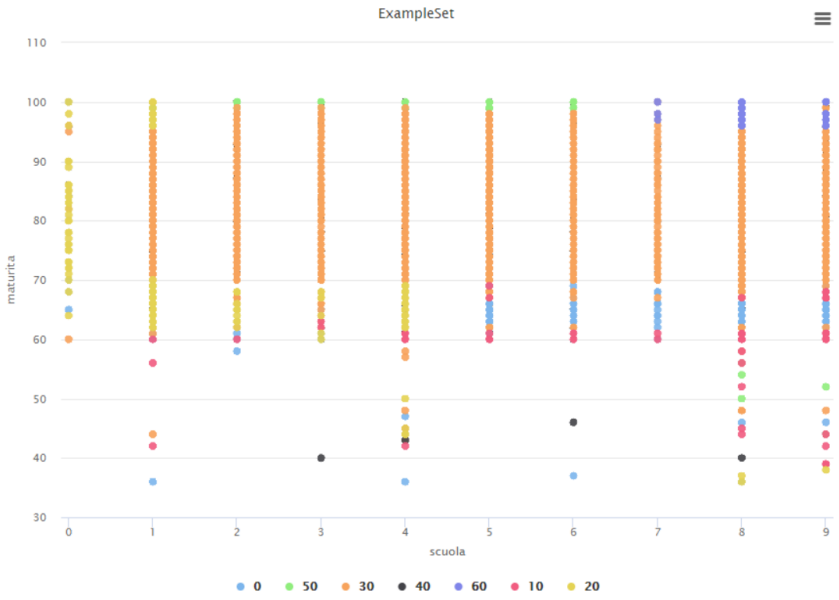


Figura 33: Classi predette

4.3 Conclusioni

Nonostante le innumerevoli prove anche utilizzando diverse tecnologie, non è stato possibile costruire un modello predittivo accurato con un basso errore di classificazione, questo a causa dell'elevata sparsità dei dati, in particolar modo perchè, sia della classificazione per voto sia in quelle dei cfu, gli studenti che vengono da diverse scuole posso avere un andamento simile tra loro, anche tenendo da conto il voto dell'esame di maturità. Per quanto riguarda invece le analisi effettuate sui dati è stato possibile ricavare diverse informazioni, in particola modo che la maggior parte degli studenti che frequentano le facoltà vengono da un Liceo Scientifico e i suddetti studenti tendono a iscriversi a quasi tutte le facoltà, al contrario di studenti provenienti da altre scuole più specializzate che tendono a scegliere pochi percorsi specifici molto spesso coerente con il loro indirizzo di scuola superiore. Un altro dato confortante deriva dall'immediata iscrizione all'università da parte di studenti appena diplomati ma che va in contrasto con l'alto numero di studenti che dopo il primo anno di studi tendono a cambiare facoltà con la conclusione che la maggior parte degli studenti intraprende fin da subito un percorso universitario anche se non sono del tutto soddisfatti della scelta iniziale presa.

5 Tecnologie Utilizzate

Abbiamo sviluppato l'intero processo di analisi, utilizzando Colab, con linguaggio Python e l'ausilio di libreria **Pandas**, **Numpy**, **sklearn**, etc., per quanto concerne la visualizzazione grafica, abbiamo utilizzato **matplotlib.pyplot**. Per costruire il modello predittivo è stato utilizzato anche il programma **Rapid Miner** oltre alle librerie e al linguaggio sopracitati.

Link Repository `https://colab.research.google.com/drive/1N7vy2HY_X9L1UlSr_XLHgB_B8tjSC8Dz?usp=sharing`

6 Leggenda

Etichette CFU Riportiamo qui di seguito le etichette utilizzate per i CFU

CFU	**LABEL CFU**
CFU $<$ 10	0
10 $\leq$ CFU $<$ 20	10
20 $\leq$ CFU $<$ 30	20
30 $\leq$ CFU $<$ 40	30
40 $\leq$ CFU $<$ 50	40
50 $\leq$ CFU $<$ 60	50
CFU $\geq$ 60	60

Tabella 7: CFU

Etichette Voto Riportiamo qui di seguito le etichette utilizzate per i Voti

VOTO	**LABEL VOTO**
voto $<$ 21	1
21 $\leq$ voto $<$ 22	2
22 $\leq$ voto $<$ 23	3
23 $\leq$ voto $<$ 24	4
24 $\leq$ voto $<$ 25	5
25 $\leq$ voto $<$ 26	6
26 $\leq$ voto $<$ 27	7
27 $\leq$ voto $<$ 28	8
28 $\leq$ voto $<$ 29	9
voto $\geq$ 29	10

Tabella 8: Voto

Etichette Scuola Riportiamo qui di seguito le etichette utilizzate per le scuole

SCUOLA	**LABEL SCUOLA**
Artistico	0
Classico	1
Linguistico	2
Professionale	3
Scientifico	4
Scientifico Scienze applicate	5
Scienze Umane	6
Scienze Umane Economico sociale	7
Tecnico - Economico	8
Tecnico - Tecnologico	9

Tabella 9: Scuola