Analisi dei risultati ottenuti

Metriche

Le metriche usate per i test sono le seguenti:

- *Specificity (TN_rate) = TN / (TN + FP)*
- *Recall (TP_rate) = TP / (TP + FN)*
- *Precision = TP / (TP + FP)*
- *F-measure = 2·Recall·Precision / (Recall+Precision)*
- *G-mean = (Recall · Specificity)$^{0.5}$*
- *MCC = (TP·TN-FP·FN) / ((TP+FP)(TP+FN)(TN+FP)(TN+FN)$^{0.5}$*

Dataset

I test sono stati condotti su tre dataset binari, tutti nel dominio della *text categorization*, con un diverso tasso di sbilanciamento: il primo "`acq`", con 7495 attributi ha un rapporto tra il numero di istanze della classe positiva ed il numero totale di istanze pari a 0,1719; il secondo "`money`" con 8302 attributi, ha un rapporto pari a 0,0563; l'ultimo, "`corn`" con 7757 attributi, ha un rapporto pari 0,0189. Tutti quanti avevano 9598 istanze. L'ordinamento con cui sono presentati i dataset rispecchia un rapporto decrescente o, se si vuole, uno sbilanciamento crescente. Quindi `acq` è il meno sbilanciato mentre `corn` è quello maggiormente sbilanciato.

Ho cercato di assegnare ai dataset dei colori diversi in modo da poterli distinguere quando si hanno i grafici davanti. Al dataset `acq` ho assegnato il celeste perché il nome mi ricorda l'acqua, al dataset `money` il rosa perché la banconota da 500€ è rosa, ed infine al dataset `corn` il giallo perché è il colore del mais.

Tabelle e grafici

A pagina 2 si trovano le tabelle, a pagina 3 i grafici raggruppati per dataset, mentre a pagina 4 i grafici raggruppati per tipologia di test, in modo da poter osservare come si comporta quel test al variare dello sbilanciamento del dataset. Nelle tabelle mostro tutte le metriche usate per i test, mentre nei grafici mostro le metriche "composte", cioè F-measure, G-mean e MCC.

I colori usati nei grafici sono stati scelti per veicolare informazioni riguardanti i test ed i dataset in modo da poter fare comparazioni visive più semplici. Ho assegnato diversi colori in base alla percentuale di features scelte nella Feature Selection: il colore rosso all'1%, il verde al 2% ed il blu al 5%; così da avere il famoso ordine "RGB" assegnato ad una percentuale crescente. Poi il grigio ad i test senza FS ed il bianco alla *baseline*, ovvero il solo classificatore *Random Forest*.

Risultati

I risultati variano parecchio in base allo sbilanciamento dei dataset.

Per quanto riguarda il dataset `acq`, che è quello con lo sbilanciamento minore, si parte già da una baseline abbastanza alta e la FS aggiunge solo qualche punto percentuale. Le FS al 2% e 5% danno un risultato migliore rispetto a quella all'1%. Le combinazioni con *Random Undersampling* non danno un grande miglioramento, mentre quelle con *Cost-Sensitive Classification* danno risultati diversi in base al costo, in particolare con un costo di 2 per i falsi negativi si nota un leggero miglioramento rispetto alla sola FS, mentre all'aumentare del costo si migliora sempre meno, fino a peggiorare per un costo pari a 5. In questo caso, dove la FS non porta a grossi miglioramenti, la si può comunque usare sia per ridurre il carico computazionale durante l'allenamento, sia per avere un'idea delle feature più importanti.

Nel dataset `money`, che è quello con sbilanciamento medio tra i tre studiati, si notano miglioramenti sia nella sola FS che FS combinata con gli altri metodi. L'unico caso è RU(1:1) dove per le metriche F-measure e MCC si hanno valori pressoché uguali alla sola FS. Inoltre il caso di Cost5 risulta migliore se preso da solo anziché combinato con FS.

Il dataset `corn` è quello con lo sbilanciamento maggiore e quello che ha portato a risultati più estremi. Si partiva da una baseline molto bassa e la FS ha migliorato notevolmente i risultati, soprattutto nel caso 1% dove si è arrivati a valori attorno a 0,9 per tutte e tre le metriche. Quello che si nota è che in combinazione con RU si hanno risultati peggiori della sola FS, soprattutto nel caso RU(1:1): questo potrebbe essere dovuto al basso numero di istanze usate per l'allenamento. Si presentano valori bassi della F-measure ma alti per la G-mean: guardando le tabelle si vede come nonostante ci siano *specificity* e *recall* molto alte, quindi pochi falsi negativi per le classi negative e positive, si ha una *precision* molto bassa. Questo vuol dire che si hanno tantissimi falsi positivi. Precision bassa la si ha soprattutto nel caso RU(1:1) e quando il campionamento viene fatto prima della FS. I metodi RU e Cost presi da soli risultano peggiori che quando combinati con la FS.

Per quanto riguarda i confronti tra test uguali effettuati su dataset diversi, nel caso di RU si hanno dei miglioramenti sicuramente nei primi due dataset `acq` e `money`, quelli con sbilanciamento rispettivamente minore e medio, mentre in `corn`, quello con lo sbilanciamento maggiore, RU combinato con FS è migliore rispetto alla baseline ma la sola FS risulta il metodo che dà risultati superiori.

Nel caso di Cost, i miglioramenti più apprezzabili si notano nel dataset `money`, quello con sbilanciamento medio. Per il dataset `acq`, quello con sbilanciamento minore, non si notano grosse differenze. Mentre per `corn`, quello con sbilanciamento maggiore, funziona malissimo da solo ma preso in combinazione con FS funziona meglio della sola FS.

acq pos/total: 0,1719

FS 1%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	0,998	0,983	0,975	0,964	0,937	0,978	0,967	0,952	0,971	0,959	0,948	0,940
Recall	0,730	0,811	0,848	0,896	0,929	0,857	0,901	0,930	0,894	0,922	0,936	0,950
Precision	0,991	0,928	0,904	0,874	0,805	0,914	0,884	0,843	0,894	0,863	0,833	0,815
F-Measure	0,841	0,866	0,875	0,885	0,862	0,884	0,893	0,884	0,894	0,892	0,881	0,877
G-Mean	0,854	0,893	0,909	0,929	0,933	0,916	0,933	0,941	0,932	0,940	0,942	0,945
MCC	0,819	0,834	0,842	0,852	0,824	0,854	0,862	0,852	0,865	0,861	0,848	0,844

FS 2%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	0,998	0,987	0,981	0,969	0,954	0,981	0,977	0,952	0,971	0,959	0,945	0,934
Recall	0,730	0,864	0,890	0,917	0,946	0,889	0,915	0,943	0,921	0,939	0,950	0,958
Precision	0,991	0,950	0,928	0,892	0,852	0,927	0,916	0,846	0,899	0,864	0,829	0,802
F-Measure	0,841	0,905	0,908	0,904	0,897	0,908	0,916	0,892	0,910	0,900	0,885	0,873
G-Mean	0,854	0,923	0,934	0,943	0,950	0,934	0,945	0,947	0,946	0,949	0,947	0,946
MCC	0,819	0,881	0,884	0,877	0,868	0,883	0,892	0,862	0,885	0,872	0,854	0,840

FS 5%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	0,998	0,991	0,988	0,981	0,957	0,988	0,985	0,963	0,976	0,959	0,941	0,924
Recall	0,730	0,860	0,905	0,937	0,975	0,905	0,929	0,962	0,947	0,975	0,983	0,987
Precision	0,991	0,966	0,953	0,931	0,862	0,953	0,945	0,878	0,915	0,868	0,823	0,783
F-Measure	0,841	0,909	0,929	0,934	0,915	0,929	0,937	0,918	0,931	0,918	0,896	0,873
G-Mean	0,854	0,923	0,946	0,959	0,966	0,946	0,957	0,962	0,961	0,967	0,962	0,955
MCC	0,819	0,888	0,910	0,916	0,892	0,910	0,920	0,896	0,912	0,896	0,870	0,842

No FS

-	Baseline	-	RU(3:1)	RU(2:1)	RU(1:1)	-	-	-	Cost2	Cost3	Cost4	Cost5
Specificity	0,998	-	0,996	0,989	0,965	-	-	-	0,984	0,963	0,933	0,900
Recall	0,730	-	0,844	0,918	0,971	-	-	-	0,951	0,983	0,990	0,994
Precision	0,991	-	0,982	0,959	0,885	-	-	-	0,942	0,880	0,805	0,735
F-Measure	0,841	-	0,908	0,938	0,926	-	-	-	0,947	0,929	0,888	0,845
G-Mean	0,854	-	0,917	0,953	0,968	-	-	-	0,967	0,973	0,961	0,946
MCC	0,819	-	0,889	0,922	0,905	-	-	-	0,932	0,910	0,861	0,810

money pos/total: 0,0561

FS 1%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	0,996	0,991	0,971	0,964	0,950	0,971	0,963	0,937	0,981	0,975	0,972	0,967
Recall	0,235	0,575	0,799	0,849	0,877	0,782	0,832	0,911	0,721	0,777	0,827	0,844
Precision	0,764	0,786	0,614	0,578	0,503	0,606	0,560	0,452	0,690	0,641	0,627	0,597
F-measure	0,359	0,665	0,694	0,688	0,640	0,683	0,670	0,604	0,705	0,702	0,713	0,699
G-mean	0,484	0,755	0,881	0,905	0,913	0,871	0,895	0,924	0,841	0,870	0,897	0,903
MCC	0,408	0,657	0,681	0,68	0,641	0,668	0,661	0,615	0,688	0,687	0,702	0,69

FS 2%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	0,996	0,991	0,975	0,971	0,953	0,977	0,972	0,953	0,984	0,978	0,972	0,967
Recall	0,235	0,536	0,804	0,855	0,905	0,838	0,855	0,888	0,732	0,782	0,872	0,894
Precision	0,764	0,780	0,646	0,630	0,524	0,676	0,638	0,518	0,720	0,670	0,645	0,606
F-measure	0,359	0,636	0,716	0,725	0,664	0,748	0,730	0,654	0,726	0,722	0,741	0,722
G-mean	0,484	0,729	0,885	0,911	0,929	0,905	0,912	0,920	0,849	0,875	0,921	0,930
MCC	0,408	0,631	0,703	0,716	0,667	0,737	0,721	0,656	0,710	0,707	0,733	0,718

FS 5%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	0,996	0,994	0,977	0,972	0,951	0,978	0,969	0,950	0,988	0,982	0,975	0,966
Recall	0,235	0,453	0,821	0,877	0,922	0,799	0,849	0,922	0,715	0,816	0,883	0,911
Precision	0,764	0,802	0,671	0,646	0,519	0,671	0,613	0,516	0,771	0,723	0,667	0,608
F-measure	0,359	0,579	0,739	0,744	0,664	0,730	0,712	0,661	0,742	0,766	0,760	0,729
G-mean	0,484	0,671	0,896	0,923	0,936	0,884	0,907	0,936	0,840	0,895	0,928	0,938
MCC	0,408	0,587	0,726	0,737	0,670	0,716	0,703	0,668	0,728	0,754	0,752	0,727

No FS

-	Baseline	-	RU(3:1)	RU(2:1)	RU(1:1)	-	-	-	Cost2	Cost3	Cost4	Cost5
Specificity	0,996	-	0,985	0,979	0,934	-	-	-	0,993	0,988	0,981	0,970
Recall	0,235	-	0,698	0,838	0,978	-	-	-	0,520	0,760	0,827	0,939
Precision	0,764	-	0,727	0,701	0,459	-	-	-	0,809	0,782	0,712	0,644
F-measure	0,359	-	0,712	0,763	0,625	-	-	-	0,633	0,771	0,765	0,764
G-mean	0,484	-	0,829	0,906	0,956	-	-	-	0,719	0,867	0,901	0,954
MCC	0,408	-	0,696	0,752	0,646	-	-	-	0,633	0,758	0,753	0,763

corn pos/total: 0,0189

FS 1%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	1,000	0,998	0,990	0,981	0,968	0,990	0,985	0,981	0,997	0,996	0,996	0,994
Recall	0,161	0,875	0,982	1,000	1,000	1,000	1,000	0,982	0,982	0,982	1,000	1,000
Precision	0,900	0,875	0,632	0,479	0,350	0,629	0,538	0,478	0,833	0,809	0,800	0,737
F-Measure	0,273	0,875	0,769	0,647	0,519	0,772	0,700	0,643	0,902	0,887	0,889	0,848
G-Mean	0,401	0,934	0,986	0,990	0,984	0,995	0,992	0,981	0,989	0,989	0,998	0,997
MCC	0,377	0,873	0,784	0,685	0,582	0,789	0,728	0,679	0,903	0,889	0,892	0,856

FS 2%

-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	1,000	0,998	0,986	0,978	0,961	0,988	0,981	0,979	0,998	0,998	0,998	0,998
Recall	0,161	0,679	0,982	0,982	1,000	0,982	1,000	0,982	0,821	0,875	0,893	0,786
Precision	0,900	0,844	0,550	0,433	0,309	0,591	0,479	0,444	0,868	0,891	0,877	0,863
F-Measure	0,273	0,752	0,705	0,601	0,473	0,738	0,647	0,611	0,844	0,883	0,885	0,822
G-Mean	0,401	0,823	0,984	0,980	0,980	0,985	0,990	0,980	0,905	0,934	0,944	0,886
MCC	0,377	0,753	0,730	0,645	0,545	0,757	0,685	0,653	0,842	0,881	0,883	0,820

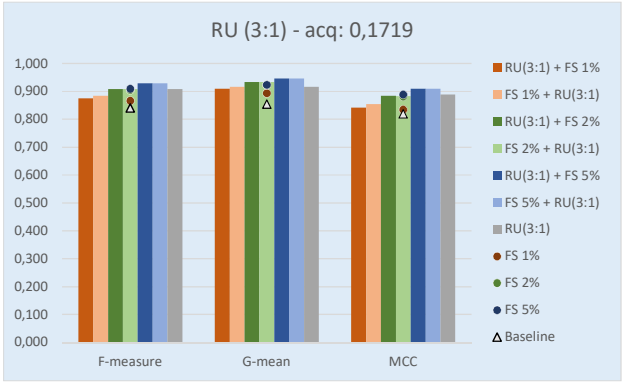
FS 5%

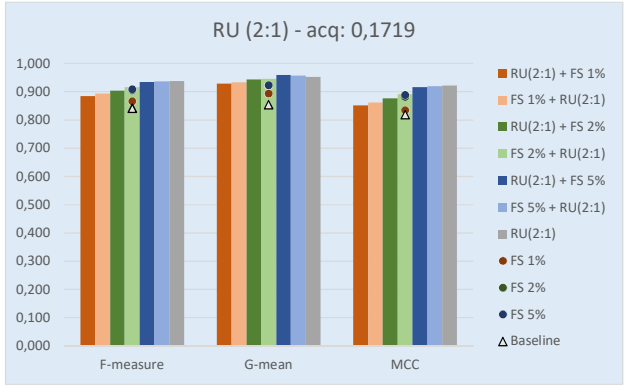
-	Baseline	FS	RU(3:1) + FS	RU(2:1) + FS	RU(1:1) + FS	FS + RU(3:1)	FS + RU(2:1)	FS + RU(1:1)	FS + Cost2	FS + Cost3	FS + Cost4	FS + Cost5
Specificity	1,000	1,000	0,981	0,964	0,949	0,981	0,974	0,972	0,998	0,995	0,994	0,990
Recall	0,161	0,518	0,929	0,964	0,964	0,946	0,946	0,982	0,821	0,982	1,000	1,000
Precision	0,900	0,967	0,456	0,318	0,244	0,469	0,387	0,374	0,885	0,786	0,727	0,644
F-Measure	0,273	0,674	0,612	0,478	0,390	0,627	0,549	0,542	0,852	0,873	0,842	0,783
G-Mean	0,401	0,720	0,955	0,964	0,956	0,963	0,960	0,977	0,905	0,988	0,997	0,995
MCC	0,377	0,704	0,643	0,543	0,472	0,659	0,596	0,597	0,850	0,876	0,850	0,798

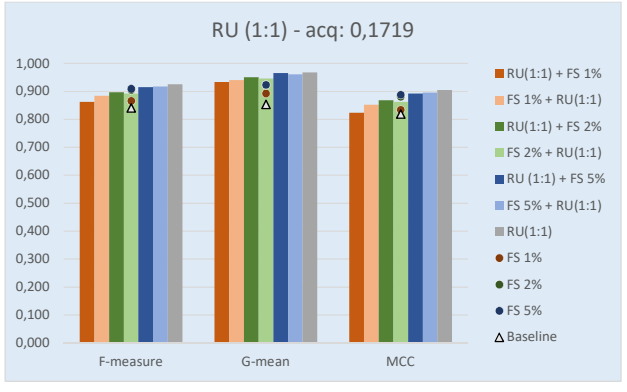
No FS

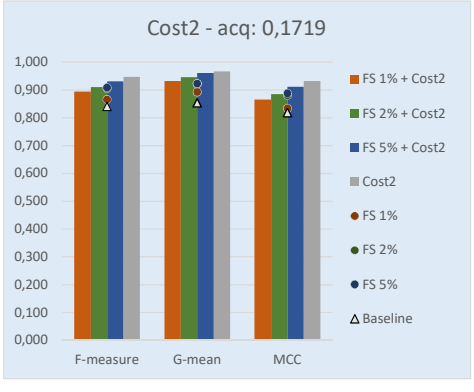
-	Baseline	-	RU(3:1)	RU(2:1)	RU(1:1)	-	-	-	Cost2	Cost3	Cost4	Cost5
Specificity	1,000	-	0,985	0,971	0,948	-	-	-	0,998	0,998	0,998	0,995
Recall	0,161	-	0,625	0,804	0,929	-	-	-	0,321	0,446	0,607	0,714
Precision	0,900	-	0,422	0,326	0,234	-	-	-	0,783	0,806	0,829	0,714
F-Measure	0,273	-	0,504	0,464	0,374	-	-	-	0,456	0,575	0,701	0,714
G-Mean	0,401	-	0,785	0,884	0,938	-	-	-	0,566	0,667	0,778	0,843
MCC	0,377	-	0,503	0,500	0,452	-	-	-	0,497	0,595	0,705	0,709

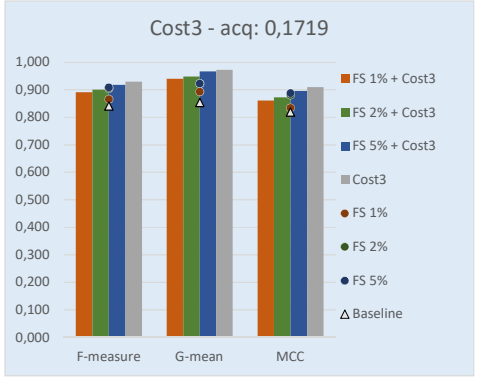
Grafici relativi al dataset `acq`

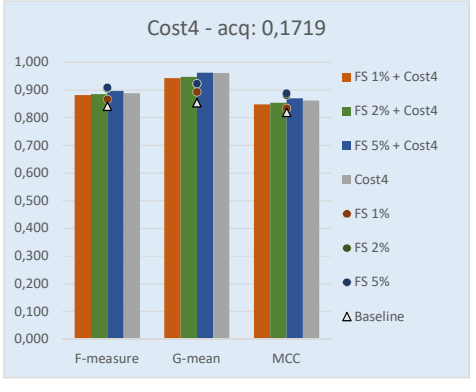


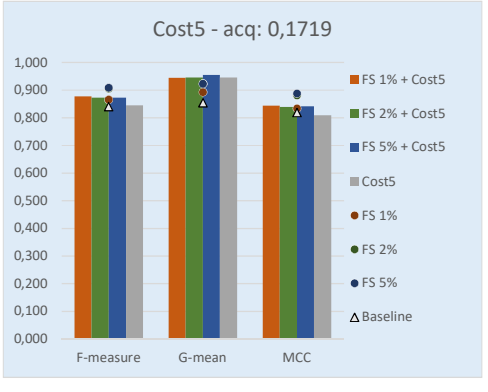




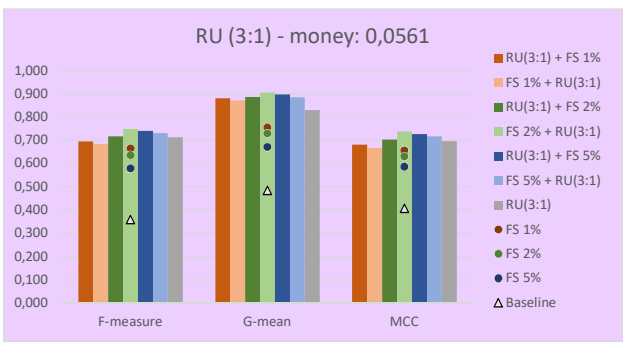


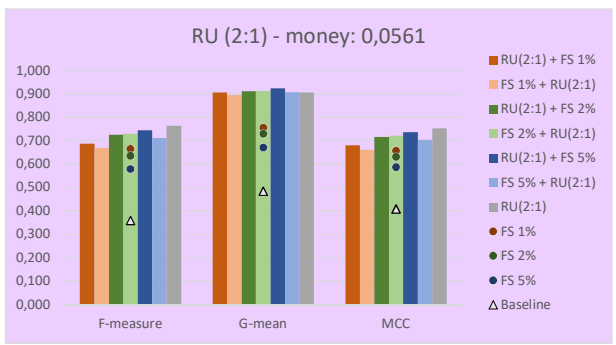


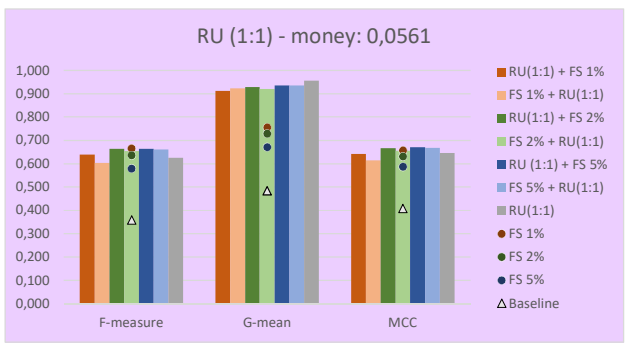


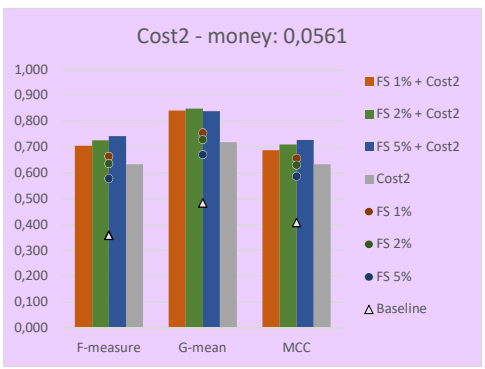


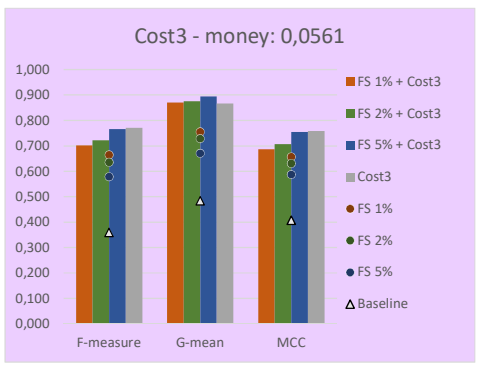
Grafici relativi al dataset `money`

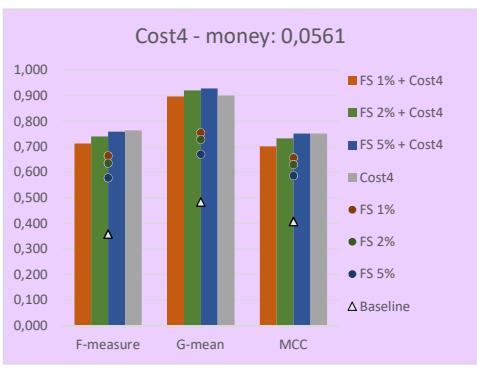


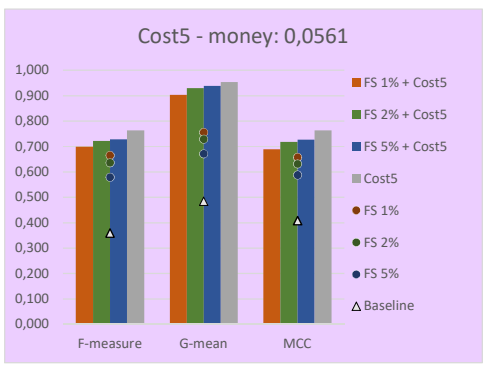




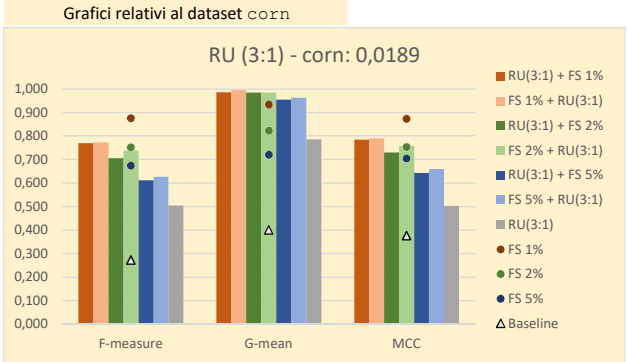


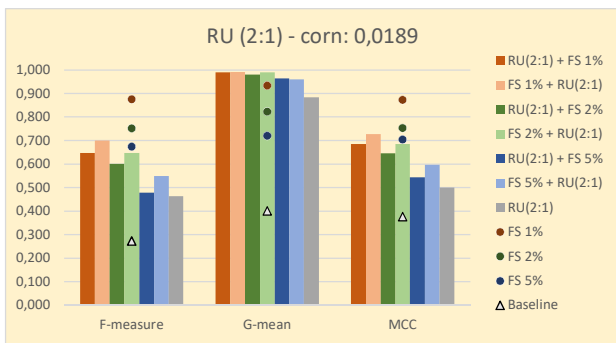


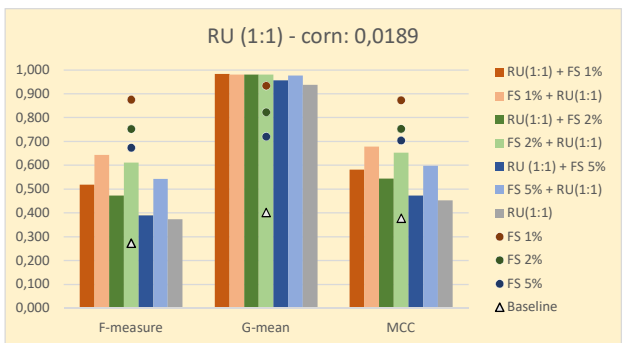


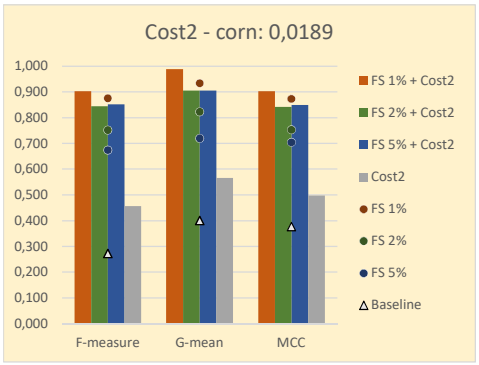


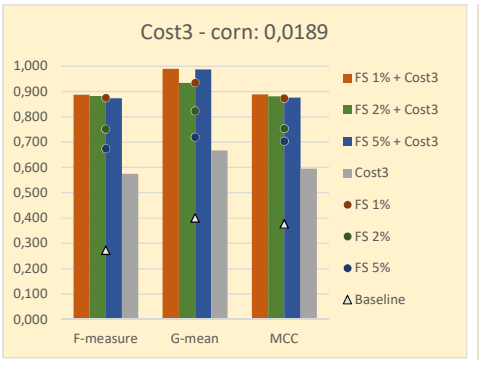
Grafici relativi al dataset `corn`

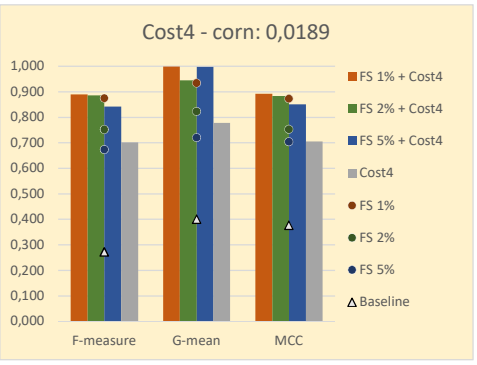


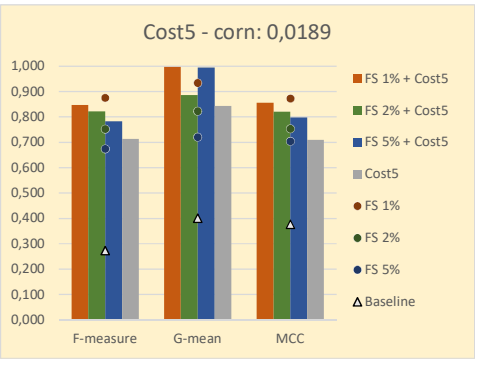




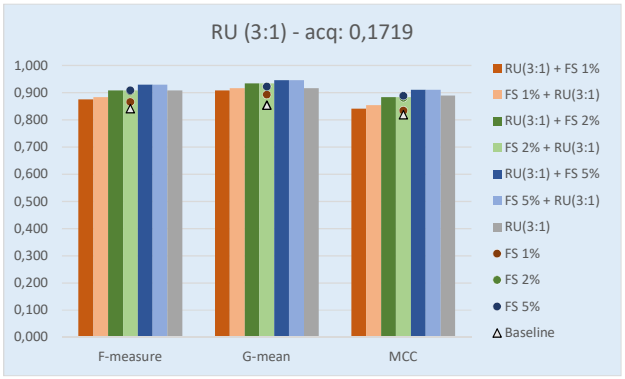


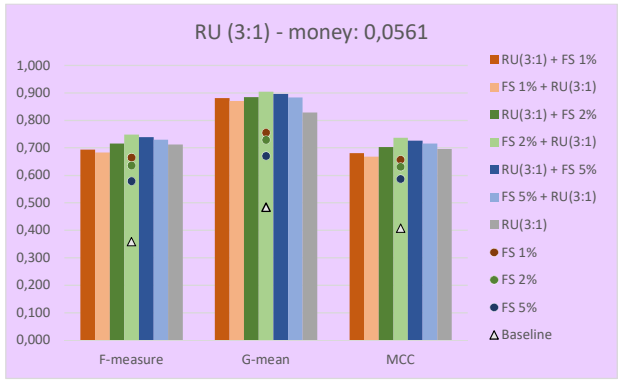


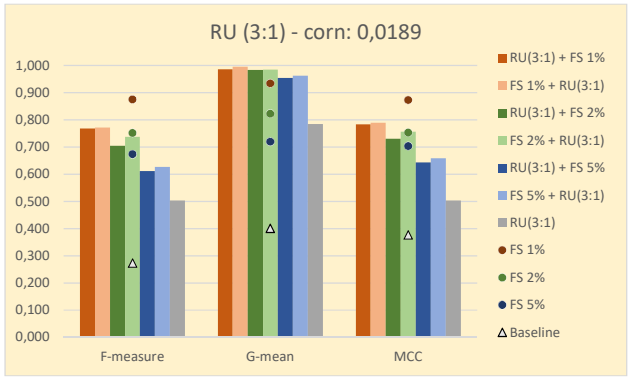


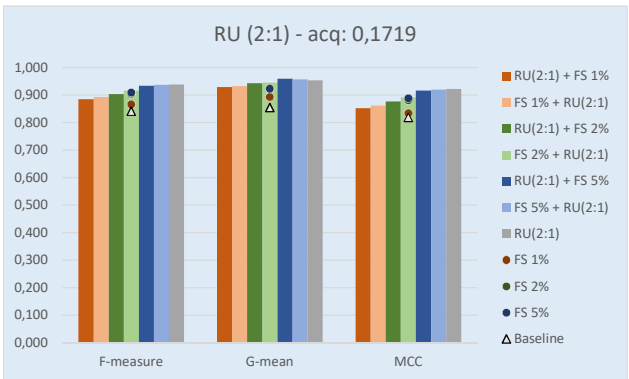


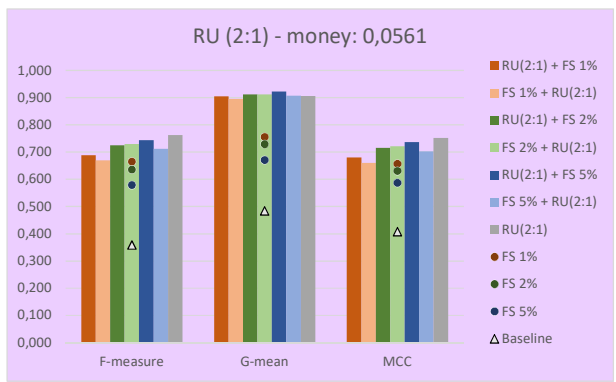
Confronto tra Random Undersampling tra diverse proporzioni e diversi dataset

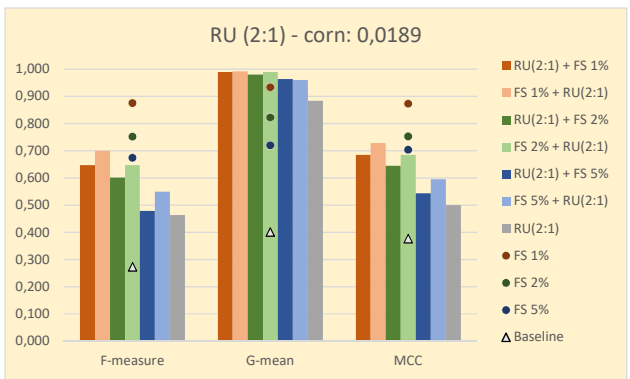


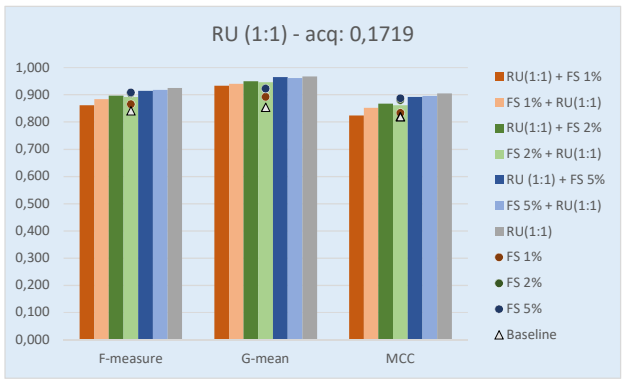


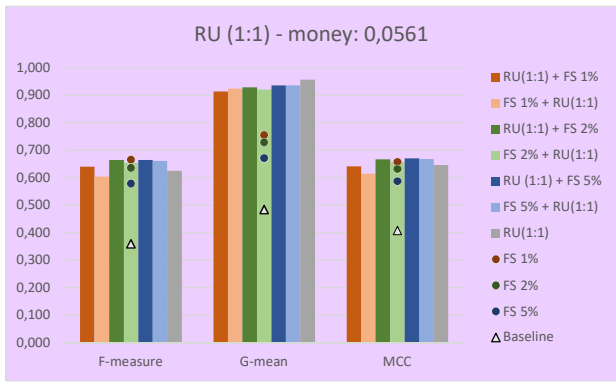


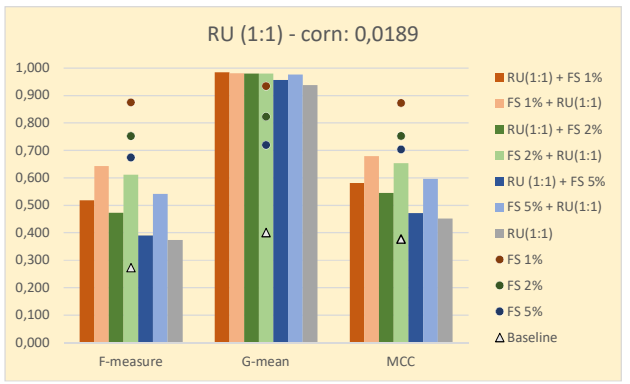












Confronto tra Cost Sensitive Learning tra diversi costi e diversi dataset