Modelli di classificazione per la previsione delle precipitazioni

Matteo Campironi, Serena Di Maggio

Università degli studi di Milano-Bicocca, CdLM Data Science

La previsione meteorologica è un'attività che da millenni l'uomo svolge in modo informale e solo formalmente dal XIX secolo. Predire le precipitazioni è una sfida complicata, ma che potrebbe avere un impatto significativo sulla società, permettendo di ridurre perdite dal punto di vista sia umano che finanziario. In questo progetto sono state implementate delle tecniche di Machine Learning al fine di costruire modelli che permettano di affermare se domani pioverà o meno, basandosi su dati meteorologici raccolti nelle più grandi città australiane.

Indice

1 Introduzione

Le previsioni meteorologiche hanno da sempre giocato un ruolo fondamentale nella vita quotidiana delle persone. Oggi queste sono facilmente consultabili su internet, sui nostri smartphone oppure in televisione, permettendo di scegliere cosa indossare la mattina, di decidere se uscire con un ombrello o meno oppure di pianificare attività come gite o feste. La loro utilità però va ben oltre il rendere la vita più facile. Le previsioni meteorologiche sono anche essenziali per salvaguardarsi da fenomeni naturali quali frane, inondazioni o valanghe. Il seguente articolo si focalizzerà sulla previsione delle precipitazioni. Questo particolare fattore climatico influenza le nostre attività, come l'agricoltura, la generazione di energia o il turismo e perciò è importante riuscire a prevederlo. Per questo scopo entrano in gioco molte variabili, come per esempio la temperatura, la velocità e la direzione del vento, l'umidità o la nuvolosità. Una variazione di questi parametri può portare o meno a precipitazioni. Oggi, grazie allo sviluppo tecnologico, possiamo tentare di approcciare il problema utilizzando tecniche avanzate e programmi sofisticati. Purtroppo le previsioni meteorologiche presentano ancora notevole incertezza, dovuta principalmente alla complessità del problema e agli errori ed imperfezioni nei dati raccolti.

In questo elaborato sono state analizzate diverse tecniche di classificazione proprie del Machine Learning il cui scopo è quello di predire se domani pioverà o meno, per poi valutare quali sono i modelli migliori. In secondo luogo si è tentato di identificare quali potessero essere le variabili più importanti al fine della classificazione. Lo sviluppo dell'intero progetto è stato reso possibile grazie all'utilizzo della piattaforma KNIME, sulla quale è stato elaborato un workflow a cui fa riferimento tutto il lavoro di seguito discusso.

2 Dataset

Il dataset preso in considerazione [3] contiene osservazioni giornaliere di natura metereologica, rilevate presso diverse stazioni australiane. Esso consiste di 23 features e di 142193 istanze. Di seguito ne riportiamo una descrizione più precisa:

- **Date:** data dell'osservazione;
- **Location:** luogo della rilevazione;
- **MinTemp:** temperatura minima (°C);
- **MaxTemp:** temperatura massima (°C);
- **Rainfall:** precipitazioni giornaliere (mm);
- **Evaporation:** evaporimetro di classe A nelle 24 ore fino alle 09:00 (mm);
- **Sunshine:** ore di luce del sole nella giornata;
- **WindGustDir:** la direzione della raffica di vento più forte nelle 24 ore fino alle 00:00;
- **WindGustSpeed:** la velocità della raffica di vento più forte nelle 24 ore fino alle 00:00;
- **WindDir9am:** direzione del vento alle 09:00;
- **WindDir3pm:** direzione del vento alle 15:00;
- **WindSpeed9am:** media ogni 10 minuti della velocità del vento prima delle 09:00 (km/h);
- **WindSpeed3pm:** media ogni 10 minuti della velocità del vento prima delle 15:00 (km/h);
- **Humidity9am:** umidità (%) alle 09:00;
- **Humidity3pm:** umidità (%) alle 15:00;
- **Pressure9am:** pressione al livello del mare alle 09:00 (hPa);
- **Pressure3pm:** pressione al livello del mare alle 15:00 (hPa);
- **Cloud9am:** frazione del cielo oscurata dalle nuvole alle 09:00;
- **Cloud3pm:** frazione del cielo oscurata dalle nuvole alle 15:00;
- **Temp9am:** temperatura alle 09:00 in gradi celsius;
- **Temp3pm:** temperatura alle 15:00 in gradi celsius;
- **RainToday:** ha piovuto oggi? (Yes/No);
- **RISK_MM:** precipitazioni del giorno successivo (mm);
- **RainTomorrow:** la variabile target. Pioverà domani? (Yes/No).

2.1 Trattamento dei dati mancanti

Il dataset presenta numerosi dati mancanti, che si suppone derivino dall'integrazione di rilevazioni provenienti da stazioni meteo posizionate in città distinte. In particolare le features *Sunshine*, *Evaporation*, *Cloud3pm* e *Cloud9am* presentano più del 40% di valori nulli, la cui causa potrebbe essere la mancanza di determinati strumenti di raccolta dati in alcune località. Di conseguenza la scelta è stata quella di rimuovere le colonne, per evitare distorsioni nel dataset dovute a supposizioni errate. Per quanto riguarda i rimanenti dati mancanti, le scelte sono state due:

- Supponendo che in periodi di tempo ravvicinati non vi siano brusche variazioni nelle rilevazioni, i valori mancanti sono stati riempiti considerando quelli corrispondenti alla data precedente e successiva ed eseguendone una media.

- Nel caso invece vi fossero dei periodi di tempo nei quali le registrazioni avessero fornito più dati nulli in successione, la scelta è stata quella di rimuovere le righe, sempre per evitare distorsioni nel dataset.

Dopo questo trattamento, il dataset consiste di 19 features e 113554 istanze.

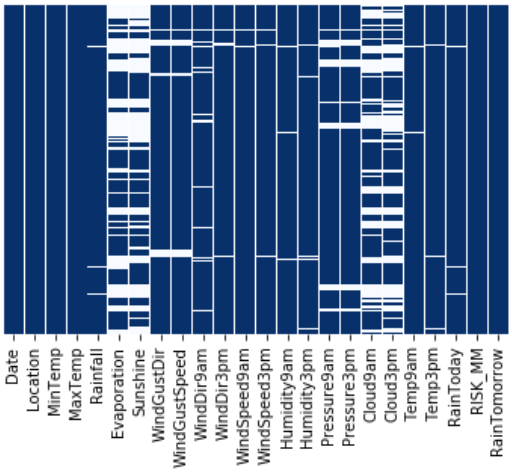


Figura 1: *Heatmap rappresentante i NaN nel dataset.*

2.2 Outliers

Alcune feature, come per esempio *Rainfall*, *WindGustSpeed*, *WindSpeed9am* o *WindSpeed3pm*, presentano dei valori estremamente alti in confronto al rispettivo terzo quartile (Q3). Le cause di questi outliers possono essere molteplici; per esempio potrebbero essere dovuti ad errori commessi durante l'inserimento dei dati, errori di misurazione oppure semplicemente potrebbero essere registrazioni veritiere molto distanti dalle altre. Dal momento che i dati non presentano anomalie così evidenti da poter essere modificati, la scelta è stata quella di tenere gli outliers. Questi infatti potrebbero essere importanti al fine della previsione della nostra variabile target.

2.3 Feature creation

Dalle variabili iniziali sono state create alcune nuove feature che saranno utilizzate nei modelli di machine learning. Da *Date* sono state estratte tre colonne: *Day*, *Month*, e *Year*. Inoltre è stata creata una nuova feature, *Ratio*, definita per ogni località nel seguente modo:

$$Ratio := \frac{\#\text{ di giorni di pioggia}}{\#\text{ rilevazioni totali}}$$

Questo ci ha permesso di effettuare un processo di discretizzazione non supervisionata e di suddividere *Ratio* in 6 intervalli di egual ampiezza.

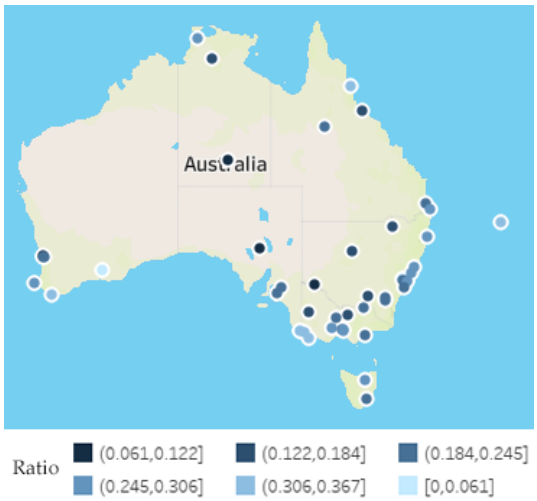
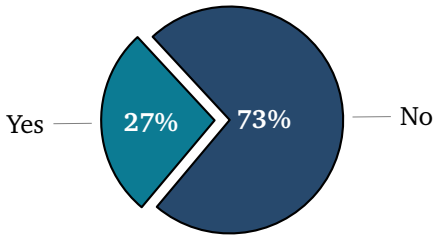


Figura 2: *Posizione geografica delle stazioni meteo raggruppate secondo l'attributo Ratio.*

2.4 Class imbalance

Da una prima analisi esplorativa è emerso un problema di class imbalance: la variabile target (*RainTomorrow*) assume per il 73% dei casi il valore *No*, mentre per il 27% il valore *Yes*.



La classe di interesse è in minoranza, e questo potrebbe causare problemi di overfitting e distorsione dei classificatori. Per questo motivo proponiamo diverse soluzioni al problema:

- **Ricampionamento:** l'idea consiste nell'utilizzare due diverse tecniche per ribilanciare il dataset. In particolare, si è scelto di utilizzare l'Equal Size Sampling per quanto riguarda l'undersampling, mentre il campionamento SMOTE per quanto riguarda l'oversampling. Il primo mantiene le osservazioni della classe minoritaria e ne campiona in modo aleatorio lo stesso numero da quella maggioritaria. Lo SMOTE invece introduce nella classe positiva istanze *sintetiche*, create da un'interpolazione casuale.

- **Analisi dei costi:** i classificatori vengono addestrati tenendo conto di una matrice di costo. È stato quindi utilizzato il nodo Weka *CostSensitiveClassifier*, basato su approccio Cost Sensitive Learning.

2.5 Data preparation

Nel dataset sono presenti sia variabili quantitative che qualitative. Quest'ultime però non possono essere utilizzate in tutti i modelli predittivi. Per ovviare a questo problema si è deciso di ricorrere alla trasformazione in variabili binarie. Per le features *WindGustDir*, *WindDir9am*, *WindDir3pm RainToday* si è scelto di utilizzare la One-Hot Encoding, in modo tale che ogni valore venga considerato indipendente dagli altri, evitando di creare correlazioni inesistenti originariamente. La feature *Location* contiene invece 49 valori distinti, da noi considerati troppi per riproporre l'approccio precedente. In alternativa alla tecnica della binarizzazione, che avrebbe aggiunto $\lceil \log_2(49) \rceil$ (ossia 6) colonne, si è preferito utilizzare la feature *Ratio* creata precedentemente, in modo tale da raggruppare le località in base alla loro piovosità. È stata inoltre rimossa *RISK_MM* come suggerito dal creatore del dataset, in quanto questa è stata usata per la creazione della variabile target *RainTomorrow* ed il suo utilizzo falsificherebbe i risultati. Otteniamo in definitiva un dataset composto da 66 feature.

Come ultimo step, è stata eseguita una normalizzazione dei dati. Se si evitasse di effettuare questo passaggio, features che hanno ordine di grandezza maggiore peserebbero di più sul calcolo delle distanze, che è il concetto alla base dei modelli di separazione. I nodi Weka di SPegasos e MultilayerPerceptron danno la possibilità di effettuare una normalizzazione interna, basata sul *Min-Max Scaling*. Questa metodologia però è influenzata dalla presenza di outliers e di conseguenza si è preferito utilizzare una normalizzazione *Z-Score*.

3 Modelli

Per questo studio sono state usate diverse tecniche di classificazione, con lo scopo di individuare i modelli più performanti. In particolare, possiamo suddividerli nelle quattro categorie seguenti:

- **Modelli Euristici:** albero di regressione J48 e classificatore Random Forest con 10 alberi;

- **Modelli di Regressione:** regressione logistica, adatto a modellare una variabile dipendente di tipo binario, come la variabile target *RainTomorrow*;

- **Modelli Probabilistici:** Naïve Bayes, basato sull'applicazione del teorema di Bayes e sull'ipotesi di indipendenza delle feature;

- **Modelli di Separazione:** SPegasos e MultiLayerPerceptron con 5 hidden layers [5], classificatori che rientrano rispettivamente nelle Support Vector Machine (SVM) e nelle reti neurali artificiali.

3.1 Misure utilizzate

Per valutare le performance dei vari modelli utilizzati, è stato deciso di basarsi sulle seguenti misure: Recall, Precision, F_1-measure, Area Under Curve (AUC) della curva Receiver Operating Characteristic (ROC). L'Accuracy, così definita:

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

è stata trascurata a causa dei problemi di imbalance del dataset. Essendo quest'ultimo sbilanciato sulla variabile target, si ottengono valori più elevati di questa misura a causa della prevalenza della classe maggioritaria su quella minoritaria. Si è scelto quindi di usare criteri di valutazione più completi, in grado di fornire una visione migliore e meno distorta dei diversi comportamenti dei modelli. Recall, definita anche come True Positive Rate, Precision e F_1-measure vengono calcolati nel seguente modo:

$$\text{Recall} := \frac{TP}{TP + FN}$$

$$\text{Precision} := \frac{TP}{TP + FP}$$

$$\text{F}_1\text{-measure} := \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

La Recall indica la frazione di record positivi che sono stati classificati correttamente, differentemente dalla Precision, la quale indica la porzione di record effettivamente positivi tra quelli predetti come tali. Elevati valori di queste due misure sono auspicabili, poiché maggiore è la Recall, più basso è il numero di record positivi classificati erroneamente, mentre più alta è la Precision e minore è il numero dei falsi negativi. Non sempre è possibile massimizzare i valori di queste due misure allo stesso tempo, poiché spesso all'aumentare dell'una diminuisce l'altra; pertanto, viene considerata la F_1-measure, definita come la media armonica tra Recall e Precision.

Nel nostro studio è stato deciso di dare maggior risalto ai valori forniti dalla Recall, dato che si è supposto che i danni provocati da una giornata di pioggia non prevista possano comportare conseguenze peggiori rispetto a un procurato falso allarme. Infine, si è considerata la curva ROC, la quale rappresenta la percentuale dei veri positivi rispetto a quella dei falsi positivi a varie impostazioni di soglia e di cui si studia l'area sotto la curva (AUC): tanto più il suo valore è grande, tanto più efficace è il modello.

4 Analisi e risultati

In questa sezione vengono riportati i principali risultati derivanti dai diversi approcci utilizzati al fine di effettuare la classificazione. In particolare, si cerca di rispondere alle seguenti research question:

- È possibile trovare un modello che riesca a predire in modo efficace se domani pioverà o meno in Australia?
- In tal caso, si è in grado di individuare le variabili che maggiormente influenzano questo fenomeno?

4.1 Analisi preliminare

Un primo passo verso la risposta al primo quesito è stato quello di addestrare i vari modelli senza alcun processo preliminare, osservando i valori delle misure di performance ottenuti.

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.526	0.628	0.572	0.753	0.826
Random Forest	0.452	0.744	0.562	0.850	0.844
Logistic	0.505	0.724	0.595	0.870	0.848
Naïve Bayes	0.556	0.545	0.551	0.811	0.799
SPegasos	0.474	0.741	0.578	0.713	0.847
MLP	0.535	0.706	0.609	0.861	0.848

Tabella 1: *Misure di performance dei diversi modelli.*

Come mostrato nella Tabella 1, i classificatori presentano elevati valori di Accuracy. È possibile giustificare questo ricordandosi che vi è uno sbilanciamento della variabile target e di conseguenza vanno prese in considerazione le altre misure. Tuttavia, da quanto si evince dai valori di Precision e Recall, nessun classificatore riesce a catturare appieno i valori realmente positivi. Essendo al corrente dei problemi di class imbalance e curse of dimensionality, si è tentato di affrontarli attraverso l'ausilio di tecniche di sampling e feature selection, ricordando la scelta della Recall come misura più significativa.

4.2 Analisi con metodo Holdout e Sampling

Dopo l'analisi preliminare, si è deciso di cercare di ovviare al problema della natura sbilanciata del dataset effettuando sia l'Equal Size Sampling che lo SMOTE. Innanzitutto, come previsto dal metodo Holdout, il dataset è stato partizionato in training set, utilizzato per addestrare i classificatori, e test set, usato per validare i modelli, costituendo rispettivamente il 67% e il 33% del totale. I risultati ottenuti in questo modo sono sintetizzati nelle Tabelle 2 e 3, che fanno riferimento ai due metodi di sampling considerati.

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.729	0.459	0.563	0.747	0.749
Random Forest	0.729	0.540	0.620	0.855	0.802
Logistic	0.771	0.526	0.625	0.870	0.795
Naïve Bayes	0.649	0.494	0.561	0.811	0.775
SPegasos	0.769	0.525	0.624	0.785	0.795
MLP	0.875	0.418	0.566	0.861	0.702

Tabella 2: *Misure di performance dei diversi modelli con Equal Size Sampling.*

Come è possibile osservare, i valori della Recall sono aumentati rispetto all'analisi precedente, al contrario della Precision e dell'Accuracy che diminuiscono per tutti i modelli. Quest'ultima ora risulta essere una misura più affidabile in quanto il dataset è stato bilanciato.

Utilizzando l'Equal Size Sampling, si ottiene che la Regressione Logistica è il modello più bilanciato per quanto riguarda le misure di performance, con una F_1-mesaure pari a 0.625 e un'AUC pari a 0.870, le più elevate.

Nel secondo caso, facendo uso dello SMOTE, i modelli euristici non sembrano rispondere ad un notevole aumento della Recall, mentre presentano valori elevati di Accuracy. Questo manifesta ancora una loro incapacità nel classificare la classe minoritaria. Situazione opposta presenta invece SPegasos, che tende a identificare eccessivamente in modo positivo le istanze. In questo caso, i modelli più convincenti risultano essere il MultiLayer Perceptron e, ancora una volta, la Regressione Logistica, che presentano valori molto simili tra loro.

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.573	0.554	0.563	0.737	0.803
Random Forest	0.553	0.669	0.606	0.858	0.840
Logistic	0.757	0.532	0.625	0.867	0.798
Naïve Bayes	0.675	0.389	0.494	0.755	0.693
SPegasos	0.967	0.312	0.472	0.680	0.520
MLP	0.735	0.556	0.633	0.861	0.811

Tabella 3: *Misure di performance dei diversi modelli con SMOTE.*

4.3 Analisi con CfsSubsetEval e Cross Validation

Avendo notato dei miglioramenti dei modelli nella sezione precedente, si è cercato di replicare i risultati provando ad utilizzare un numero inferiore di variabili. Infatti, essendo il dataset composto da 66 attributi, si rischia di incorrere nel problema della curse of dimensionality: un numero elevato di features potrebbe portare ad una diminuzione delle prestazioni.

Un primo metodo adottato per la feature selection è il CfSubsetEval impostato in direzione forward sul training set creato precedentemente. Il metodo ha restituito come attributi ottimali i seguenti: *Humidity3pm*, *Pressure9am*, *No_RainToday* e *N_WindDir9am*. Attraverso un nodo *Column Filter* sono state selezionate solo le variabili precedenti e si è scelto di validare i modelli utilizzando una 5-fold Cross Validation. Questa tecnica consiste nel suddividere il dataset in 5 parti di uguale numerosità, ognuna delle quali viene usata una sola volta come test set, mentre il resto come training. Le misure di performance ottenute con questo schema di validazione sono la media delle stesse misure di ogni passo. Il metodo di sampling è stato inserito nel ciclo della Cross Validation, per evitare overfitting. [1]

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.748	0.484	0.587	0.828	0.767
Random Forest	0.692	0.411	0.515	0.774	0.712
Logistic	0.757	0.480	0.588	0.841	0.764
Naïve Bayes	0.730	0.489	0.586	0.836	0.771
SPegasos	0.752	0.483	0.588	0.761	0.767
MLP	0.752	0.483	0.588	0.841	0.767

Tabella 4: *Misure di performance dei diversi modelli con Filter e Equal Size Sampling*

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.692	0.518	0.592	0.826	0.789
Random Forest	0.500	0.523	0.585	0.761	0.788
Logistic	0.762	0.474	0.585	0.840	0.760
Naïve Bayes	0.728	0.468	0.570	0.825	0.756
SPegasos	0.977	0.264	0.415	0.600	0.390
MLP	0.735	0.490	0.589	0.839	0.772

Tabella 5: *Misure di performance dei diversi modelli con Filter e SMOTE.*

Dai risultati ottenuti utilizzando l'undersampling (illustrati nella Tabella 4) emerge nuovamente che la Regressione Logistica e il MultiLayer Perceptron hanno i più alti valori di Recall, F_1-measure e AUC e questo porta a considerarli i modelli migliori.

In Tabella 5 si può osservare che l'uso della tecnica SMOTE ha fatto emergere un altro modello con performance paragonabili a quelle della Regressione Logistica e MultiLayer Perceptron: il Naïve Bayes. D'altra parte, SPegasos si comporta peggio del classificatore ZeroR, il quale seleziona la classe con la più grande probabilità a priori.

4.4 Analisi con WrapperSubsetEval e Cross Validation

Una seconda tecnica adottata per ridurre la dimensionalità del dataset è stata quella del Wrapper. Si tratta di una feature selection differente da quella precedentemente illustrata nella Sezione 4.3, poiché anziché sfruttare caratteristiche generiche per cercare correlazioni con la variabile target ed individuare gli attributi ottimali, questa utilizza un classificatore al fine di massimizzare una misura prefissata, in questo caso la F_1-measure. I nodi AttributeSelectedClassifier sono stati impostati con metodo di ricerca GreedyStepwise e direzione ancora forward. Quest'ultimi sono stati addestrati su un Validation set, definito come il 30% dell'iniziale training set, ossia circa il 20% del dataset. Per ognuno dei sei modelli di classificazione scelti è stato adoperato un Wrapper che si basa sullo stesso classificatore. Successivamente si è deciso di procedere in modo analogo a quanto visto nella Sezione 4.3.

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.753	0.500	0.601	0.835	0.779
Random Forest	0.702	0.471	0.563	0.811	0.759
Logistic	0.770	0.527	0.625	0.869	0.796
Naïve Bayes	0.741	0.511	0.605	0.850	0.785
SPegasos	0.762	0.534	0.628	0.786	0.800
MLP	0.781	0.524	0.627	0.869	0.794

Tabella 6: *Misure di performance dei diversi modelli con Wrapper e Equal Size Sampling.*

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.622	0.570	0.595	0.820	0.812
Random Forest	0.689	0.484	0.568	0.811	0.768
Logistic	0.766	0.524	0.622	0.866	0.794
Naïve Bayes	0.728	0.480	0.579	0.830	0.765
SPegasos	0.971	0.300	0.458	0.663	0.492
MLP	0.779	0.513	0.619	0.865	0.787

Tabella 7: *Misure di performance dei diversi modelli con Wrapper e SMOTE.*

Pur essendo computazionalmente più pesante, il metodo Wrapper si è rivelato migliore rispetto al Filter. Inoltre, i risultati sono paragonabili a quelli i cui modelli non sono stati sottoposti a tecniche di feature selection, ma solamente di sampling.

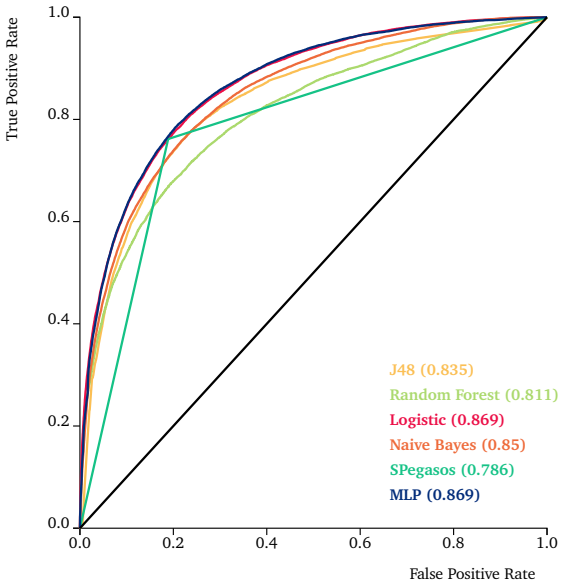


Figura 3: *Curve ROC dei classificatori con Equal Size Sampling e Wrapper.*

In definitiva, è stato scelto come approccio migliore, tra quelli proposti fino ad ora, l'Equal Size Sampling unito ad una feature selection con metodo Wrapper. Il motivo risiede nelle migliori misure di performance e nel costo computazionale decisamente inferiore a quello richiesto dallo SMOTE. In Figura 3 vengono confrontate le curve ROC dei diversi classificatori, confermando come modelli preferibili la Regressione Logistica e il MultiLayer Perceptron.

4.5 Feature significative

Attraverso le due tecniche di feature selection illustrate sono stati individuati diversi set di attributi ottimali. Nel primo caso, quello con Filter, le variabili selezionate sono già state esposte nella Sezione 4.3. Nel secondo invece, si sono ottenuti sei differenti insiemi di attributi a seconda del classificatore utilizzato nel Wrapper. Innanzitutto, si riscontra sempre la scelta dell'attributo *Humidity3pm*, mentre quella di *Pressure3pm* e *WindGustSpeed* in cinque casi, poiché fa eccezione il Random Forest. La variabile *No_RainToday*, già ottenuta con il metodo Filter, è selezionata in quattro casi. Anche gli attributi che riguardano la direzione del vento, sia nei due orari 09:00 e 15:00 sia della raffica più forte, vengono individuati come significativi per l'analisi predittiva della variabile target, pur non mostrando il prevalere di alcun punto cardinale sugli altri. Al contrario, *MinTemp* e *MaxTemp* non sembrano feature rilevanti, risultando solo dal Wrapper che utilizza i classificatori Naïve Bayes e MultiLayer Perceptron rispettivamente.

4.6 Analisi dei costi

Un ultimo e differente approccio per affrontare il problema della *class imbalance* è l'analisi dei costi. I classificatori sono stati addestrati utilizzando una matrice di costo, creata appositamente per ogni modello. Essendo la *Recall* la misura ritenuta più importante da massimizzare, si è scelto di utilizzare una matrice della seguente forma:

$$\begin{pmatrix} 0 & 1 \\ \mathrm{K} & 0 \end{pmatrix} \text{ con } K \in \{n \in \mathbb{N}^+ \mid n \leq 10\}.$$

In questo modo l'osservazione è classificata come positiva se

$$P_{\text{Yes}} \geq \frac{1}{K+1},$$

dove P_{Yes} indica la probabilità che *RainTomorrow*=Yes data come output dal classificatore [6]. Al crescere di K aumenta la *Recall* e diminuisce la *Precision* e per mantenerle in equilibrio è stato scelto per ogni modello il valore di K che massimizza la F_1-*measure*.

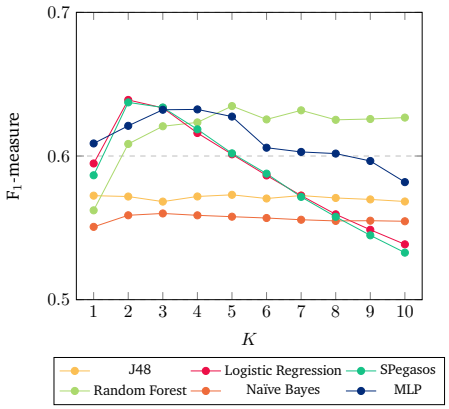


Figura 4: *Valori della F_1-measure al variare di K.*

	J48	Random Forest	Logistic	Naïve Bayes	SPegasos	MLP
K	5	5	2	3	2	4

Tabella 8: *I valori di K selezionati per ogni modello.*

Nella tabella 9 sono stati riportati i risultati ottenuti. Quello che è possibile osservare è che la Regressione Logistica e il Random Forest sono i modelli che hanno ottenuto tra i più alti valori di AUC e F_1-measure, bilanciando la Recall e la Precision. Ciò viene evidenziato anche in Figura 5, nella quale sono rappresentate le curve ROC dei diversi classificatori.

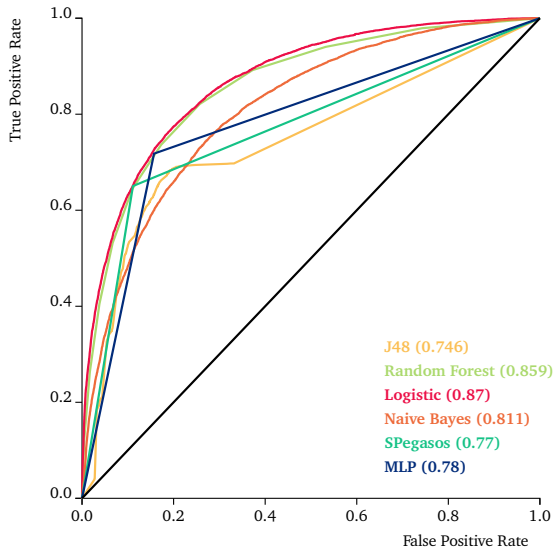


Figura 5: *Curve ROC dei classificatori con analisi dei costi.*

Tuttavia è possibile osservare che i classificatori non raggiungono livelli di performance migliori di quelli ottenuti con i metodi precedenti. È possibile imputare ciò ad una scelta non accurata della matrice di costo, ricordando che la selezione dei coefficienti è stata soggettiva.

	Recall	Precision	F_1-measure	AUC	Accuracy
J48	0.688	0.491	0.573	0.746	0.773
Random Forest	0.645	0.625	0.635	0.859	0.836
Logistic	0.661	0.619	0.639	0.870	0.835
Naïve Bayes	0.636	0.500	0.560	0.811	0.778
SPegasos	0.650	0.625	0.637	0.770	0.836
MLP	0.718	0.565	0.632	0.780	0.815

Tabella 9: *Misure di performance dei diversi modelli con Cost Analysis.*

5 Conclusioni

L'obiettivo principale di questo elaborato è stato quello di analizzare diverse tecniche di classificazione al fine di predire se domani pioverà o meno in Australia. Il dataset utilizzato non era però esente da differenti complicazioni, che hanno reso necessario un corposo lavoro di data preprocessing. In particolare, sono stati affrontati i problemi di presenza di numerosi valori mancanti, class imbalance e variabili categoriche. La trasformazione di quest'ultime, utilizzando la One-Hot Encoding, ha portato ad un aumento della dimensionalità del dataset, che avrebbe potuto influenzare negativamente le prestazioni dei modelli. Per questo motivo, si è scelto di effettuare una feature selection, usando le tecniche Filter e Wrapper. Si è cercato inoltre di migliorare il rendimento dei classificatori introducendo delle matrici di costo, che non han dato però i risultati sperati.

In conclusione, come era lecito aspettarsi, la previsione delle precipitazioni non è un problema di facile soluzione, in quanto entrano in gioco molteplici fattori. Nonostante i modelli non diano risultati pessimi, si è ancora lontani dall'affidabilità necessaria per questo genere di situazioni. Tra i classificatori utilizzati quelli che hanno performato meglio sono stati la Regressione Logistica e le reti neurali artificiali, nello specifico il MultiLayer Perceptron. Non sono state individuate delle variabili effettivamente rilevanti, oltre quelle già ritenute importanti, quali l'umidità e la pressione. Si potrebbero suggerire alcune idee per migliorare i modelli, come fornire ulteriori misurazioni di questi ultimi due attributi o rilevare con maggiore accuratezza i dati riguardanti le ore di sole, l'evaporazione e la nuvolosità, le quali sono state eliminate a causa della limitata quantità di informazione a disposizione.

Riferimenti bibliografici

[1] Marco Altini. *Dealing with imbalanced data: undersampling, oversampling and proper cross-validation*. URL: https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation.

[2] Tan Pang-Ning, Steinbach Michael e Kumar Vipin. *Introduction to Data Mining*. Pearson College.

[3] *Rain in Australia | Kaggle*. URL: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.

[4] Stacey Ronaghan. *Data Preparation for Machine Learning: Cleansing, Transformation Feature Engineering*. URL: https://towardsdatascience.com/data-preparation-for-machine-learning-cleansing-transformation-feature-engineering-d2334079b06d.

[5] *The Number of Hidden Layers*. URL: https://www.heatonresearch.com/2017/06/01/hidden-layers.html.

[6] Ian H. Witten. *More Data Mining with Weka - Class 4*. URL: https://www.cs.waikato.ac.nz/ml/weka/mooc/moredataminingwithweka/slides/Class4-MoreDataMiningWithWeka-2014.pdf.