Classificazione Type_Weather

Studente: Salvatore Paparella

Matricola: 779573

Email:

s.paparella25@studenti.uniba.it

Studentessa: Sara Paparella

Matricola: 786607

Email:

s.paparella26@studenti.uniba.it

Anno: 2025/2026

Link github: https://github.com/Paparella-Salvo/Icon-2025-26.git

Indice

1. Introduzione

Il progetto *Type Weather* ha l'obiettivo di sviluppare e valutare un sistema di classificazione del tipo di meteo (Sunny, Cloudy, Rainy, Snowy) a partire da un dataset meteorologico sintetico. L'interesse principale è l'analisi comparativa di diversi modelli di

apprendimento e, la costruzione e valutazione di una **Knowledge Base probabilistica** basata su Reti Bayesiane.

Il lavoro si articola in tre direzioni principali:

- **Analisi esplorativa del dataset**, con studio delle distribuzioni e delle correlazioni tra le variabili per comprendere quali feature risultano maggiormente informative.
- **Sperimentazione supervisionata**, confrontando più modelli di apprendimento (KNN, Decision Tree, Random Forest, SVM) tramite una procedura di valutazione robusta basata su più run indipendenti e K-Fold Cross-Validation.
- **Costruzione e analisi di Reti Bayesiane**, sia tramite apprendimento automatico della struttura (Search & Score con K2 e BIC) sia tramite una struttura definita manualmente sulla base di conoscenza esperta, con valutazione dell'impatto della complessità strutturale sull'inferenza.

Si è utilizzata la seguente metodologia per le varie fasi;

1. **pre-elaborazione dei dati**, con pulizia, codifica delle variabili categoriche e discretizzazione specifica per l'apprendimento probabilistico;
2. **scelte tecniche motivate** per ogni modello, con particolare attenzione ai parametri, ai criteri di scoring e ai metodi di inferenza;
3. **valutazione quantitativa rigorosa**, basata su medie e deviazioni standard di Accuracy e F1-score, evitando analisi su singoli run;
4. **analisi della Knowledge Base**, considerando rappresentazione, struttura, complessità delle CPT e costo computazionale dell'inferenza.

L'obiettivo finale è mostrare come modelli supervisionati e modelli probabilistici affrontino in modo diverso lo stesso problema, evidenziando vantaggi, limiti e implicazioni in termini di rappresentazione della conoscenza e capacità di generalizzazione.

2. Analisi Dataset

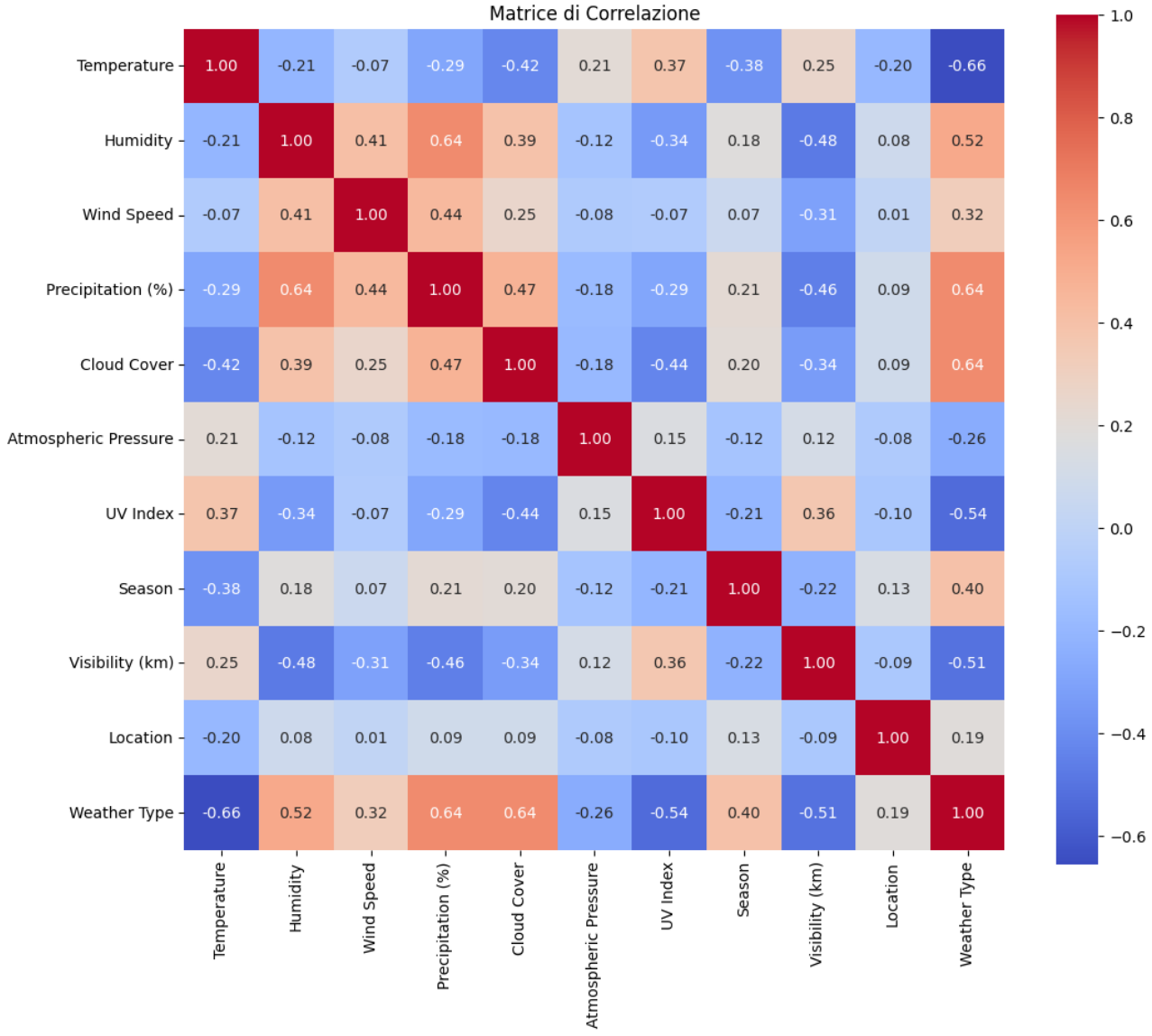
Il dataset utilizzato proviene da una raccolta sintetica disponibile su kaggle composto d 13.200 record. Ogni istanza descrive lo stato atmosferico tramite 10 variabili (numeriche e categoriche) e una variabile target (Weather Type) con quattro tipi di valori: *Sunny*, *Cloudy*, *Rainy*, *Snowy*. La distribuzione delle classi è ben bilanciata.

Le feature includono misure fisiche (Temperature, Humidity, Wind Speed, Atmospheric Pressure, Visibility), indicatori ambientali (UV Index, Precipitation), e variabili contestuali (Season, Location, Cloud Cover). Le variabili categoriche sono state convertite in valori simbolici numerici per l'addestramento dei modelli supervisionati, mentre le variabili continue sono state successivamente discretizzate per l'apprendimento delle Reti Bayesiane.

3. Pre-elaborazione dei dati

La selezione delle feature è un passaggio chiave per massimizzare la qualità del dataset, assicurando così l'efficienza e la precisione dei modelli. Questa fase consiste nel:

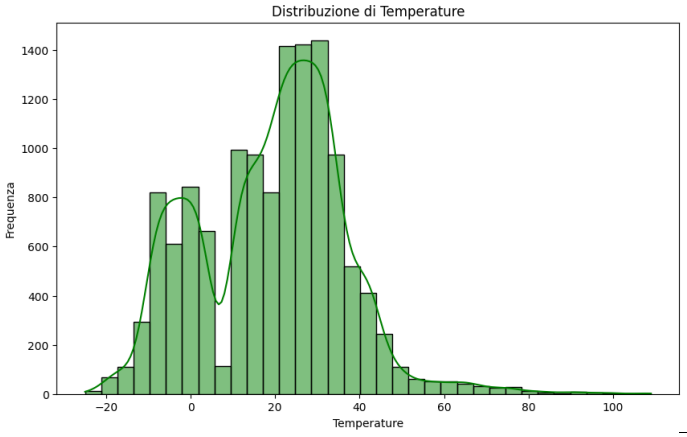
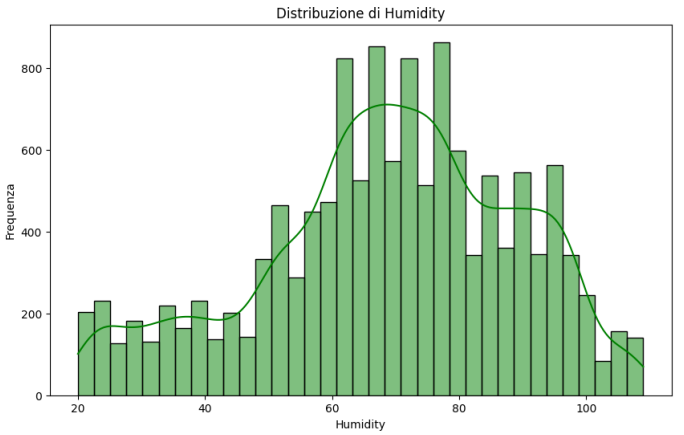
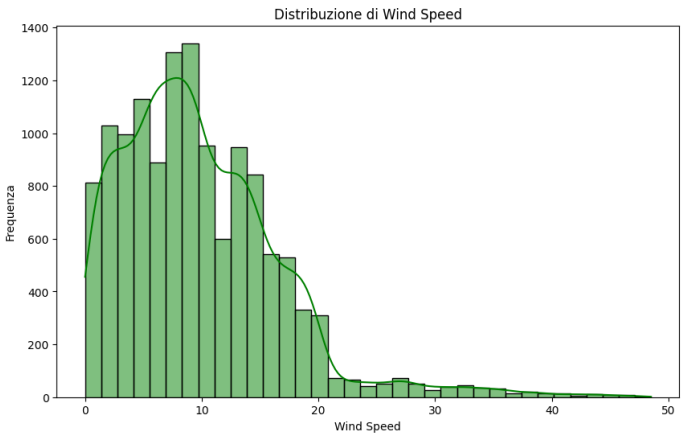
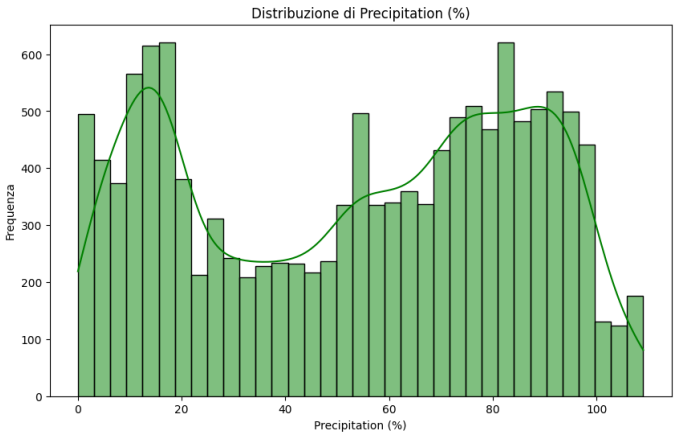
- pulizia del dataset, eliminare feature irrilevanti o ridondanti;
- conversione delle variabili categoriche in codici numerici simbolici (Cloud Cover, Season, Location, Weather Type);
- analisi delle correlazioni tra feature, tramite matrice di correlazione e visualizzazione grafica.

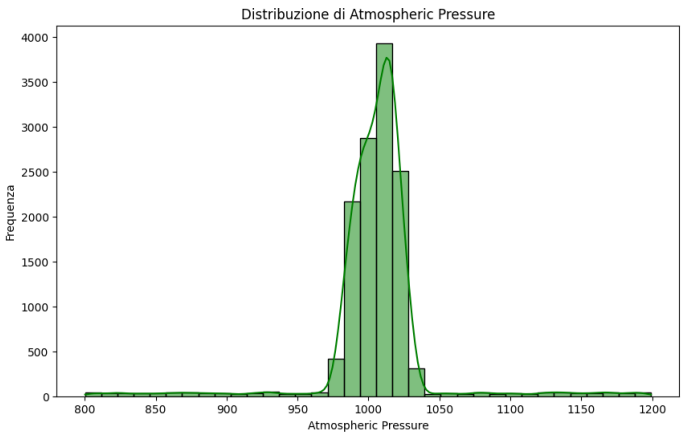
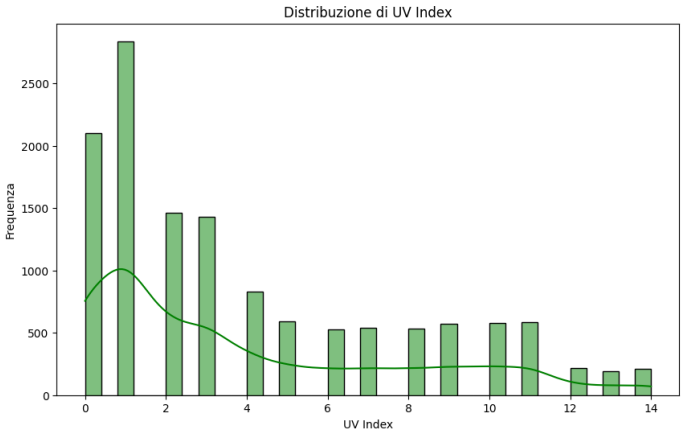
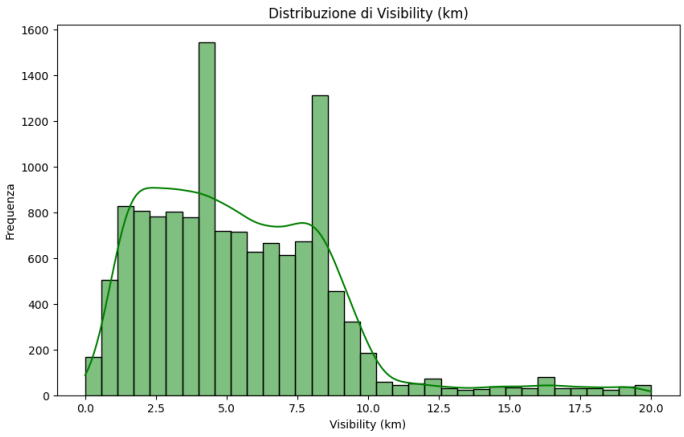


Si può osservare che non c'è nessuna coppia di feature altamente correlata, tutte le feature apportano informazioni distinte, e non è stato necessario eliminare variabili per ridondanza. Le correlazioni più significative sono:

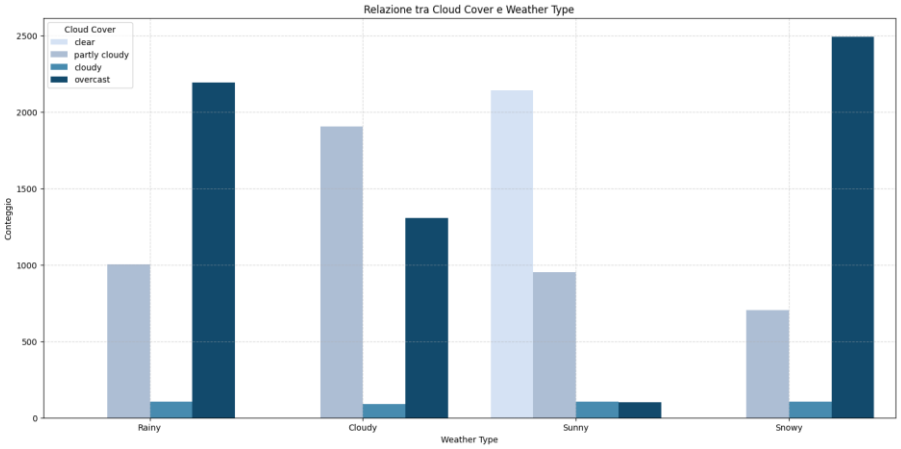
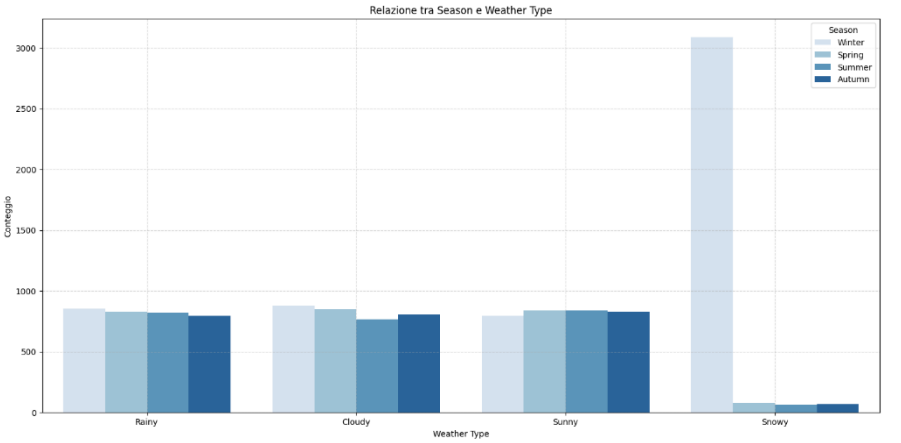
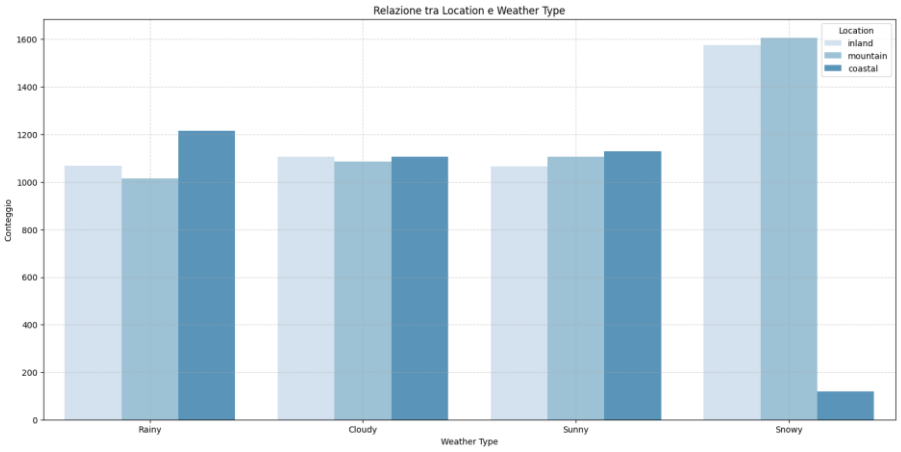
- tra *Cloud Cover* e *Weather Type* (0.60),
- tra *Precipitation (%)* e *Weather Type* (0.51),
- tra *Temperature* e *Weather Type* (-0.54).

Le distribuzioni delle variabili numeriche sono state analizzate tramite istogrammi e KDE plots

	Il grafico evidenzia una chiara **bimodalità** nella distribuzione delle temperature, con due picchi distinti: uno attorno agli **0°C**, rappresentativo delle stagioni fredde, e un altro intorno ai **30°C**, tipico delle stagioni calde. A separare i le due tipologie di temperature vi è un calo drastico per quanto riguarda temperature intoro ai **10°C.**
	La distribuzione dell'umidità è **unimodale** e **asimmetrica verso sinistra**, con la maggior parte dei valori concentrati tra il **70%** e il **90%**, indicando condizioni generalmente umide. I valori più bassi (20–40%) compaiono raramente, come mostra la coda sinistra della curva KDE, segnalando che livelli di umidità bassa sono poco frequenti nel dataset.
	La distribuzione della velocità del vento mostra una **forte asimmetria positiva**, con la maggior parte delle osservazioni concentrate su **venti deboli (0–20 km/h)** e una **lunga coda destra** che rappresenta **episodi rari di vento forte**. Questa forma, tipica di una distribuzione **log-normale**, indica che nel dataset prevalgono **condizioni di calma o brezza leggera**, mentre gli **eventi di vento intenso** sono sporadici e meno frequenti.
	La distribuzione delle precipitazioni è **bimodale e molto variabile**, con due concentrazioni principali di valori: una tra il **10–20%** (giorni secchi) e una tra l'**80–90%** (giorni piovosi). Le **precipitazioni moderate (30–60%)** risultano poco frequenti, evidenziando una netta distinzione tra **condizioni asciutte e condizioni molto umide** nel dataset.

	La distribuzione della pressione atmosferica è **quasi normale**, con la maggior parte dei valori concentrati attorno ai **1000 hPa** e una **bassa variabilità** complessiva (intervallo principale 980–1020 hPa). Gli **outlier** agli estremi, rari e isolati, riflettono probabilmente **condizioni anomale o errori di misura**, confermando nel complesso un **andamento stabile e regolare**, tipico della pressione in situazioni meteorologiche normali.
	La distribuzione dell'indice UV è **fortemente asimmetrica a destra**, con la maggior parte delle osservazioni concentrate su **valori molto bassi (0–1)** e una **lunga coda verso valori elevati** (fino a 15), che rappresentano **eventi rari di alta radiazione solare**. Questo pattern evidenzia che **condizioni di basso UV sono la norma**, mentre i livelli moderati o alti sono **eccezionali** — un risultato coerente con le altre variabili del dataset, che descrivono scenari prevalentemente freddi, umidi e poco ventosi.
	La distribuzione della **visibilità** risulta **quasi uniforme**, senza picchi dominanti: tutti i livelli, da **scarsa (0 km)** a **ottima (10 km)**, compaiono con **frequenza simile**. Questo indica che **non esiste una condizione di visibilità tipica o prevalente**, ma piuttosto una **rappresentazione equilibrata** di tutte le situazioni possibili — un aspetto che, come per la precipitazione, suggerisce un **campionamento omogeneo** delle diverse condizioni atmosferiche nel dataset.

Successivamente alle feature numeriche esaminiamo le feature categoriche con la feature target

	Il grafico evidenzia una **correlazione chiara e coerente** tra la **copertura nuvolosa** e il **tipo di tempo**: le condizioni **soleggiate** si associano quasi esclusivamente a cieli **sereni (clear)**, mentre le situazioni **piovose o nevose** corrispondono prevalentemente a **copertura totale (overcast)**. Le categorie intermedie risultano meno frequenti, suggerendo classificazioni meteorologiche piuttosto nette.
	Le condizioni Rainy, Cloudy e Sunny presentano una distribuzione stagionale piuttosto omogenea, mentre la condizione ‘Snowy’ è quasi esclusivamente associata all’inverno.
	Il grafico mostra una distribuzione abbastanza equilibrata delle tre Location per le condizioni Cloudy e Sunny. Si osserva un incremento della categoria **Coastal** in corrispondenza del tempo **Rainy**. La principale differenza riguarda la condizione Snowy, dove la presenza di osservazioni nella categoria Coastal è quasi nulla, mentre Inland e Mountain risultano nettamente più rappresentate.

Queste analisi hanno guidato la successiva discretizzazione per l'apprendimento probabilistico.
Il dataset è stato suddiviso in **training e test** con stratificazione della variabile target, mantenendo la distribuzione bilanciata delle classi (2.640 campioni per ciascuna).
Le feature sono state standardizzate con *StandardScaler*, adattato solo sul training set per evitare data leakage.
Nonostante il bilanciamento iniziale, è stato applicato **SMOTE** sul training set per aumentare la robustezza dei modelli e migliorare la separabilità locale. Il test set è rimasto inalterato per garantire una valutazione realistica delle capacità di generalizzazione.

4. Scelte Tecniche

Le scelte progettuali sono state guidate da esigenze pratiche e dai principi di rappresentazione e ragionamento trattati nel corso.

Il dataset meteorologico, composto da 13.200 feature eterogenee, permette di modellare un dominio ricco di dipendenze. L'analisi delle correlazioni non ha evidenziato ridondanze significative, per cui tutte le feature sono state mantenute.

Le variabili continue sono state discretizzate tramite soglie meteorologiche significative (temperature sotto 0°C, livelli di umidità, classi di vento) così da rendere i dati compatibili con l'apprendimento strutturale della BN e incorporare conoscenza esperta nella definizione dei range.

Sono stati selezionati quattro modelli con comportamenti complementari: KNN, Decision Tree, Random Forest, SVM. Questo permette di confrontare i risultati dei vari modelli.

Sono stati adottati due criteri di scoring:

- **K2**, che massimizza la probabilità del modello,
- **BIC**, che penalizza la complessità.

Il confronto tra i due consente di valutare l'impatto della complessità strutturale sulla capacità predittiva. Inoltre, è stata, anche, definita anche una **struttura esperta**, realizzata sulla base di relazioni causali note nel dominio meteorologico (ad esempio *Season* → *Temperature*, *Cloud Cover* → *Weather Type*).

L'inferenza è stata effettuata tramite **Variable Elimination**, metodo esatto che permette di calcolare la distribuzione MAP e di analizzare direttamente il costo computazionale legato alla struttura appresa.

5. Rappresentazione della Conoscenza

Nel progetto, la rappresentazione della conoscenza è affidata alla **Rete Bayesiana**, che costituisce la Knowledge Base (KB) del sistema.Ogni nodo rappresenta un concetto meteorologico rilevante (Temperature, Humidity, Cloud Cover, Atmospheric Pressure, Weather Type, ecc.), ottenuto dopo la discretizzazione delle variabili continue.
Gli archi rappresentano dipendenze dirette tra nodi. Nella struttura esperta, tali dipendenze riflettono relazioni causali note nel dominio (es. *Season* → *Temperature*).
Nelle strutture apprese automaticamente, invece, gli archi derivano da correlazioni statistiche individuate tramite Search & Score (K2, BIC).
Le tabelle di probabilità condizionata (CPT) non sono definite manualmente, ma vengono apprese automaticamente dai dati tramite *Maximum Likelihood Estimation*, utilizzando il

metodo *model.fit()* di *pgmpy*. Esse consentono di stimare la componente quantitativa della Knowledge Base e descrivono come la probabilità di ciascun nodo dipenda dai suoi genitori.
La discretizzazione trasforma variabili continue in categorie semantiche, rendendo la rappresentazione più interpretabile e compatibile con l'apprendimento strutturale, facilitando la costruzione di una KB coerente con il dominio meteorologico.

6. Apprendimento Supervisionato

L'obiettivo dell'apprendimento supervisionato è sviluppare modelli in grado di classificare il tipo di tempo (Weather Type) a partire dalle variabili atmosferiche e contestuali. Il problema si configura come un compito di classificazione multi-classe.

Per garantire una valutazione robusta, è stato adottato un approccio comparativo basato su quattro algoritmi con caratteristiche complementari:

- **K-Nearest Neighbors (KNN)**
- **Albero Decisionale (Decision Tree)**
- **Random Forest**
- **Support Vector Machine (SVM)**

La selezione degli iperparametri e la valutazione dei modelli sono state effettuate tramite una **K-Fold Cross-Validation** (K=5) applicata al training set, normalizzato e bilanciato tramite SMOTE.

Per ogni modello è stata definita una lista di configurazioni di iperparametri. Ogni configurazione è stata valutata seguendo il processo:

1. suddivisione del training set in 5 sottoinsiemi;
2. addestramento su 4 sottoinsiemi e validazione sul sottoinsieme rimanente;
3. calcolo della **F1-score pesata** per ogni sottoinsieme;
4. media delle prestazioni sui 5 sottoinsiemi.

La configurazione con la F1-score media più alta è stata selezionata come migliore.

Per stimare la stabilità dei modelli, dopo la selezione degli iperparametri ogni algoritmo è stato eseguito **5 volte in modo indipendente**. In ciascuna iterazione è stato generato un nuovo **train/test split stratificato**, così da ottenere training set e test set differenti a ogni run. Questo approccio permette di valutare la variabilità delle prestazioni e di ottenere stime più affidabili di Accuracy e F1-score rispetto a una singola suddivisione dei dati.

I risultati sono stati rappresentati tramite **grafici comparativi**, che permettono di osservare in modo immediato la stabilità e la consistenza delle prestazioni dei diversi algoritmi, evidenziando eventuali differenze di comportamento tra i modelli.

6.1. k-Nearest Neighbors (KNN)

Il modello **k-Nearest Neighbors** (**KNN**) è stato valutato mediante **K-Fold Cross-Validation (K = 5)** sul training set, normalizzato e bilanciato, al fine di selezionare la configurazione ottimale degli iperparametri. Sono state testate diverse combinazioni del numero di vicini (*n_neighbors* $\in$ {5, 7}) e della strategia di pesatura (*uniform*, *distance*).

La configurazione migliore è risultata:

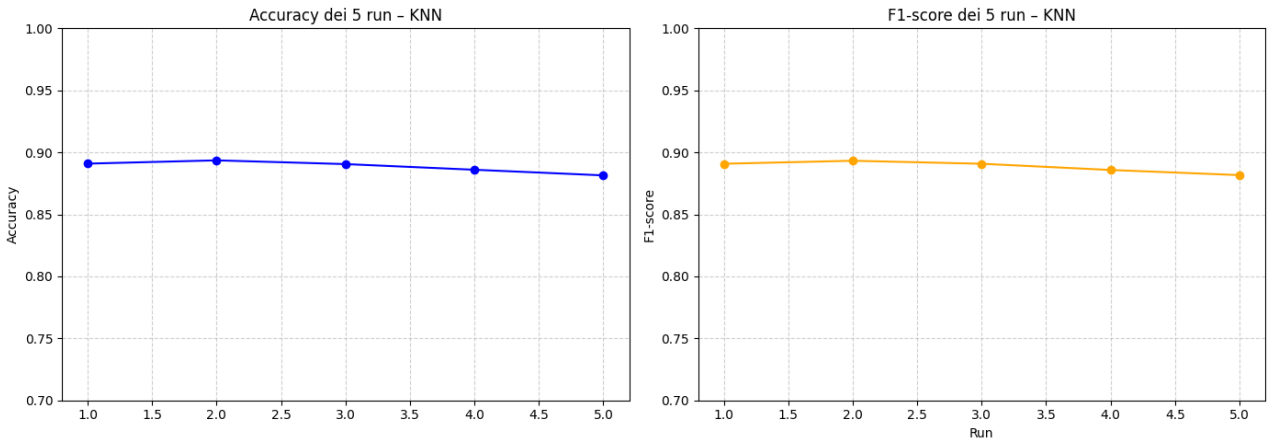
- *n_neighbors* = 7
- *weights* = 'distance'

Questa scelta consente di privilegiare i vicini più prossimi, migliorando la capacità del modello di cogliere relazioni locali tra osservazioni simili.

Dopo la selezione degli iperparametri, il modello è stato eseguito 5 volte in modo indipendente, ciascuna con un nuovo train/test split stratificato. I risultati ottenuti sul test set evidenziano una buona stabilità delle prestazioni:

- **Accuracy media**: 0.8885
- **F1-score pesata media**: 0.8885

Le metriche aggregate sono state rappresentate tramite **grafici comparativi**, che mostrano l'andamento di Accuracy e F1-score sui 5 run. Il modello si è dimostrato affidabile e consistente, con variazioni contenute tra le diverse esecuzioni.



6.2. Albero Decisionale (Decision Tree)

Il modello **Decision Tree** è stato ottimizzato tramite **K-Fold Cross-Validation (K = 5)** applicata al training set, normalizzato e bilanciato. Sono state valutate due configurazioni di iperparametri, variando esclusivamente la profondità massima dell'albero:

- *max_depth = 5*
- *max_depth = 10*

Il parametro *min_samples_split* è stato mantenuto fisso a 2, mentre *random_state = 42* ha garantito la riproducibilità dei risultati. La configurazione con la F1-score media più elevata è stata selezionata come modello ottimale.

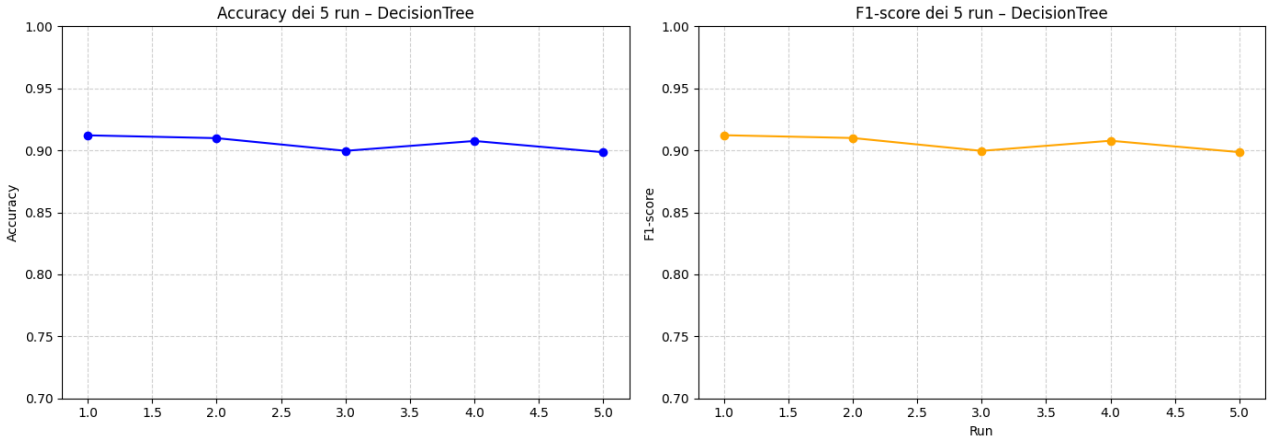
Per stimare la stabilità del modello, il Decision Tree è stato successivamente eseguito 5 volte in modo indipendente, ciascuna con un nuovo train/test split stratificato. Questo approccio consente di valutare la variabilità delle prestazioni e di ottenere una stima più affidabile della capacità di generalizzazione.

I risultati aggregati sui 5 run mostrano prestazioni solide e consistenti:

- **Accuracy media**: 0.9055

- **F1-score pesata media**: 0.9056

Le metriche sono state rappresentate tramite **grafici comparativi**, che evidenziano un comportamento stabile del modello tra le diverse esecuzioni.



6.3. Random Forest

Il modello **Random Forest** è stato ottimizzato tramite **K-Fold Cross-Validation (K = 5)** sul training set, normalizzato e bilanciato. Sono state valutate due configurazioni di iperparametri, variando il numero di alberi nella foresta:

- *n_estimators = 50*
- *n_estimators = 100*

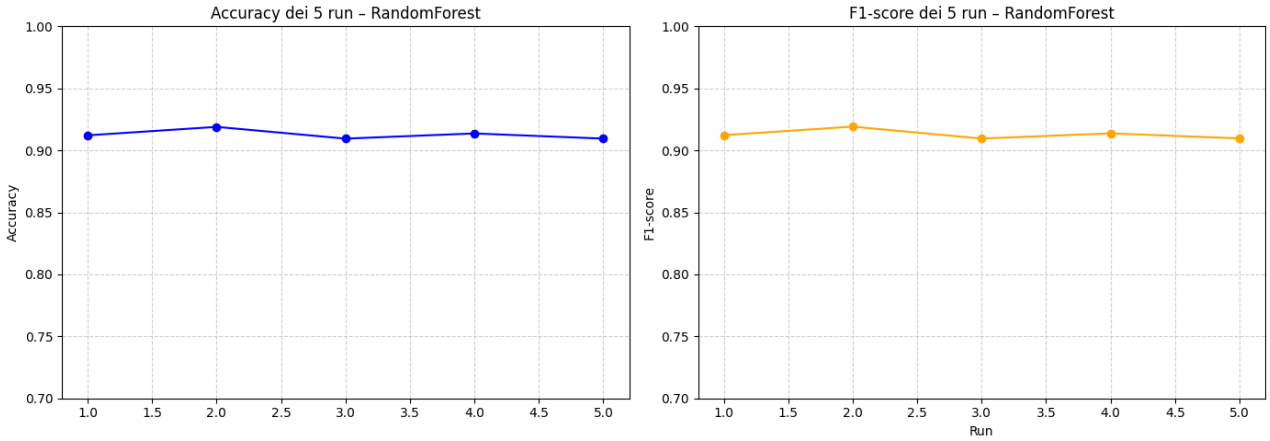
La profondità massima è stata fissata a *max_depth = 10*, mentre *random_state = 42* ha garantito la riproducibilità. L'approccio ensemble consente di combinare più alberi decisionali, riducendo la varianza rispetto a un singolo Decision Tree e migliorando la capacità di generalizzazione.

Dopo la selezione degli iperparametri, il modello è stato eseguito 5 volte in modo indipendente, ciascuna con un nuovo train/test split stratificato. Questo ha permesso di stimare la variabilità delle prestazioni e di ottenere una valutazione più robusta.

I risultati aggregati sui 5 run evidenziano prestazioni elevate e stabili:

- **Accuracy media**: 0.9127
- **F1-score pesata media**: 0.9129

Le metriche sono state rappresentate tramite **grafici comparativi**, che mostrano la consistenza del modello tra le diverse esecuzioni.



6.4. Support Vector Machine (SVM)

Il modello **Support Vector Machine (SVM)** è stato ottimizzato tramite **K-Fold Cross-Validation (K = 5)** sul training set, normalizzato e bilanciato. Sono state testate due configurazioni del parametro di regolarizzazione:

- *C = 1*
- *C = 10*

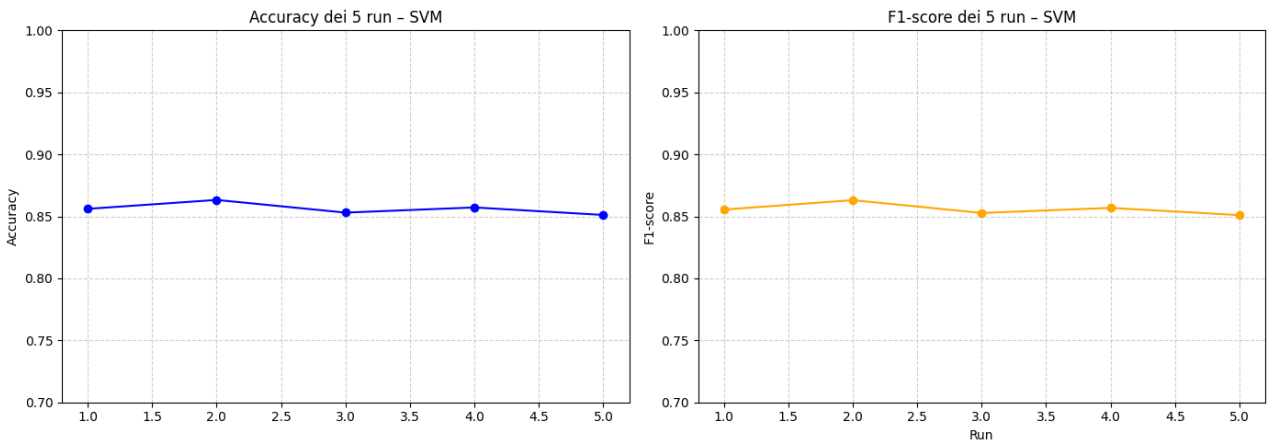
In entrambi i casi è stato utilizzato il *kernel RBF*, adatto a catturare relazioni non lineari tra le variabili. La configurazione con la F1-score media più elevata è stata selezionata come modello ottimale.

Dopo la selezione degli iperparametri, il modello è stato eseguito 5 volte in modo indipendente, ciascuna con un nuovo train/test split stratificato. Questo ha permesso di stimare la variabilità delle prestazioni e di ottenere una valutazione più robusta.

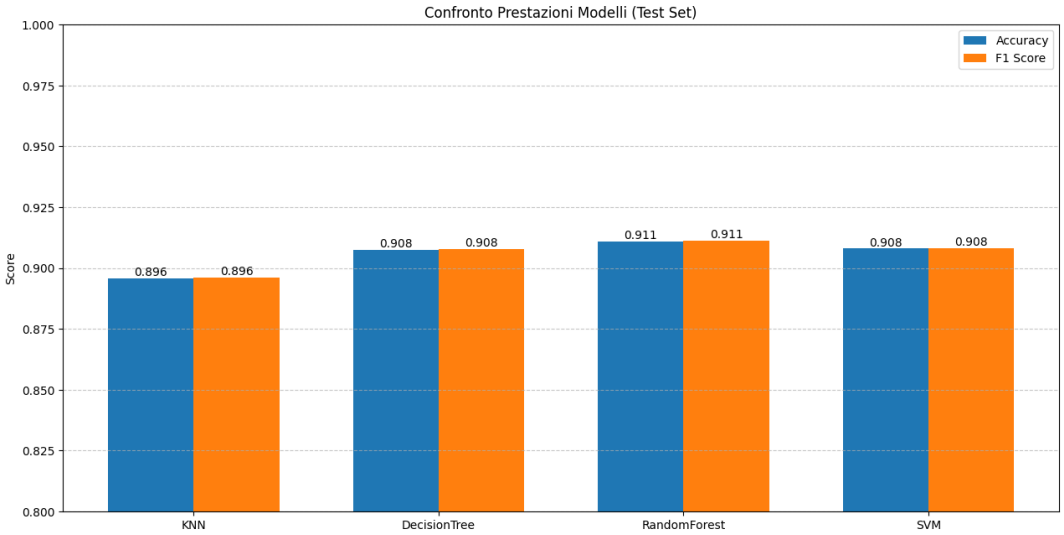
I risultati aggregati sui 5 run evidenziano prestazioni buone e coerenti:

- **Accuracy media**: 0.8561
- **F1-score pesata media**: 0.8558

Le metriche sono state rappresentate tramite **grafici comparativi**, che mostrano la stabilità del modello tra le diverse esecuzioni.



6.5. Confronto tra Modelli



Al termine della fase di sperimentazione è stato effettuato un confronto quantitativo tra i quattro algoritmi implementati. Il grafico riassume le prestazioni medie sul test set, calcolate su **5 esecuzioni indipendenti** per ciascun modello. Le metriche di riferimento sono l'**Accuracy** e la **F1-score pesata**.

Dall'analisi dei risultati emergono le seguenti considerazioni:

1. **Random Forest** si conferma il modello più performante, con **Accuracy media = 0.9127** e **F1-score media = 0.9129**. L'approccio ensemble ha dimostrato un'elevata capacità di generalizzazione e una notevole stabilità tra le diverse esecuzioni; si distingue grazie alla capacità di ridurre la varianza tramite l'aggregazione di molti alberi deboli: questo la rende particolarmente efficace in domini con interazioni non lineari e feature eterogenee come quello meteorologico. La sua stabilità sui 5 run conferma una buona robustezza rispetto alla variabilità del training set.
2. **Decision Tree** segue a breve distanza, con **Accuracy media = 0.9055** e **F1-score media = 0.9056**. Pur mostrando prestazioni leggermente inferiori rispetto alla Random Forest, mantiene un buon equilibrio tra efficacia predittiva e interpretabilità.
3. **K-Nearest Neighbors (KNN)** ottiene risultati solidi, con **Accuracy media = 0.8885** e **F1-score media = 0.8885**. Mostra una buona coerenza interna, ma la sua natura "instance-based" lo rende sensibile alla distribuzione locale dei dati e meno competitivo in domini con molte variabili. Il modello si dimostra affidabile, anche se leggermente meno efficace rispetto agli approcci basati su alberi.
4. **Support Vector Machine (SVM)** presenta le prestazioni più contenute tra i modelli analizzati, con **Accuracy media = 0.8561** e **F1-score media = 0.8558**; pur utilizzando un kernel RBF adatto a catturare relazioni non lineari, soffre la dimensionalità del problema e la presenza di feature con scale e distribuzioni molto diverse. Le prestazioni restano comunque solide, ma inferiori rispetto ai modelli ensemble.

In conclusione, tutti i modelli supervisionati hanno mostrato prestazioni elevate e consistenti. Tuttavia, sulla base delle evidenze sperimentali, la Random Forest rappresenta il miglior compromesso tra accuratezza, robustezza e capacità di generalizzazione, risultando il modello più affidabile per il problema di classificazione meteorologica affrontato

7. Bayesian Network

Nel progetto è stata utilizzata una **Rete Bayesiana** per modellare le relazioni tra le variabili del dataset meteorologico e il *Weather Type*. L'obiettivo è di **costruire e confrontare diverse strutture** per valutare come la scelta della rete influenzi la capacità predittiva del modello.

Ogni variabile del dataset è stata rappresentata come un nodo della rete e le connessioni tra nodi sono state determinate tramite tre approcci diversi:

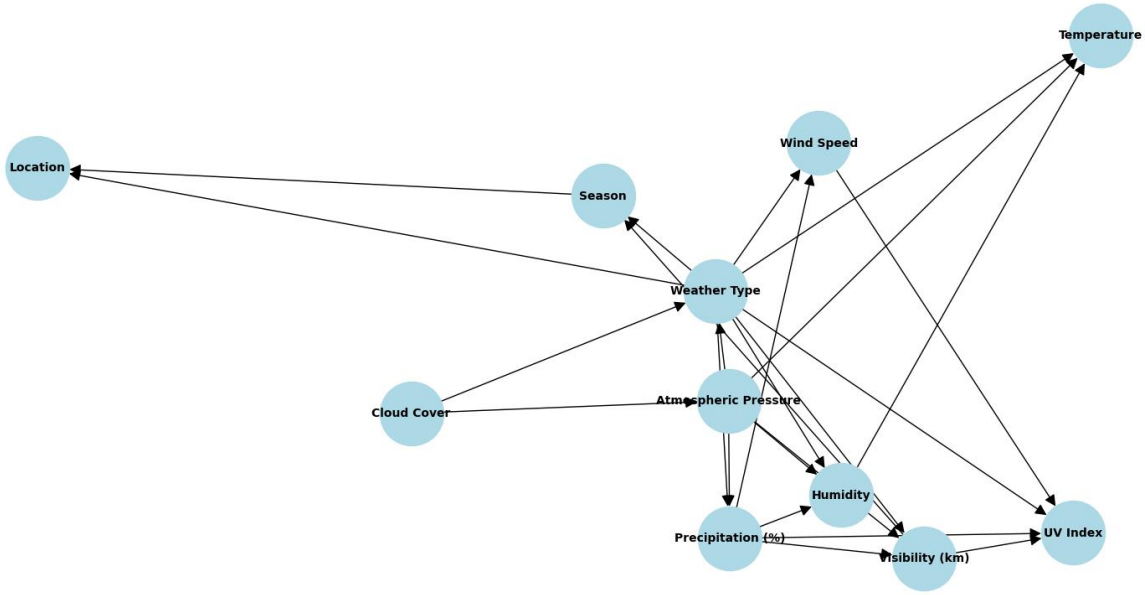
1. **Apprendimento della struttura con metodo K2**
2. **Apprendimento della struttura con punteggio BIC**
3. **Struttura definita manualmente (Expert)** basata su relazioni considerate plausibili nel contesto meteorologico;

Definita la struttura, la rete è stata addestrata sui dati discretizzati e valutata tramite accuracy sul test set. Questa parte del progetto ha permesso di confrontare modelli con strutture diverse, osservando sia la loro accuratezza sia la complessità della rete (numero di archi).

7.1. Modello Search & Score

La struttura della Rete Bayesiana è stata appresa automaticamente dai dati discretizzati tramite un approccio **Search & Score**, utilizzando l'algoritmo di **Hill Climbing** con funzione di punteggio **K2**. Questo metodo esplora lo spazio delle possibili strutture in modo greedy, selezionando gli archi che massimizzano la coerenza statistica rispetto ai

dati di addestramento.



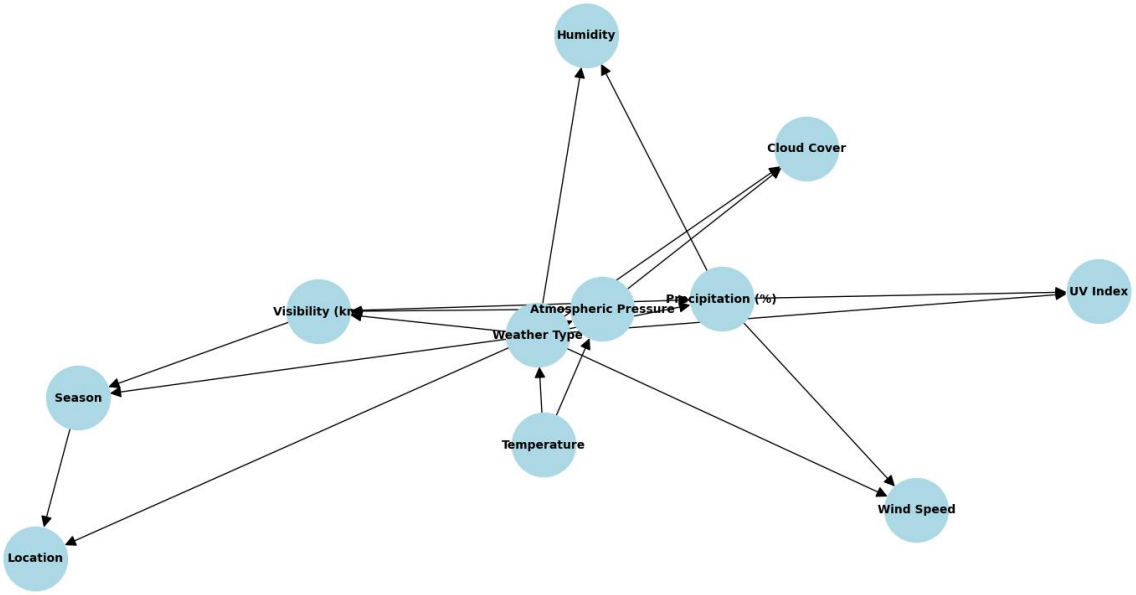
La rete risultante include tutte le variabili del dataset e presenta una struttura **ricca e articolata**, con numerose connessioni tra variabili meteorologiche e contestuali. Il nodo *Weather Type* emerge come **variabile centrale**, influenzando direttamente molte altre variabili come *Season*, *Cloud Cover*, *Atmospheric Pressure*, *Wind Speed*, *Temperature*, *Humidity*, *Precipitation*, *Visibility* e *UV Index*. Questo riflette la natura multifattoriale del fenomeno atmosferico e la capacità del modello di cogliere relazioni complesse tra le variabili.

La valutazione è stata effettuata su un test set indipendente (20%), utilizzando inferenza esatta tramite **Variable Elimination** e predizione MAP per ciascuna istanza. Il modello ha raggiunto un'**Accuracy pari a 0.8966** nella classificazione della variabile *Weather Type*.

7.2. Modello Search & Score (BIC)

La struttura della Rete Bayesiana è stata appresa automaticamente dai dati discretizzati tramite **Hill Climbing**, utilizzando come criterio di selezione il **Bayesian Information Criterion (BIC)**. Questo punteggio penalizza le strutture eccessivamente complesse,

favorendo modelli più parsimoniosi e meno soggetti a overfitting.

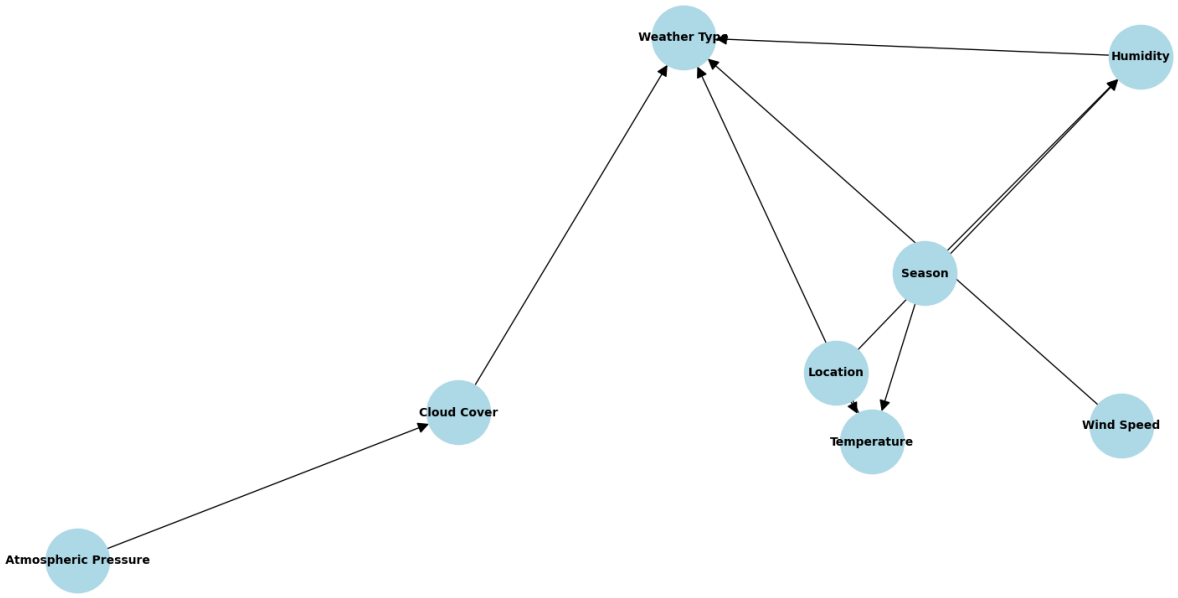


La rete appresa risulta **più contenuta** rispetto a quella ottenuta con il punteggio K2, con un numero ridotto di archi e una struttura più selettiva. Il nodo *Weather Type* mantiene un ruolo centrale, risultando influenzato da variabili chiave come *Atmospheric Pressure*, *Temperature* e *Season*. Queste connessioni riflettono relazioni meteorologiche plausibili e statisticamente significative, selezionate in base al contributo informativo effettivo.

La valutazione è stata effettuata su un test set indipendente (20%), utilizzando inferenza esatta tramite **Variable Elimination** e predizione MAP. Il modello ha raggiunto un'**Accuracy pari a 0.8939** nella classificazione della variabile *Weather Type*.

7.3. Modello Esperto

Il modello Expert utilizza una **struttura definita manualmente**, basata su relazioni meteorologiche considerate plausibili e interpretabili. La rete è stata costruita specificando direttamente gli archi tra le variabili, senza ricorrere ad algoritmi di apprendimento automatico.



La struttura proposta è **semplificata e coerente** con il dominio meteorologico:

- *Season* e *Location* forniscono il contesto generale, influenzando *Temperature*, *Humidity* e *Cloud Cover*.
- *Atmospheric Pressure* agisce su *Cloud Cover*, mentre *Location* ha un impatto diretto anche su *Weather Type*.
- Il nodo *Weather Type* integra le principali variabili atmosferiche finali, ricevendo input da *Temperature*, *Wind Speed*, *Cloud Cover* e *Location*.

La rete è stata addestrata sui dati discretizzati e valutata su un test set indipendente (20%) tramite inferenza esatta con **Variable Elimination**. Il modello ha raggiunto un'**Accuracy pari a 0.7920** nella classificazione della variabile *Weather Type*.

7.4. Complessità della KB e Rete Bayesiana

La complessità della Knowledge Base dipende dalla struttura della Rete Bayesiana e dalla dimensione delle sue tabelle di probabilità condizionata (CPT). Sono state confrontate tre reti con strutture diverse: **K2**, **BIC** e **Expert**.
Il numero di archi emersi nei tre modelli evidenzia differenze significative in termini di complessità:

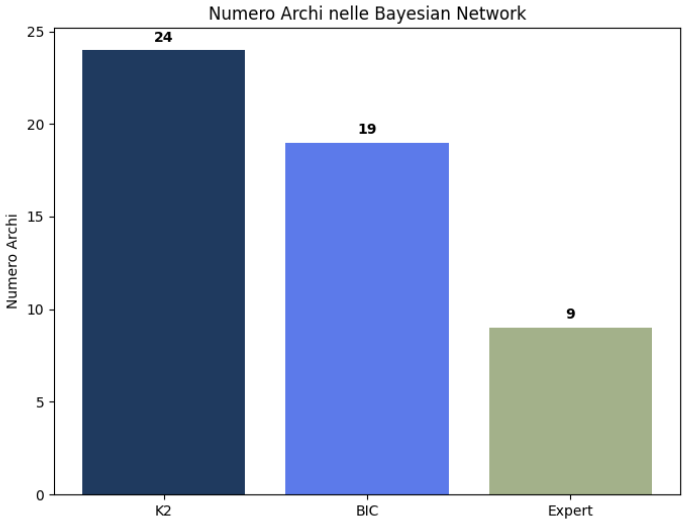
- **K2**: 24 archi
- **BIC**: 19 archi
- **Expert**: 9 archi

La rete appresa con K2 risulta la più densa, con numerose connessioni tra variabili. BIC produce una struttura più contenuta, mentre la rete esperta è la più semplice e interpretabile.

La complessità delle tabelle di probabilità condizionata (CPT) dipende dal numero di genitori di ciascun nodo. Nelle reti apprese automaticamente, il nodo *Weather Type* riceve input da molte variabili, il che implica una struttura condizionata più articolata. Nella rete esperta, invece, il numero di genitori è limitato, riducendo la complessità parametrica in fase di addestramento e inferenza.

L’inferenza è stata eseguita tramite Variable Elimination, il metodo previsto dalla libreria utilizzata. Poiché questo algoritmo risente della struttura del grafo, le reti più dense (K2 e BIC) risultano teoricamente più costose da elaborare, mentre la rete esperta, avendo meno connessioni, comporta una complessità inferenziale inferiore. Nel progetto non sono state effettuate misurazioni dei tempi di esecuzione, ma la differenza di complessità strutturale permette comunque di trarre considerazioni qualitative.

Le reti K2 e BIC sono state apprese tramite Hill Climbing, che esplora uno spazio di strutture molto ampio. L’uso di criteri di scoring ha guidato la ricerca, ma il processo rimane computazionalmente impegnativo. La rete esperta, invece, è stata definita manualmente, eliminando la fase di apprendimento strutturale.



7.5. Analisi Comparativa delle Reti Bayesiane

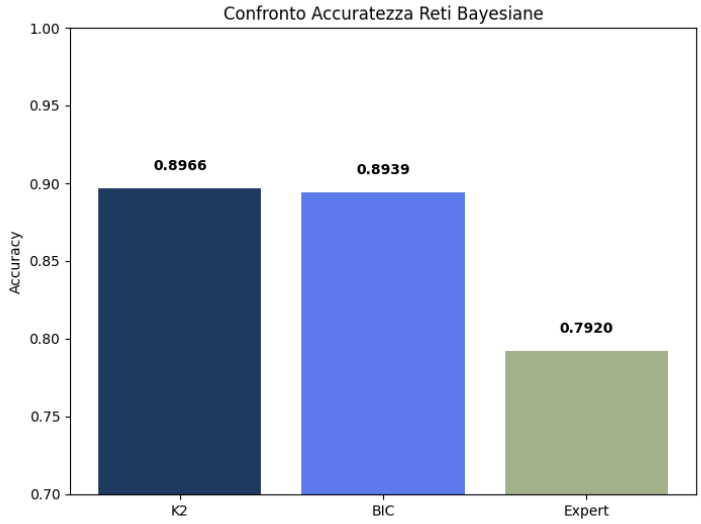
Modello	Metodo di Apprendimento Struttura	Nodi Utilizzati	Accuracy
BN_K2	Automatico (HillClimb + K2 Score)	11 (Tutti)	89.66%
BN_BIC	Automatico (HillClimb + BIC Score)	11 (Tutti)	89.39%
BN_Expert	Manuale (Expert Knowledge)	8 (Parziale)	79.20%

I modelli con struttura appresa automaticamente raggiungono le prestazioni migliori, con **K2** che ottiene l’accuratezza più elevata (**89.66%**) e **BIC** segue a breve distanza (**89.39%**). La lieve differenza indica che la penalizzazione sulla complessità introdotta dal BIC riduce marginalmente la capacità predittiva, pur mantenendo una rappresentazione strutturalmente solida.

Il modello **Expert**, basato su una struttura manuale e su un sottoinsieme di variabili, ha mostrato prestazioni significativamente inferiori (**79.20%**), evidenziando come una

modellazione eccessivamente semplificata non riesca a catturare adeguatamente le interazioni tra le variabili meteorologiche.

Nel complesso, i risultati confermano che, in un dominio caratterizzato da forti dipendenze e relazioni non lineari l'apprendimento automatico della struttura rappresenta la soluzione più efficace. I modelli K2 e BIC offrono un buon compromesso tra accuratezza predittiva e capacità descrittiva, mentre la rete esperta si distingue per la sua semplicità ma con prestazioni leggermente inferiori.



8. Ragionamento con Vincoli (CSP)

Oltre alla componente probabilistica, è stato integrato un modulo di ragionamento simbolico basato su Constraint Satisfaction Problems (CSP). Il CSP utilizza come input la predizione MAP ottenuta dalla Rete Bayesiana appresa con il metodo K2, che fornisce tre variabili discrete:

- WeatherType
- UVLevel
- TemperatureClass

Questi valori, prodotti dalla BN_K2 tramite inferenza MAP, vengono utilizzati per determinare:

- un'attività consigliata,
- un outfit adeguato,
- rispettando vincoli meteorologici.

Variabili

- Activity ∈ {Hiking, Beach, IndoorGym, CityWalk}

- Outfit ∈ {LightClothes, Jacket, Waterproof, UVProtection}
- WeatherType, UVLevel, TemperatureClass (dalla BN)

Vincoli

- Se WeatherType = Rainy → Activity ≠ Hiking
- Se WeatherType = Rainy → Outfit = Waterproof
- Se WeatherType = Snowy → Outfit = Jacket
- Se UVLevel = high → Outfit = UVProtection
- Se TemperatureClass = very_cold → Outfit = Jacket

Questi vincoli rappresentano conoscenza simbolica esplicita, non appresa dai dati, e mostrano come il sistema combini ragionamento probabilistico (BN) e ragionamento deterministico (CSP).

Esempio di output:

Predizione BN:

WeatherType = Rainy

UVLevel = low

TemperatureClass = mild

Soluzione CSP:

Activity = IndoorGym

Outfit = Waterproof

9. Connessione ai Temi del Corso

Il progetto integra diversi concetti fondamentali del corso di Ingegneria della Conoscenza.

La Knowledge Base del sistema è costituita da una Rete Bayesiana, che fornisce una rappresentazione formale del dominio meteorologico attraverso:

- variabili discrete ottenute tramite discretizzazione semantica,
- dipendenze tra concetti modellate come archi del grafo,
- distribuzioni condizionate (CPT) apprese dai dati.

La distinzione tra struttura qualitativa (grafo) e struttura quantitativa (CPT) riflette i principi fondamentali della rappresentazione della conoscenza probabilistica.

Il sistema effettua inferenza probabilistica tramite:

- Variable Elimination, metodo esatto che permette di calcolare distribuzioni posteriori e MAP,

- MAP Query, utilizzata per la predizione del tipo di meteo.

Questi meccanismi mostrano come la KB possa essere utilizzata non solo per descrivere il dominio, ma anche per ragionare su di esso.

Il progetto integra due forme di apprendimento:
-Apprendimento strutturale:

- Hill Climbing con punteggi K2 e BIC, che esplorano lo spazio delle strutture possibili.
- Confronto tra modelli appresi automaticamente e modello esperto.

Questo evidenzia il ruolo del trade-off tra complessità e capacità predittiva.

-Apprendimento parametrico:

- Stima delle CPT tramite Maximum Likelihood Estimation (MLE).

Questo collega la fase di apprendimento alla costruzione della KB.

Per la Valutazione dei modelli di ragionamento sono stati confrontati:

- modelli appresi automaticamente, più accurati ma più complessi,
- modello esperto, più interpretabile ma meno performante.
  Il confronto mostra come la complessità strutturale influenzi:
- dimensione delle CPT,
- costo dell'inferenza,
- accuratezza predittiva.

Il progetto combina tre forme di ragionamento:

- discriminativo (modelli supervisionati),
- probabilistico (Reti Bayesiane),
- simbolico (CSP).

10. Limiti e Possibili Estensioni

Limiti del progetto

- Il dataset è sintetico e non riflette pienamente la variabilità dei dati reali.
- La discretizzazione manuale può introdurre soglie arbitrarie.
- La struttura esperta è semplificata e non cattura tutte le dipendenze reali.
- La parte supervisionata è più estesa della parte di ragionamento.

Possibili estensioni

- Utilizzo di dati reali provenienti da stazioni meteorologiche.
- Introduzione di una ontologia meteorologica (OWL).

11. Conclusioni

Il progetto *Type Weather* ha analizzato diverse tecniche di classificazione per la previsione delle condizioni meteorologiche, confrontando approcci supervisionati tradizionali e modelli probabilistici basati su Reti Bayesiane. I risultati mostrano come il dominio meteorologico sia caratterizzato da forti interazioni tra variabili, richiedendo modelli in grado di catturare dipendenze complesse.

Tra i classificatori supervisionati, il **Random Forest** ha ottenuto le migliori prestazioni complessive (≈91.1%) (grazie alla capacità di ridurre la varianza e catturare interazioni non lineari.), seguito da **Decision Tree** e **SVM**, che hanno mostrato risultati comparabili e stabili. Il **KNN**, pur risultando leggermente meno performante, ha comunque confermato una buona capacità discriminativa, confermando la coerenza informativa del dataset.

Nel contesto probabilistico, le **Reti Bayesiane con struttura appresa automaticamente** hanno raggiunto accuratezze prossime all'89–90%, dimostrando una buona capacità di modellare le relazioni statistiche tra le variabili meteorologiche. Il modello basato su **conoscenza esperta**, pur essendo più interpretabile, ha invece mostrato prestazioni inferiori (≈79%), evidenziando i limiti di una struttura troppo semplificata in un dominio caratterizzato da molteplici dipendenze. Il confronto tra K2 e BIC ha inoltre mostrato il classico trade-off tra complessità strutturale e capacità predittiva.

Nel complesso, l'analisi sperimentale evidenzia come i modelli ensemble e le Reti Bayesiane apprese dai dati rappresentino le soluzioni efficaci e complementari: i primi eccellono nella predizione, mentre le seconde offrono una rappresentazione esplicita della conoscenza e un meccanismo di inferenza interpretabile.

L'integrazione di un modulo di ragionamento simbolico (CSP) ha ulteriormente mostrato come la conoscenza probabilistica possa essere utilizzata per supportare decisioni basate su vincoli, arricchendo il sistema con un livello di ragionamento deterministico.

I risultati ottenuti costituiscono una base solida per sviluppi futuri, quali l'impiego di dati reali, l'estensione a modelli temporali, l'integrazione con conoscenza esplicita tramite ontologie e un ulteriore affinamento della rappresentazione delle feature.