SISTEMA PREDITTIVO PER IL BIKE SHARING

MODELLI BAYESIANI IN R PER INFERIRE IL TASSO DI UTILIZZO DELLE BICI

Progetto di Modelli Probabilistici per le Decisioni 2017/2018

Pennati Mattia 793375
Virgilio Luca 794866

INTRODUZIONE

I servizi di bike sharing sono sempre più diffusi in tutto il mondo. Essi hanno ripercussioni nell'ambito della salute, del traffico e in quello ambientale

Per questo motivo, abbiamo costruito alcuni modelli che si pongono come obiettivo la stima della quantità di bici utilizzate in un dato momento, sotto determinate condizioni, in un servizio di bike sharing

DATASET

Per il nostro progetto, abbiamo scelto un dataset disponibile sul sito UCI Machine Learning, all'indirizzo https://archive.ics.uci.edu/ml/datasets/bike+sharing+data set

Il dataset è composto da dati raccolti dai primi due anni del servizio *Capital Bikeshare* di Washington, DC, dal sito *meteofree* e dal calendario del dipartimento di risorse umane degli USA

DATASET

Feature	Descrizione	Continuo o discreto
instant	indice univoco del record	discreto
dteday	data	discreto
season	stagione	discreto
yr	anno	discreto
mnth	mese	discreto
hr	ora	discreto
holiday	giorno festivo	discreto (binario)
weekday	giorno della settimana	discreto
workingday	giornata lavorativa (sabato e domenica esclusi)	discreto
weathersit	condizioni metereologiche	discreto

DATASET

Feature	Descrizione	Continuo o discreto
temp	temperatura in Celsius normalizzata	continuo
atemp	temperatura percepita in Celsius normalizzata	continuo
hum	valore di umidità in % normalizzato	continuo
windspeed	velocità del vento in m/s normalizzata	continuo
casual	numero di bici prelevate da utenti non registrati	discreto
registered	numero di bici prelevate da utenti registrati	discreto
cnt	numero totale di bici prelevate	discreto

PRE-PROCESSING

Prima della definizione del modello, abbiamo eseguito una fase di pulizia dei dati:

1. Eliminazione di features: instant, dteday, casual e registered
2. Discretizzazione di features continue: hum, windspeed, temp, atemp
3. ulteriore discretizzazione della variabile target cnt

Le scelte finali per queste operazioni sono state effettuate tramite l'ausilio di: matrice di correlazione, conoscenze del dominio, numerosità dei vari intervalli delle discretizzazioni e risultati ottenuti dai modelli con diverse discretizzazioni

DEFINIZIONE DEI MODELLI

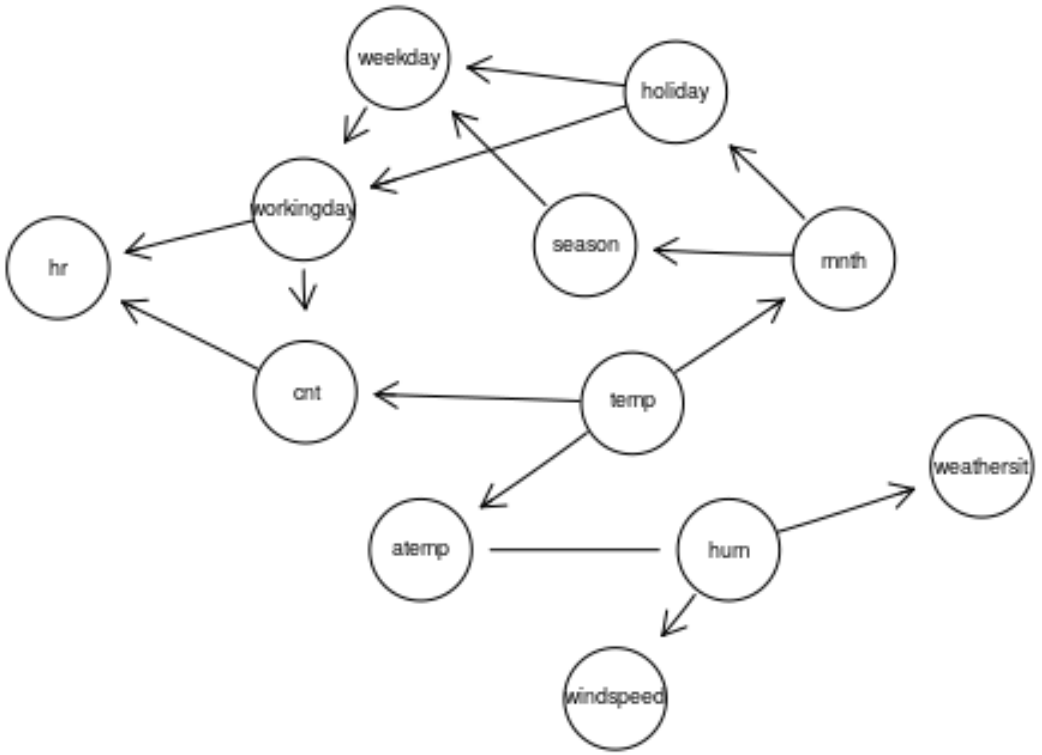
Per definire il modello delle reti bayesiane sono stati confrontati diversi metodi e algoritmi appartenenti alle classi di:

1. Constraint-Based structure learning algorithms
2. Score-Based structure learning algorithms
    1. Score-based structure learing algorithms applicati a reti esistenti
3. Hybrid structure learning algorithms

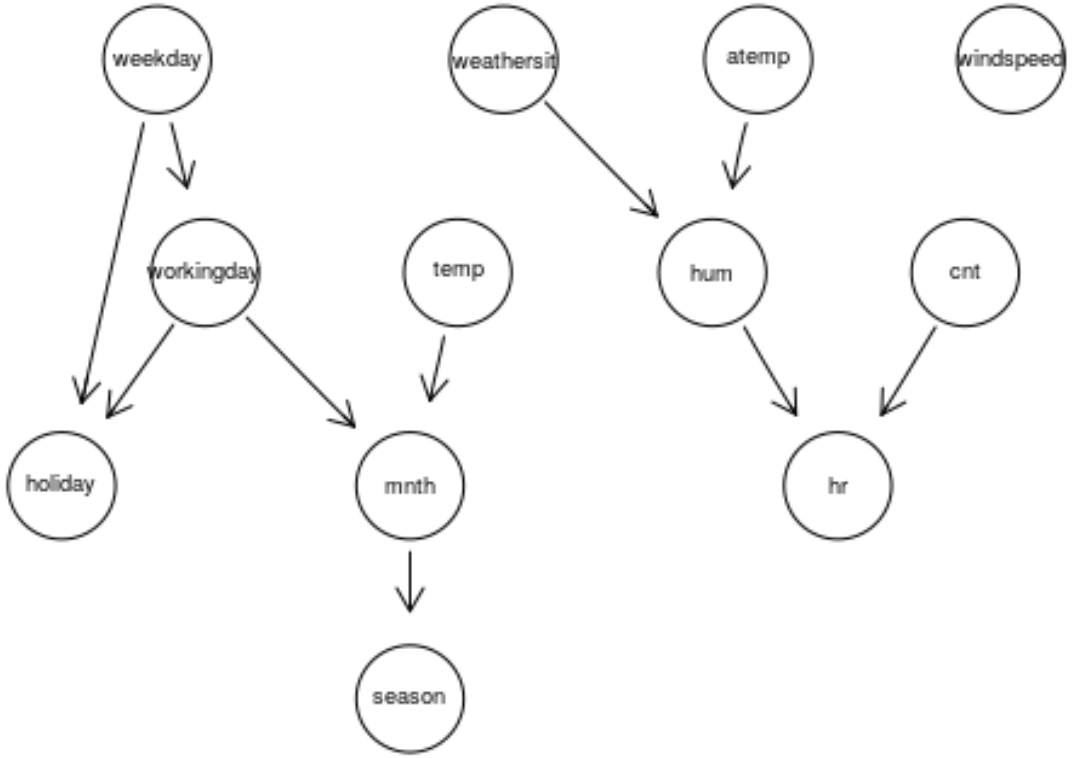
Combinando i risultati di questi algoritmi, la conoscenza personale e l'analisi della correlazione si otterranno 3 modelli finali distinti

MODELLI A CONFRONTO

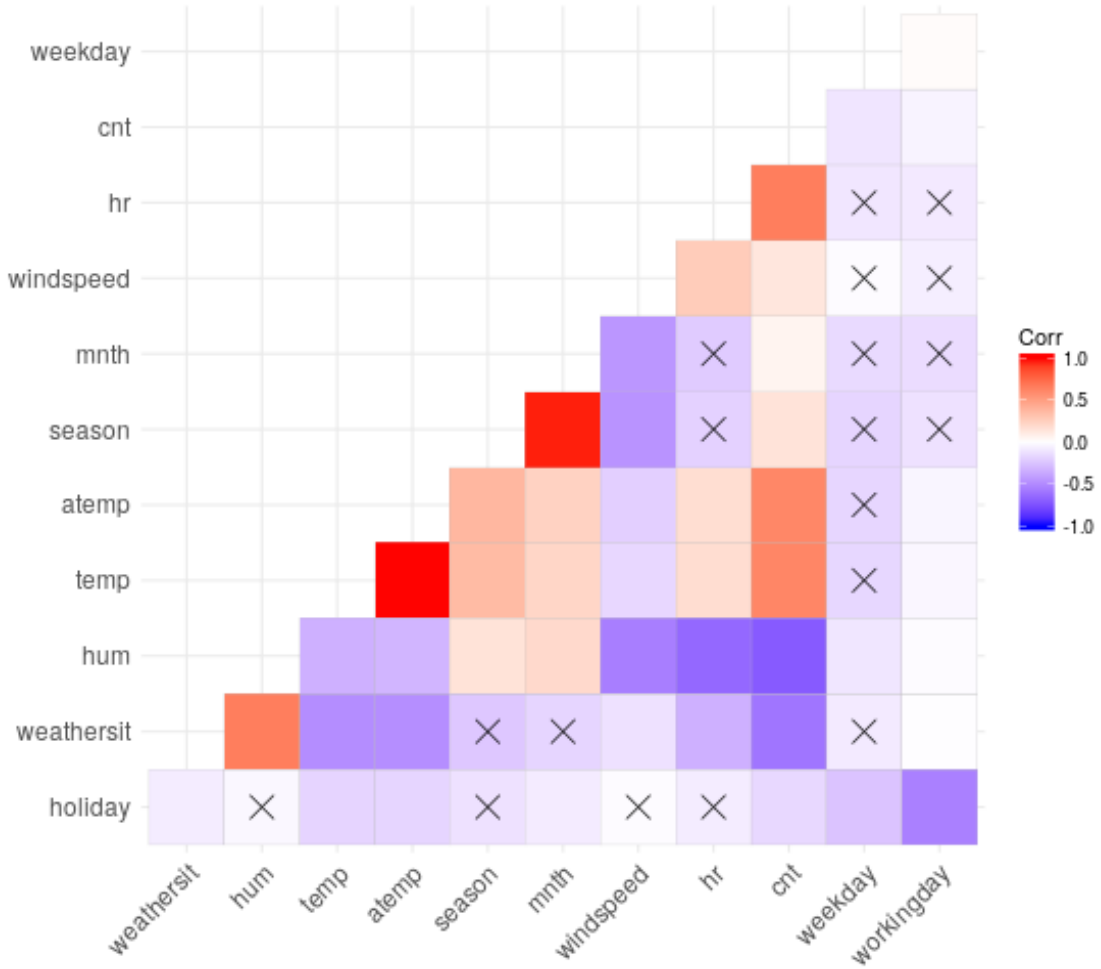
Tabu partendo da Grow-Shrink

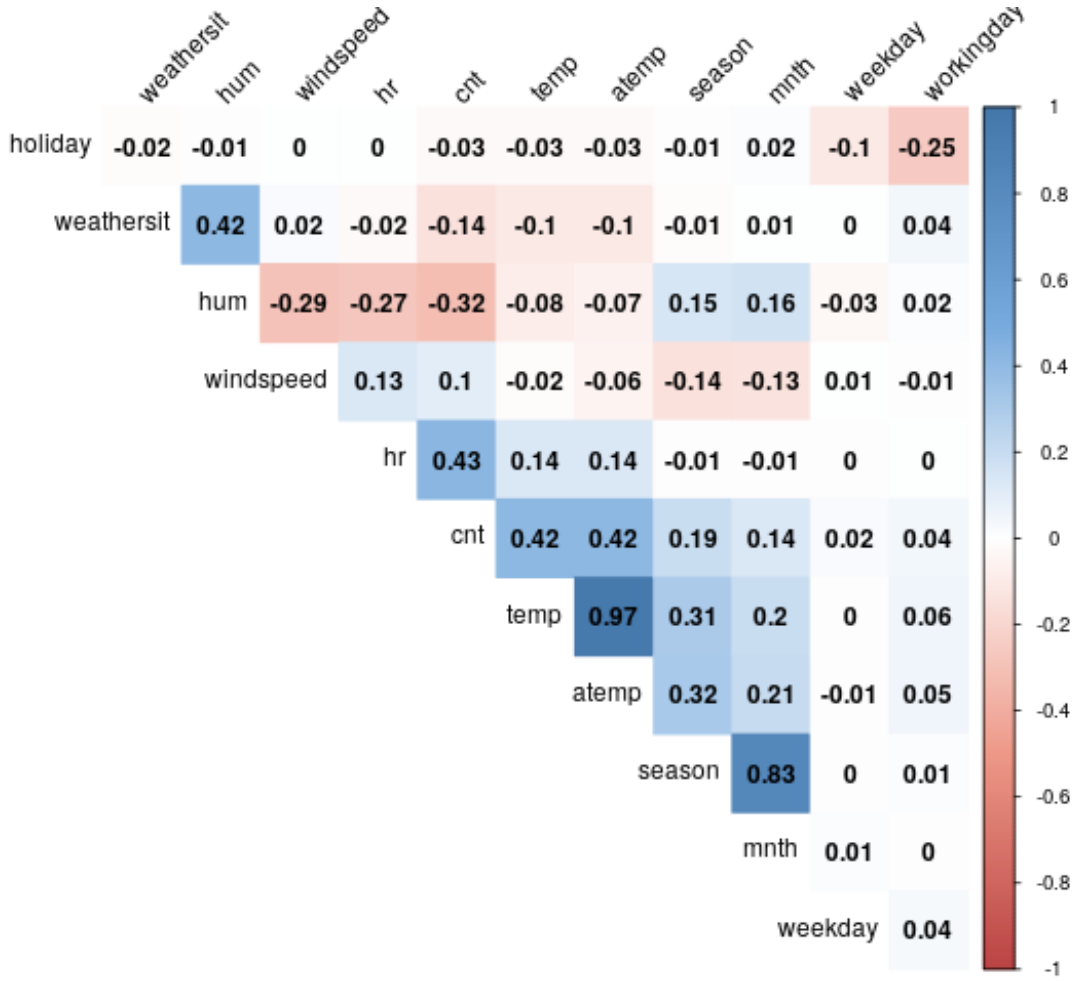


Inter-IAMB

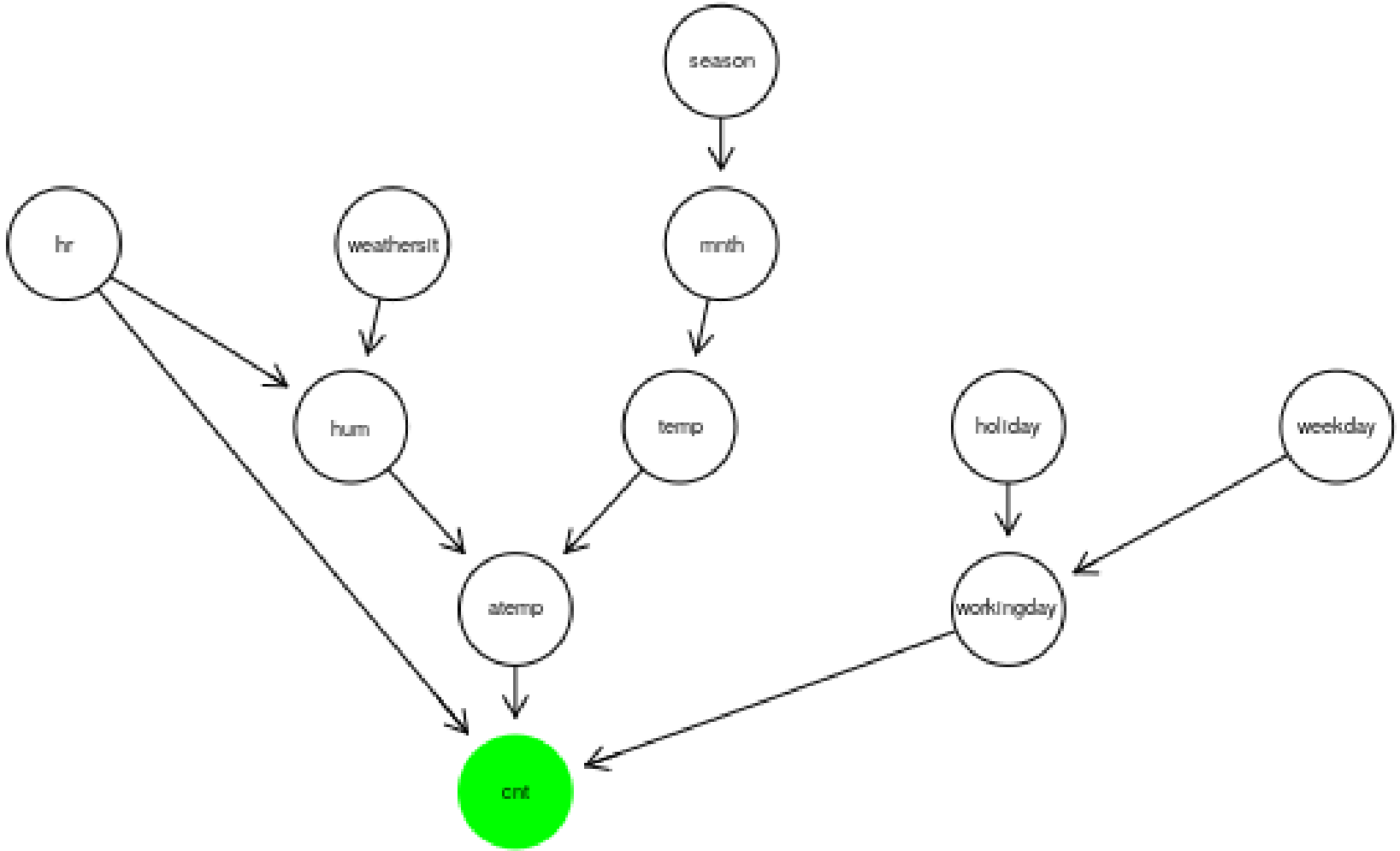


CORRELAZIONE

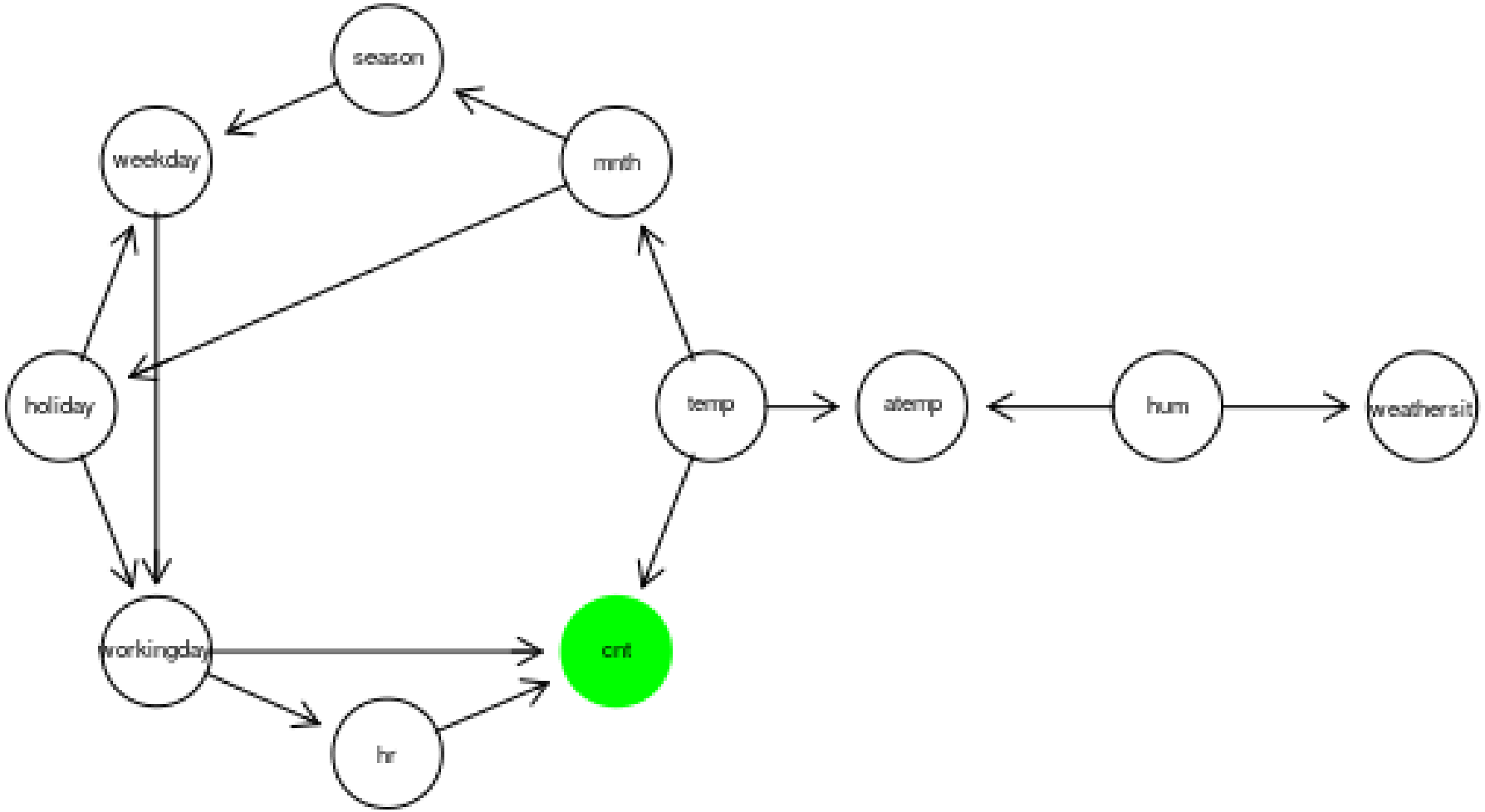




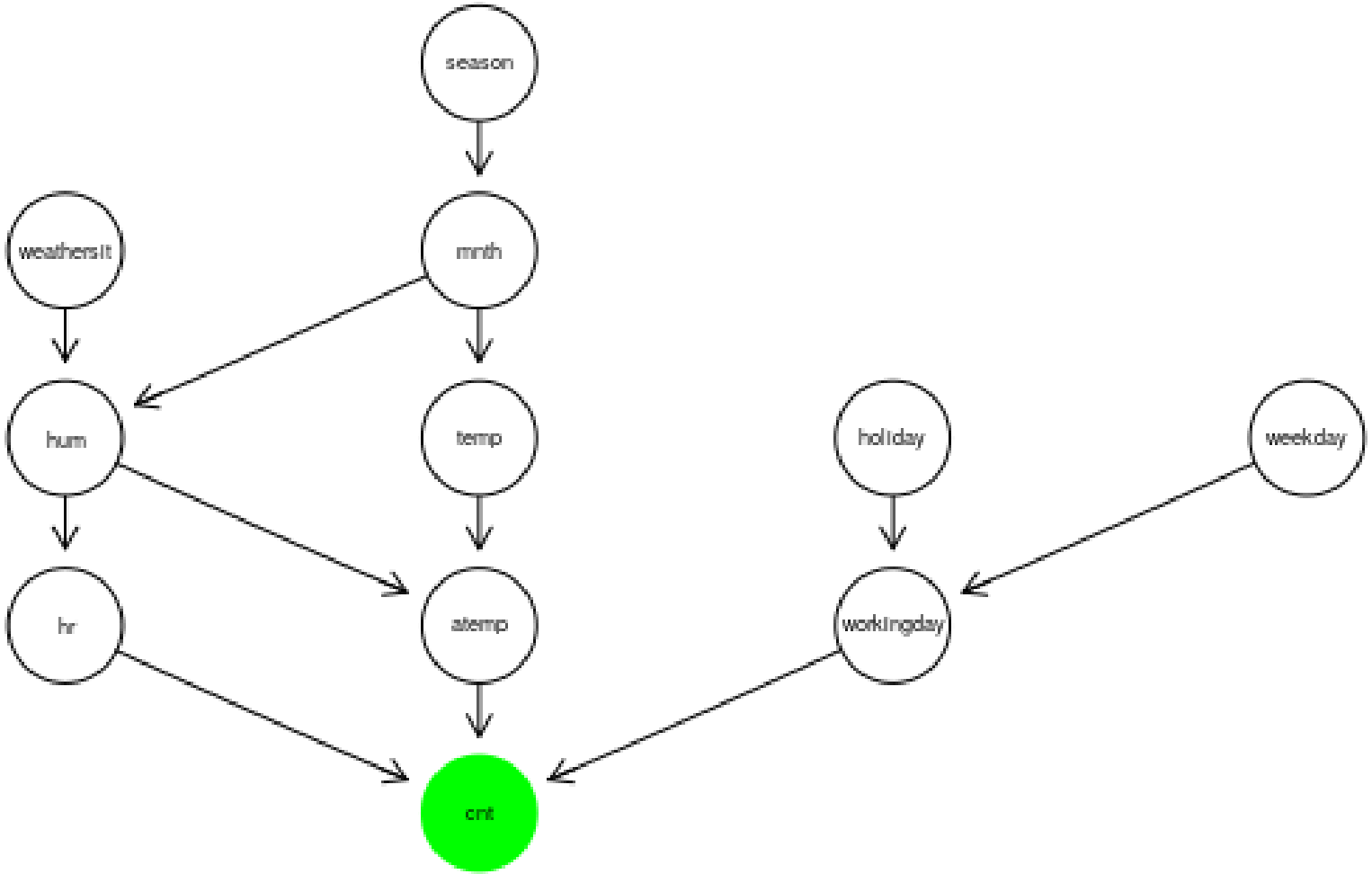
MODELLO FINALE 1



MODELLO FINALE 2



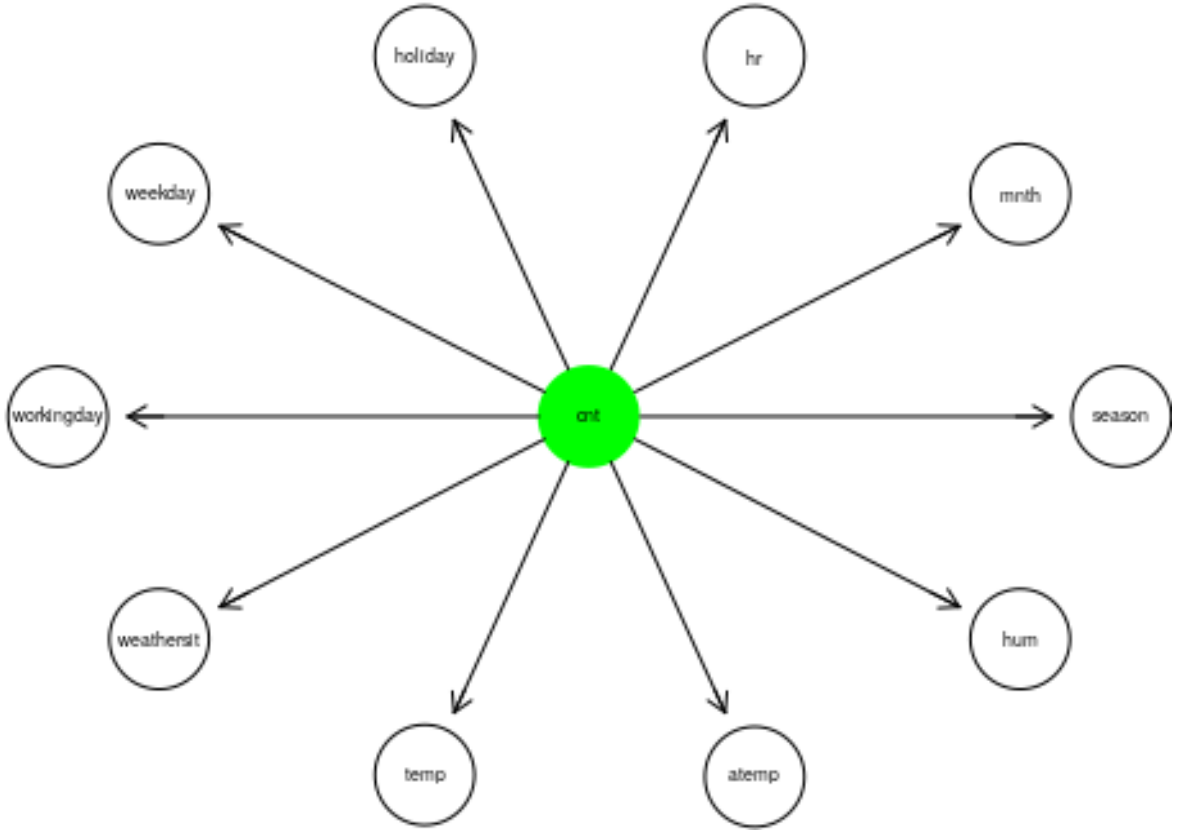
MODELLO FINALE 3



MODELLO NAIVE BAYES

Abbiamo deciso di creare anche un modello Naive-Bayes per effettuare un confronto con il nostro modello e valutarne le performance

Un modello Naive-Bayes prevede che la variabile di interesse sia causa di tutte le altre variabili, che si assumono indipendenti tra di loro



CONVALIDA DEL MODELLO

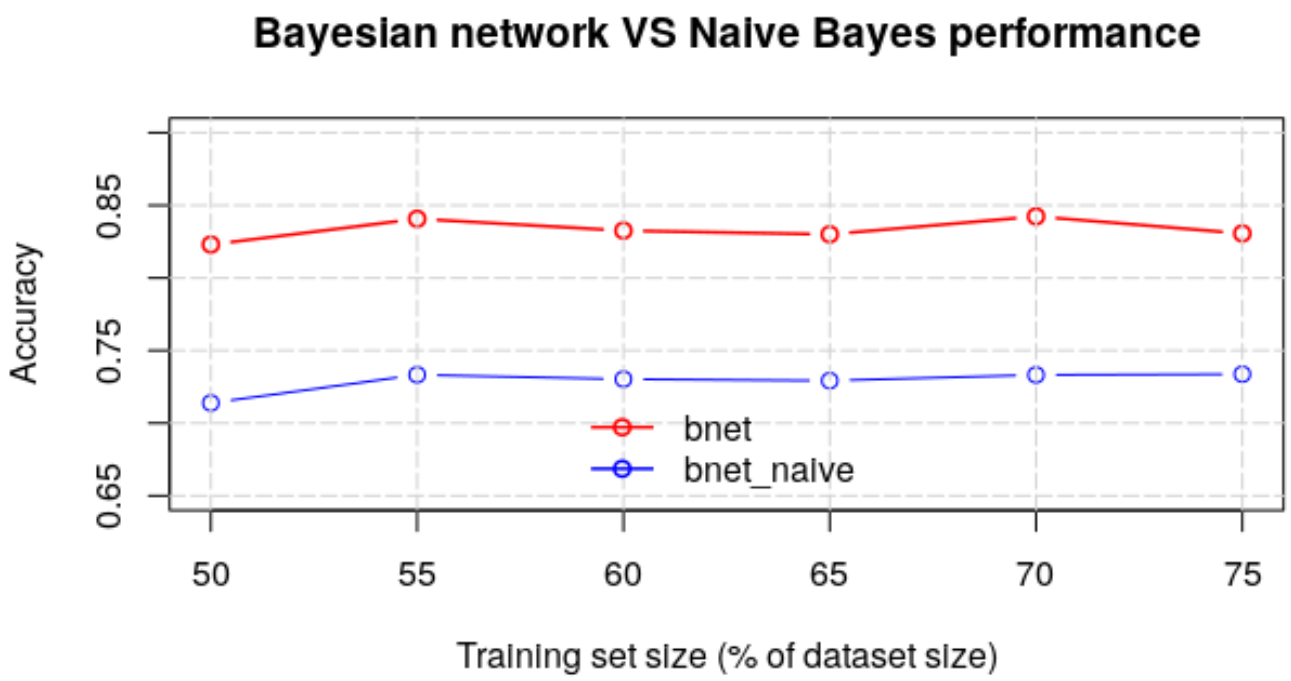
Per valutare i modelli, abbiamo:

1. Utilizzato diversi test e train set per vederne il comportamenti al crescere del training set
2. Eseguito 5-fold cross validation
3. Visualizzato le curve ROC per valutare le predizioni delle singole classi

I singoli risultati, presentati solo per il migliore dei 3 modelli (modello 2), sono presentati nelle slide a seguire.

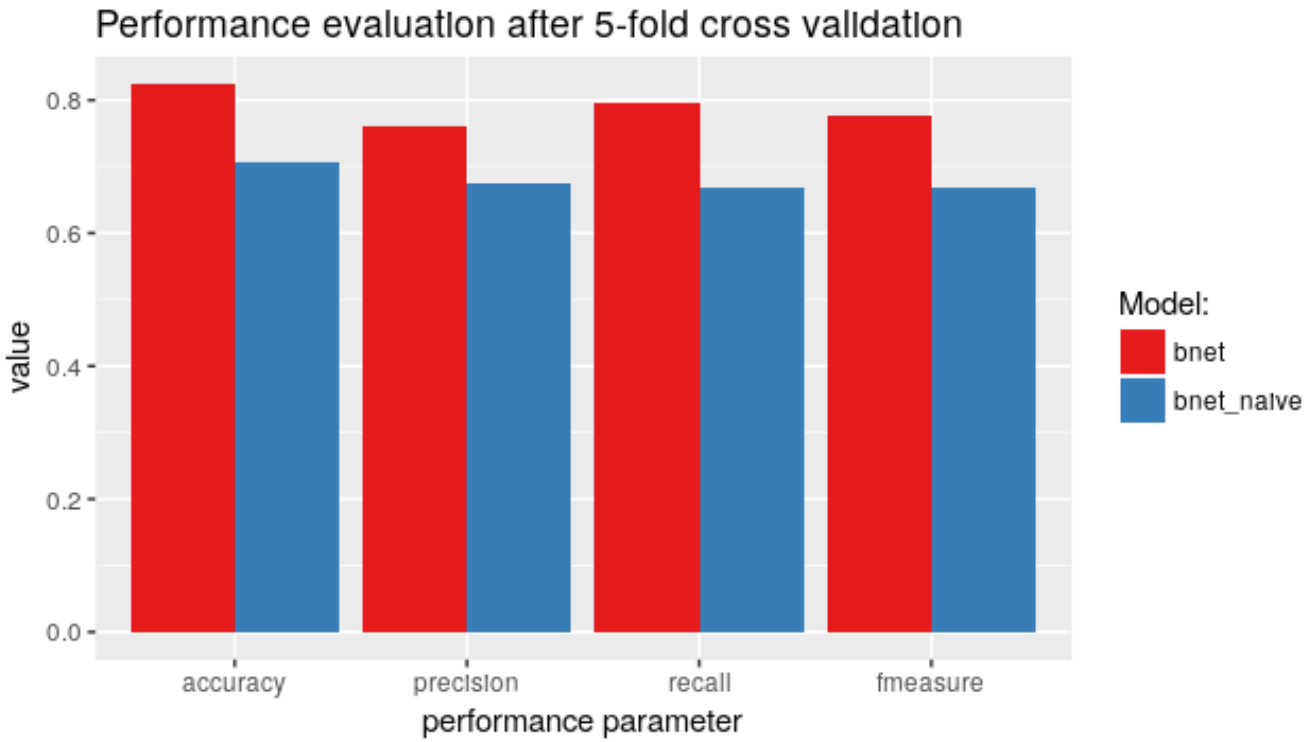
RISULTATI: INFLUENZA DEL TRAINING SET SULL'ACCURATEZZA

Train e test sets	Bayesian Network	Naive Bayes
test 1	0.82	0.71
test 2	0.84	0.73
test 3	0.83	0.73
test 4	0.83	0.73
test 5	0.84	0.73
test 6	0.83	0.73

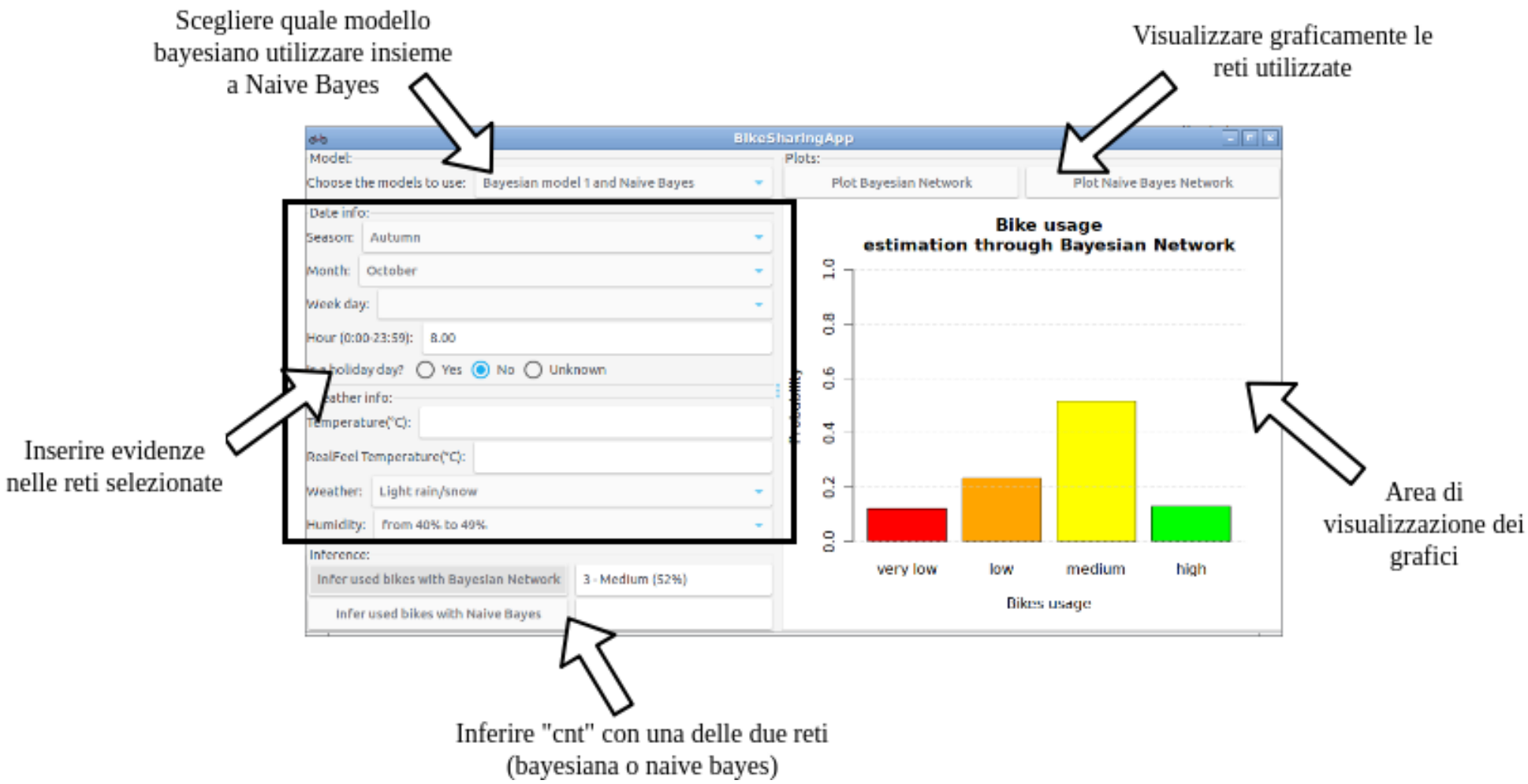


RISULTATI: MISURE DI QUALITÀ

Misura di qualità	Bayesian Network	Naive Bayes
accuracy	0.82	0.71
precision	0.76	0.67
recall	0.80	0.67
f-measure	0.78	0.67



INTERFACCIA



CONCLUSIONI

La dimensione del train set non influenza in maniera determinante le performance dei modelli

Le reti bayesiane forniscono prestazioni nei termini di accuracy, precision, recall e f-measure migliori del 10% circa rispetto al modello Naive-Bayes

Questo lavoro mostra dei risultati soddisfacenti e può essere un buon punto di partenza per ulteriori analisi che hanno come obiettivo il miglioramento dei servizi di bike sharing. In futuro si potrebbe:

1. Incrementare il dataset e rieseguire l'analisi
2. Utilizzare i dati relativi ad anni più stabili rispetto ai primi due anni del servizio