

Progetto di Sistemi Context Aware:
Smart Parking System
A.A. 2021-2022

Luca Genova 1038843
Marco Benito Tomasone 1038815
Simone Boldrini 1038792

2022-2023

Indice

1	Introduzione	2
2	Valutazione degli Algoritmi di Classificazione	4
2.1	Dataset, data mining e tecnologie utilizzate	4
2.2	Gli algoritmi	5
2.2.1	KNN	6
2.2.2	Random Forest	6
2.2.3	Gaussian Naive Bayes	7
2.3	Confronto dei risultati	8
3	Data Mining	9
3.1	Dataset	9

Capitolo 1

Introduzione

Il parcheggio è una delle sfide più comuni che i conducenti affrontano quotidianamente nelle città. Trovare un posto auto libero in una zona affollata può richiedere molto tempo e aumentare il traffico, causando stress e frustrazione per i conducenti. Per risolvere questi problemi, è stato proposto un sistema di parcheggio intelligente che monitora la disponibilità dei posti auto in una zona specifica della città di Bologna. Il sistema fornisce informazioni sul numero di posti auto liberi, aiutando i conducenti a trovare rapidamente un posto dove parcheggiare.

Il sistema si basa sul riconoscimento dell'attività dell'utente e della posizione per determinare lo stato di guida o cammino dell'utente e fornire informazioni sullo stato dei parcheggi disponibili. Questo è il cuore del progetto e consente di tenere conto degli eventi di ingresso (WALKING \rightarrow DRIVING) e uscita (DRIVING \rightarrow WALKING) dall'auto.

L'esecuzione di questo progetto è stato suddiviso in due fasi:

- La prima fase consiste nel realizzare il sistema di parcheggio intelligente che monitora la disponibilità dei posti auto in una zona specifica della città di Bologna. Esso è diviso in tre parti:
 - **Un'app:** l'applicazione mobile che permette ai conducenti di visualizzare le informazioni sullo stato dei posti auto in una zona specifica. La funzionalità dell'app è basata sul riconoscimento della posizione e dell'attività dell'utente.
 - **Un frontend:** il sito web che permette agli amministratori di visualizzare le informazioni sulle richieste di parcheggio in una zona specifica, di utilizzare l'algoritmo K-means per visualizzare gli eventi di parcheggio in cluster e anche di visualizzare la heatmap di essi.
 - **Un backend:** il server che gestisce le richieste dell'applicazione e fornisce le informazioni sullo stato dei posti auto interrogando un database PostgreSQL (con estensione PostGIS) tramite query spaziali.

- La seconda fase consiste nell'analizzare le prestazioni di tre diversi algoritmi di clustering: K-means, Random Forest e Gaussian Naive Bayes. Questi algoritmi vengono, poi, confrontati con le prestazioni del sistema integrato all'interno dell'applicazione, per determinare quale sia il più adatto per essere implementato all'interno di essa. In questo modo, è possibile ottenere una soluzione più precisa e affidabile per fornire informazioni sullo stato dei posti auto disponibili.

Il presente report, quindi, fornisce una descrizione dettagliata del funzionamento del sistema di parcheggio intelligente e analizza le sue prestazioni e la sua fattibilità. Verrà presentato il design del sistema, le tecnologie utilizzate e i risultati delle prove effettuate. Inoltre, verrà fornita una valutazione complessiva del sistema e verranno proposte possibili soluzioni per eventuali problemi riscontrati.

Capitolo 2

Valutazione degli Algoritmi di Classificazione

Nella presente fase del progetto, è stata effettuata la valutazione di tre algoritmi di classificazione: **KNN**, **Random Forest** e **Gaussian Bayes**. L'obiettivo era quello di eseguire un confronto tra i tre algoritmi e di determinare quale di essi fosse più preciso nel classificare i dati utilizzando il dataset HAR fornito.

La valutazione è stata effettuata attraverso la misurazione dell'accuratezza degli algoritmi e il confronto dei risultati ottenuti. Questo ha permesso di determinare quale algoritmo fosse il più adatto per il problema di riconoscimento delle attività svolte dall'utente.

2.1 Dataset, data mining e tecnologie utilizzate

Il dataset utilizzato per l'analisi ci è stato fornito dal professor Marco Di Felice. Il dataset è composto da 62.584 osservazioni e 13 colonne riguardanti il valore dei sensori di uno smartphone android e l'attività dell'utente, come seguente:

accelerometer#mean	la media delle osservazioni dell'accelerometro.
accelerometer#min	l'osservazione minima dell'accelerometro.
accelerometer#max	l'osservazione massima dell'accelerometro.
accelerometer#std	la deviazione standard delle osservazioni dell'accelerometro.
gyroscope#mean	la media delle osservazioni del giroscopio.
gyroscope#min	l'osservazione minima del giroscopio.
gyroscope#max	l'osservazione massima del giroscopio.
gyroscope#std	la deviazione standard delle osservazioni del giroscopio.
gyroscopeuncalibrated#mean	la media delle osservazioni del giroscopio non calibrato.
gyroscopeuncalibrated#min	l'osservazione minima del giroscopio non calibrato.
gyroscopeuncalibrated#max	l'osservazione massima del giroscopio non calibrato.
gyroscopeuncalibrated#std	la deviazione standard delle osservazioni del giroscopio non calibrato.
target	l'attività svolta dall'utente.

Per il data mining è stato utilizzato il linguaggio di programmazione *Python* e per il training e prediction dei modelli la libreria *Scikit-learn*.

Non è stato fatto un grande preprocessing dei dati, in quanto il dataset è stato fornito abbastanza pulito. I dati sono stati lavorati al fine di eliminare i valori nan e di normalizzare i dati. Inoltre sono state eliminate tutte le tuple che non riguardavano le attività di interesse (WALKING e DRIVING).

Per tutti e tre gli algoritmi il dataset è stato suddiviso in training e testing set. Per il training è stato utilizzato l'80% dei dati e per il testing il 20%.

2.2 Gli algoritmi

Come già detto precedentemente, in questa fase di valutazione degli algoritmi di classificazione, sono stati considerati tre differenti approcci: il **KNN**, il **Random Forest** e il **Gaussian Naive Bayes**.

Questi algoritmi rappresentano tre approcci distinti alla risoluzione del problema di classificazione, ciascuno con i suoi punti di forza e limiti. Il **KNN** utilizza una logica basata sulla vicinanza dei dati, il **Random Forest** è un algoritmo di ensemble che utilizza molteplici alberi di decisione e il **Gaussian Naive Bayes** si basa sull'assunzione che le feature sono distribuite normalmente.

In questa sezione, verranno descritti in dettaglio ciascuno di questi algoritmi e verrà presentato un confronto tra i loro risultati e le loro prestazioni.

2.2.1 KNN

Per quanto riguarda il KNN è stato deciso di effettuare un tuning degli iperparametri per il KNN, con lo scopo di ottenere una prestazione ottimale del modello.

I parametri che sono stati modificati sono stati:

- **n_neighbors**: che rappresenta il numero di vicini considerati nella classificazione
- **weights**: che determina il peso dei vicini nella classificazione
- **leaf-size**: che rappresenta la dimensione massima di una foglia dell'albero di ricerca
- **p**: che determina la potenza della metrica utilizzata nei calcoli della distanza tra i vicini

Il tuning degli iperparametri è stato effettuato utilizzando la funzione *GridSearchCV* di *Scikit-learn*. Questa funzione permette di eseguire una ricerca su una griglia di valori per gli iperparametri, testando tutte le possibili combinazioni e selezionando quella che produce la migliore prestazione in termini di accuratezza.

L'obiettivo di questa ottimizzazione era quello di migliorare la precisione del modello KNN e di fornire una soluzione ottimale per la classificazione dei dati. Il risultato ottenuto ha dimostrato che l'ottimizzazione degli iperparametri è stata efficace e ha contribuito a migliorare la performance del modello.

Risultati

I risultati del tuning degli iperparametri utilizzando il dataset precedentemente descritto sono riportati nella tabella seguente:

	Best value
n_neighbors	5
weights	uniform
leaf_size	30
p	2

Con questi parametri siamo riusciti ad ottenere un'accuratezza finale del 90.5%.

2.2.2 Random Forest

Come per il KNN, anche per il Random Forest è stato effettuato un tuning degli iperparametri, sempre con *GridSearchCV*, per ottenere una prestazione ottimale del modello.

I parametri che sono stati modificati sono stati:

- **n_estimators**: che rappresenta il numero di alberi decisionali nella foresta
- **max_depth**: che rappresenta la profondità massima degli alberi decisionali
- **min_samples_split**: che rappresenta il numero minimo di campioni richiesti per dividere un nodo interno
- **min_samples_leaf**: che rappresenta il numero minimo di campioni richiesti per costruire una foglia dell'albero

Il tuning degli iperparametri è stato effettuato utilizzando la funzione *GridSearchCV* di *Scikit-learn*.

L'obiettivo di questa ottimizzazione era quello di migliorare la precisione del modello Random Forest e di fornire una soluzione ottimale per la classificazione dei dati. Il risultato ottenuto ha dimostrato che l'ottimizzazione degli iperparametri è stata efficace e ha contribuito a migliorare la performance del modello.

Risultati

I risultati del tuning degli iperparametri utilizzando il dataset precedentemente descritto sono riportati nella tabella seguente:

	Best value
n_estimators	100
max_depth	5
min_samples_split	2
min_samples_leaf	1

Con questi parametri siamo riusciti ad ottenere un'accuratezza finale del 91.2%.

2.2.3 Gaussian Naive Bayes

Per quanto riguarda il Gaussian Naive Bayes, abbiamo deciso di utilizzare questo algoritmo per la classificazione dei dati in quanto si tratta di un metodo semplice e veloce per la classificazione basato sull'assunzione che le feature siano distribuite secondo una distribuzione normale.

In questo caso, non era necessario effettuare un tuning degli iperparametri, in quanto il Gaussian Naive Bayes non ha parametri che possono essere ottimizzati per migliorare le sue prestazioni.

Risultati

il modello Gaussian Naive Bayes ha fornito un'accuratezza del 81.82%, che risulta essere un risultato inferiore rispetto ai modelli KNN e Random Forest.

2.3 Confronto dei risultati

Illustriamo di seguito le confusion matrix dei modelli KNN, Random Forest e Gaussian Naive Bayes, ottenute utilizzando il dataset precedentemente descritto.

Capitolo 3

Data Mining

3.1 Dataset

In nostro possesso è stato dato un dataset in formato *.csv* contenente 62.584 osservazioni e 13 colonne riguardanti il valore dei sensori di uno smartphone android e l'attività dell'utente.

Le etichette di questo dataset assumono 5 valori differenti:

- **STILL**: l'utente non si muove.
- **WALKING**: l'utente cammina.
- **CAR**: l'utente è in auto.
- **BUS**: l'utente è in bus.
- **TRAIN**: l'utente è in treno.

Le feature del dataset sono 12 e sono le seguenti:

- **accelerometer#mean**: la media delle osservazioni dell'accelerometro.
- **accelerometer#min**: l'osservazione minima dell'accelerometro.
- **accelerometer#max**: l'osservazione massima dell'accelerometro.
- **accelerometer#std**: la deviazione standard delle osservazioni dell'accelerometro.
- **gyroscope#mean**: la media delle osservazioni del giroscopio.
- **gyroscope#min**: l'osservazione minima del giroscopio.
- **gyroscope#max**: l'osservazione massima del giroscopio.

- **gyroscope#std**: la deviazione standard delle osservazioni del giroscopio.
- **gyroscopeuncalibrated#mean**: la media delle osservazioni del giroscopio non calibrato.
- **gyroscopeuncalibrated#min**: l'osservazione minima del giroscopio non calibrato.
- **gyroscopeuncalibrated#max**: l'osservazione massima del giroscopio non calibrato.
- **gyroscopeuncalibrated#std**: la deviazione standard delle osservazioni del giroscopio non calibrato.