POLITECNICO DI TORINO

Corso di Laurea Magistrale

Ingegneria della Produzione Industriale e dell'Innovazione Tecnologica

Tesi di Laurea Magistrale

Modello predittivo per la stima del flusso dei visitatori in un punto vendita finalizzato ad applicazioni in ambito strategico

Relatore

Prof. Maurizio Schenone

Candidata

Francesca Margherita Chifari

ANNO ACCADEMICO 2017/2018

Indice

Indice delle figure e delle tabelle **4**

Ringraziamenti **7**

1 INTRODUZIONE **9**

2 VIDEO ANALISI **11**

2.1 Introduzione alla video analisi 11

2.2 Funzionalità della Video Analisi 12

2.3 Individuazione e tracciamento di oggetti in movimento 14

2.3.1 Segmentazione del movimento 14

2.3.2 Rimozione del rumore di sottofondo 15

2.3.3 Tracciamento degli oggetti 16

3 ANALISI DI COMPORTAMENTO BASATA SULLA VIDEO ANALISI **19**

3.1 Introduzione all'Analisi di Comportamento 19

3.2 Funzionamento dell'Analisi di Comportamento 22

3.2.1 Algoritmi di video analisi 22

3.2.2 Modello statistico e di inferenza 25

3.2.3 Extract, Transform and Load 27

3.3 Applicazione del modello a un caso reale 32

3.4 Limiti del modello 34

4 STRUMENTO DI ANALISI E CORREZIONE DELL'ERRORE **37**

4.1 Presupposti iniziali 37

4.2 Raccolta dei dati 38

4.3 Interpretazione dei dati 40

4.4 Analisi dell'andamento dell'errore 43

4.5 Distribuzione dei dati 44

4.6 Modello di correzione 50

4.7 Considerazioni sullo strumento realizzato 54

5 MODELLO PREDITTIVO 55
5.1 Machine Learning 55
5.1.1 Introduzione al Machine Learning 55
5.1.2 Azure Machine Learning Studio 57
5.2 Scenario 60
5.3 Raccolta dati 60
5.3.1 Inout 60
5.3.2 Negozio 62
5.3.3 Meteo 64
5.4 Struttura del modello 65
5.4.1 Conversione files 66
5.4.2 Preparazione dei dati 67
5.4.3 Combinazione dei dati 71
5.4.4 Suddivisione dei dati per il training e la predizione 71
5.4.5 Algoritmo di previsione 72
5.5 Lettura e interpretazione dei dati 74
6 CONCLUSIONE 79
6.1 Comparazione tra il modello di correzione attuale e quello elaborato 79
6.1.1 Correzione giornaliera con confronto per fascia oraria 79
6.1.2 Correzione mensile con confronto per giorno 81
6.2 Prerequisiti del modello 83
6.3 Applicazioni del modello 83
6.3.1 Ambito strategico 83
6.3.2 Ambito gestionale 84
7 BIBLIOGRAFIA 86
7.1 Video Analisi 86
7.2 Machine Learning 86
8 APPENDICE 88
8.1 Tabella creata dal modulo Score Model di Azure Machine Learning Studio ... 88

8.2 Tabella utilizzata per calcolare il grado di precisione dei risultati di predizione 89
8.3 Tabella che include il modello di correzione attuale applicato ad ogni fascia oraria del 2 maggio 2016 90
8.4 Tabella che include il modello di correzione realizzato applicato ad ogni fascia oraria del 2 maggio 2016 91
8.5 Tabella che include il modello di correzione attuale applicato mensilmente . 92
8.6 Tabella che include il modello di correzione realizzato applicato mensilmente 93

Indice delle figure e delle tabelle

Figura 2.1 – Immagine estrapolata dalla telecamera installata presso un salone di bellezza che rileva il passaggio di persone davanti ad una vetrina e ne traccia la traiettoria 12
Figura 2.2 – Esempio di classificazione in base ai tratti somatici 13
Figura 2.3 – Diagramma a blocchi che rappresenta il sistema 14
Figura 2.4 – Esempio di identificazione degli oggetti 15
Figura 2.5 – Tracciamento del movimento degli oggetti in un fotogramma 17
Figura 3.1 – Key Performance Indicator a seconda del settore 21
Figura 3.2 – Schema sul funzionamento dell'Analisi di Comportamento 22
Figura 3.3 – Numero delle persone entrate dal varco 28
Figura 3.4 – Numero delle persone uscite dal varco 29
Tabella 3.5 – Tabella che mostra come viene applicato il fattore di correzione 29
Figura 3.6 – Numero delle persone ancora presenti nel punto vendita considerando gli ingressi e le uscite amplificate dal fattore correttivo .. 30
Figura 3.7 – Stima dei visitatori su base oraria 30
Figura 3.8 – Esempio di schermata della dashboard che rappresenta il numero e la direzione di passaggi in una giornata davanti ad una vetrina 31
Figura 3.9 – Esempio di schermata della dashboard che rappresenta il numero di visitatori in un'area monitorata in un giorno 31
Figura 3.10 – Esempio di schermata della dashboard che rappresenta il tempo di permanenza minimo, massimo e medio dei visitatori 32
Tabella 3.11 – Tabella che riporta il conteggio degli ingressi e delle uscite per il 1 maggio 2016 33
Tabella 3.12 – Numero di ingressi, di uscite e persone presenti nell'area monitorata 35
Figura 4.1 – Schermata principale di pgAdmin III 38
Tabella 4.2 – Tabella Areas 39
Tabella 4.3 – Estratto dalla Tabella InOut 40

Tabella 4.4 – Estratto dalla Tabella InOut che contiene solo l'area 27 40
Figura 4.5 – Grafico che delinea l'andamento dell'errore..... 43
Tabella 4.6 – Tabella per costruire l'istogramma con 20 classi 45
Figura 4.7 – Istogramma della distribuzione dei dati considerando 20 classi 46
Tabella 4.8 – Tabella per costruire l'istogramma con 40 classi 47
Figura 4.9 – Istogramma della distribuzione dei dati considerando 40 classi 48
Figura 4.10 – Istogramma che descrive un errore nel conteggio delle entrate 48
Figura 4.11 – Istogramma che descrive un errore nel conteggio delle uscite..... 49
Figura 4.12 – Istogramma che mostra come l'errore tra gli ingressi e le uscite si compensa 49
Figura 4.13 – Istogramma che si avvicina al caso ideale 50
Figura 4.14 – Istogramma descrive una distribuzione casuale..... 50
Tabella 4.15 – Estratto dalla tabella Maggio16 per il giorno 1 maggio 2016..... 51
Tabella 4.16 – Tabella con i dati del mese di maggio 2016 52
Tabella 4.17 – Tabella che illustra il risultato ottenuto dopo aver applicato la correzione..... 53
Figura 5.1 – Interfaccia grafica di Azure Machine Learning Studio 58
Tabella 5.2 – Tabella InOut prima di essere filtrata..... 61
Tabella 5.3 – Estratto dalla tabella che indica la combinazione di video proiettato in vetrina, addetto vendite e spot televisivo 63
Tabella 5.4 – Estratto sul meteo di Roma..... 64
Figura 5.5 – Modello creato su Azure Machine Learning Studio..... 66
Figura 5.6 – Impostazioni del modulo select columns in dataset per la tabella inout_previsione 67
Figura 5.7 – Impostazioni del modulo Group Categorical Values: trasformazione del lunedì e del martedì in feriale, del sabato e della domenica in festivo.. 69
Figura 5.8 – Impostazioni del modulo Edit Metadata per cambiare il nome alla colonna mese in stagione..... 69
Figura 5.9 – Impostazioni del modulo Clean Missing Data per eliminare le righe con elementi mancanti della tabella meteo_roma 70

Figura 5.10 – Impostazioni del modulo Decision Forest Regression 73
Figura 5.11 – Esempio di albero creato dal regressore di foresta decisionale..... 73
Figura 5.12 – Estratto dalla tabella creata automaticamente dal modulo Score Model che mostra i risultati di predizione 74
Tabella 5.13 – Estratto dalla tabella utilizzata per calcolare il grado di precisione dei risultati di predizione 75
Tabella 5.14 – Numero di previsioni per categoria 76
Figura 5.15 – Grafico che rappresenta la suddivisione dei risultati in base alla categoria corrispondente..... 76
Figura 5.16 – Andamento del grado di precisione nel mese di febbraio 77
Tabella 5.17 – Estratto della tabella che illustra i risultati della predizione con un grado di precisione basso..... 78
Tabella 5.18 – Estratto dalla tabella creata automaticamente dal modulo score model che mostra i risultati di predizione 78
Tabella 5.19 – estratto dalla tabella utilizzata per calcolare il grado di precisione dei risultati di predizione 78
Tabella 6.1 – Estratto della tabella che include il modello di correzione attuale applicato alle fasce orarie (2 maggio 2016) 80
Tabella 6.2 – Estratto della tabella che include il modello di correzione realizzato applicato alle fasce orarie (2 maggio 2016) 81
Tabella 6.3 – Estratto della tabella che include il modello di correzione attuale applicato ai giorni del mese di febbraio..... 82
Tabella 6.4 – Estratto della tabella che include il modello di correzione realizzato applicato ai giorni del mese di febbraio..... 82

Ringraziamenti

La mia carriera universitaria è giunta al termine ed è importante per me ringraziare tutte le persone che mi hanno motivato e accompagnato durante questi cinque anni e che hanno reso possibile il raggiungimento di questo grande traguardo.

Vorrei ringraziare il Professor Schenone, che per la seconda volta è stato il mio relatore seguendomi passo dopo passo durante la stesura di questo elaborato e che si è sempre dimostrato pronto ad aiutarmi, dandomi consigli preziosi, rispondendo sempre ai miei messaggi, colmando le mie perplessità e leggendo, correggendo e ricorreggendo la mia tesi.

Un ringraziamento va anche al Professor Zucca, mio tutor accademico, per essere sempre stato disponibile a incontrarmi ogni qual volta avessi bisogno del suo intervento, ascoltarmi e aver supportato noi tutti durante il nostro percorso di studi.

Ci tengo molto a ringraziare Gaetano Torrisi, Partner di Discovery Reply, che mi ha offerto la grande opportunità di scrivere la tesi nella sua azienda e di prendere parte ad uno dei suoi progetti più interessanti ed innovativi.

Ringrazio Andrea Ceciarelli, il mio manager, per avermi guidato durante il mio tirocinio.

Ringrazio di cuore Alessandro Casu, per avermi preso sotto la sua ala protettrice fin dall'inizio della mia esperienza lavorativa, aver indirizzato il mio lavoro e avermi sempre motivato.

Grazie a tutti i miei colleghi che hanno reso la mia permanenza in questa azienda costruttiva e divertente.

Un ringraziamento speciale a mia sorella e ai miei genitori, i quali mi hanno sempre incoraggiato e guidato nelle mie scelte, consigliandomi e condividendo la loro esperienza affinchè ne potessi trarre il meglio. Un pensiero va anche a mia nonna Concetta, che non c'è più, ma che sarebbe molto orgogliosa del mio percorso e di vedermi laureare oggi e a mio nonno Piero.

Grazie a tutti gli amici e compagni di corso del Politecnico di Torino e dell'AIT, che hanno studiato, viaggiato e lavorato con me in questi cinque anni.

Grazie a Vito Alberto di Modica e Roberto Laganaro, che mi hanno accolta come coinquilina durante la mia permanenza a Milano facendomi sentire a casa.

Infine, un grazie dal cuore a Roberto, compagno di vita e di sogni, l'unica ragione per cui avrei voluto che il mio ultimo semestre durasse un po' di più. Grazie per avermi sempre spinta a dare il massimo e puntare sempre più in alto. Grazie per essermi stato sempre vicino.

1 INTRODUZIONE

Il presente elaborato è frutto di un'approfondita ricerca sull'elaborazione di un modello predittivo in grado di stimare il numero di visitatori futuri in un punto vendita partendo dall'analisi dello storico dei dati estrapolati da un sistema di video analisi.

La motivazione che mi ha spinto ad approfondire questa tematica nasce dal desiderio di combinare un forte interesse per lo studio sul comportamento di acquisto dei consumatori con la volontà di imparare a utilizzare uno dei nuovi trend tecnologici di maggiore rilievo, il *Machine Learning*.

La tesi è articolata in cinque capitoli.

Nel primo capitolo viene fornita una spiegazione teorica della video analisi, descrivendone le applicazioni, le funzionalità e l'approfondimento di una di esse al fine di introdurre il lettore nell'ambito dell'argomento.

Il secondo capitolo si concentra sull'Analisi del Comportamento basata sulla Video Analisi, ovvero il sistema realizzato da Discovery Reply Srl che combina la video analisi con i sistemi di video sorveglianza. Viene illustrato il modello attuale, il suo funzionamento con particolare attenzione al flusso dei dati e all'algoritmo impiegato, il suo utilizzo in un caso reale ed i limiti rilevati a seguito della sua applicazione.

Nel terzo capitolo viene descritta la soluzione concepita al fine di migliorare il meccanismo di stima e rendere i dati estrapolati più verosimili e attendibili. In primo luogo vengono valutate le condizioni di partenza. Quindi, vengono descritte le modalità di raccolta e l'interpretazione dei dati. Infine vengono introdotte le metodologie utilizzate per individuare la distribuzione dei dati, verificare l'efficacia del modello e correggere il delta dell'errore.

Nel quarto capitolo vengono introdotti il concetto di *Machine Learning* e la piattaforma impiegata per la realizzazione del modello predittivo. Viene spiegato come la soluzione realizzata sia stata applicata ad un caso pratico, descrivendo le tre tabelle contenenti i

dati analizzati, la progettazione del modello creato sulla piattaforma di *Machine Learning* e l'algoritmo utilizzato. Infine vengono letti ed interpretati i risultati ottenuti.

Nell'ultimo capitolo vengono esposti in modo dettagliato i risultati conclusivi di questo studio, riportando un confronto tra il modello di correzione attuale e quello realizzato, un riepilogo del modello predittivo, considerandone i limiti e possibili applicazioni future in ambito strategico.

Seguono la bibliografia e l'appendice contenente le tabelle in versione integrale dei risultati ottenuti in questo studio.

2 VIDEO ANALISI

In questo capitolo viene introdotta la tecnologia alla base di questo studio, la video analisi.

Dopo una breve presentazione sul funzionamento di questo sistema e le sue applicazioni, vengono descritte le principali funzionalità della video analisi e viene approfondito il processo di tracciamento automatico per identificare gli oggetti in movimento.

2.1 Introduzione alla video analisi

La video analisi consiste nella capacità di analizzare fotogrammi automaticamente per determinare e descrivere eventi temporali e spaziali grazie ad algoritmi di riconoscimento dell'immagine e sistemi di rilevazione e apprendimento. Per funzionare necessita dell'integrazione tra la tecnologia video e i software di analisi.

La parte hardware è costituita da una o più telecamere IP, ovvero videocamere digitali in rete che trasmettono i dati tramite un collegamento Fast Ethernet, e le apparecchiature di rete necessarie per l'acquisizione e la trasmissione dei dati.

l software di analisi invece vengono eseguiti sul sistema per interpretare i video acquisiti e aiutare a prendere decisioni per l'esecuzione di attività automatizzate.

I sistemi di video analisi possono essere definiti come soluzioni video che sfruttano la tecnologia senza l'intervento umano in grado di processare, manipolare e compiere azioni basate su informazioni archiviate delle immagini video o in diretta.

Può essere applicata in diversi contesti, tra cui: il riconoscimento, il tracciamento e il conteggio di una persona o un oggetto in movimento, l'identificazione della posizione, il rilevamento di determinati eventi come l'intromissione in un'area non consentita e l'avviamento automatico di attività come l'attivazione di un allarme, la chiusura di porte di uscita, etc... .

2.2 Funzionalità della Video Analisi

Qui di seguito vengono riportare le principali funzioni applicative di questa tecnologia.

- Rilevamento di movimento: individua la presenza di movimento per ogni fotogramma della scena osservata.
- Tracciamento di movimento: utilizzato per individuare la traiettoria compiuta da persone o oggetti che si spostano in riferimento ad una scena che fa da sfondo che rimane invariata nel tempo (figura 2.1).

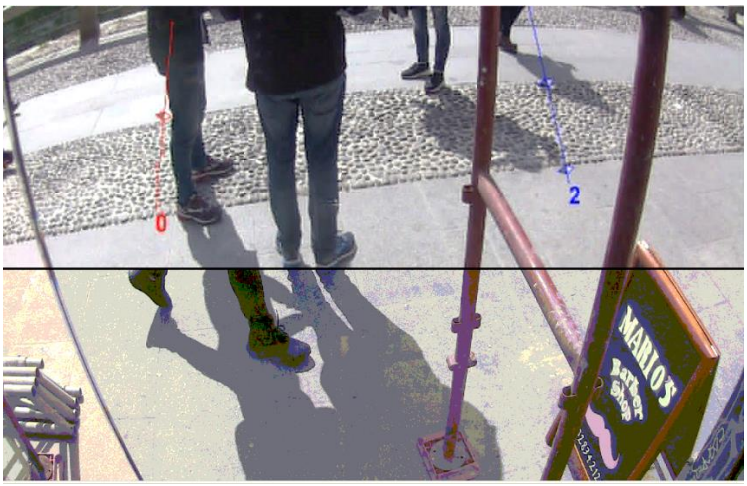

Figura 2.1 – Immagine estrapolata dalla telecamera installata presso un salone di bellezza che rileva il passaggio di persone davanti ad una vetrina e ne traccia la traiettoria

- Riconoscimento della forma: riconosce le forme nel video di input come, ad esempio, cerchi o quadrati. Viene generalmente integrato a funzionalità più avanzate per il rilevamento di oggetti.
- Riconoscimento degli oggetti: determina la presenza di un oggetto o entità, come un simbolo o un logo.
- Riconoscimento facciale: utilizzato per identificare volti. Questo sistema viene spesso integrato con algoritmi finalizzati alla classificazione delle persone (figura 2.2).

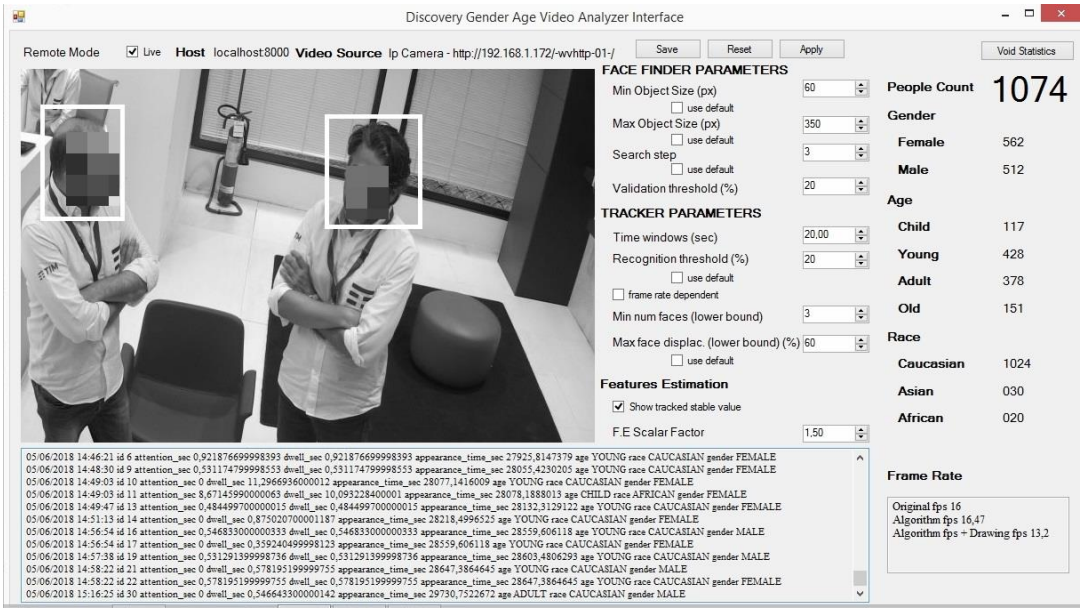

Figura 2.2 – Esempio di classificazione in base ai tratti somatici

- Egomotion: utilizzato per determinare la posizione di una telecamera in movimento relativamente ad una scena fissa analizzando il segnale di uscita.
- Mascheramento dinamico dei dati: consiste nel bloccare una parte del segnale video in base al segnale stesso, ad esempio nel caso in cui si verifichino problemi di privacy.
- Rilevazione di fiamme e fumo: funzione che permette di rilevare fiamme e fumo in meno di 15-20 secondi per mezzo di un chip DPS integrato. Il chip è in grado di elaborare gli algoritmi che analizzano i fotogrammi e che individuano un incendio basandosi sulle tonalità cromatiche dell'immagine, sulla rilevazione di tremolii, sulla forma e la direzione.
- Rilevamento di manomissione: questa funzione determina se la telecamera, il segnale di uscita o il contenuto registrato sono stati manomessi.

Le funzionalità della video analisi utilizzate per questo studio sono il rilevamento e tracciamento di oggetti in movimento.

2.3 Individuazione e tracciamento di oggetti in movimento

Il processo di tracciamento automatico degli oggetti inizia con l'identificazione di oggetti in movimento.

Una volta individuati i pixels in movimento, è necessario raggrupparli in regioni ed eliminare eventuali rumori di sottofondo.

Quindi si ricerca la corrispondenza tra gli oggetti individuati tra un fotogramma e quello successivo (figura 2.3).

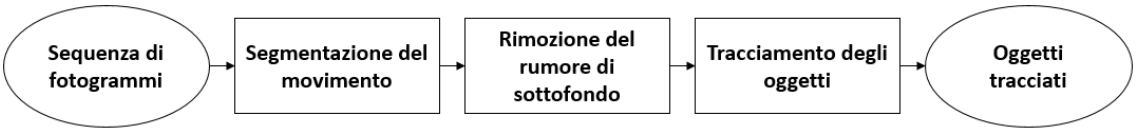

Figura 2.3 – Diagramma a blocchi che rappresenta il sistema

2.3.1 Segmentazione del movimento

Il primo passo nel processo di tracciamento automatico degli oggetti in movimento è l'identificazione gli elementi di interesse per la sequenza video e il raggruppamento per ogni fotogramma i pixels di questi oggetti.

La sottrazione dello sfondo è un metodo ampiamente utilizzato per il rilevamento di oggetti in movimento e consiste nel distinguere il *background*, ovvero quella parte della scena che si può considerare fissa o tende a cambiare lentamente nel tempo, dal *foreground* che contiene le porzioni di immagine in movimento.

Nel caso in cui sia possibile individuare l'immagine dello sfondo, il compito di rilevare gli oggetti è semplice e può essere attuato sottraendo il fotogramma corrente direttamente dall'immagine dello sfondo.

Tuttavia, l'immagine di sfondo non è sempre nota e spesso deve essere generata dal sistema automaticamente e in modo continuativo. In questo caso il sistema riconosce, partendo da un'insieme di fotogrammi, l'insieme di pixels che non variano nel corso del tempo e li memorizza come sfondo per decine di secondi. Il *background*

viene quindi sottratto dall'immagine di input al fine di estrarre gli elementi estranei (figura 2.4).

La sottrazione dello sfondo è un metodo che si fonda sulla mistura delle gaussiane, una tecnica che considera il *background* come una matrice WxH, dove W rappresenta la lunghezza dell'immagine, mentre H la sua altezza. Ogni cella della matrice contiene fino a 7 potenziali valori, ognuno dei quali identifica un colore. In questo modo il pixel può continuare ad essere considerato parte dello sfondo, ma può assumere più di un valore in base alla variazione di determinati parametri quali la luce, i riflessi, le condizioni atmosferiche, etc...

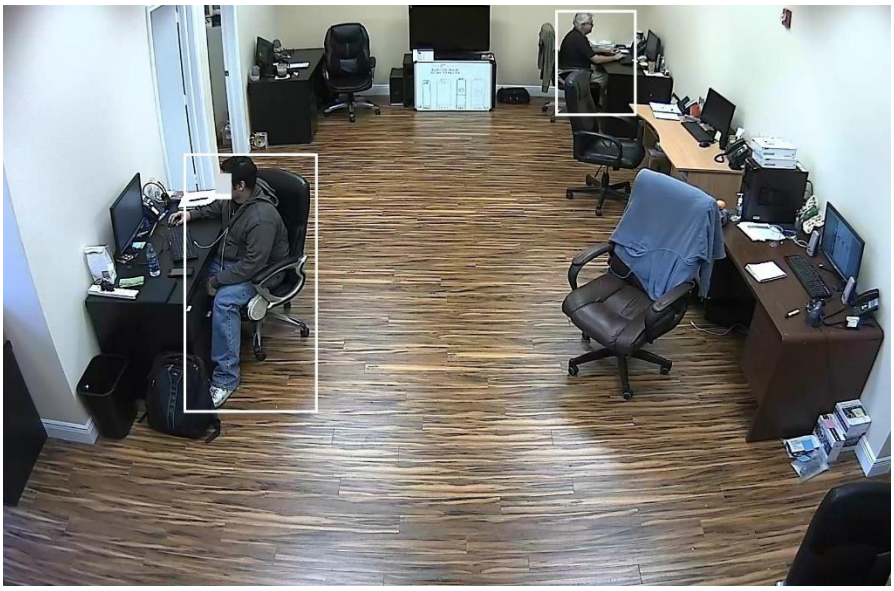

Figura 2.4 – Esempio di identificazione degli oggetti

2.3.2 Rimozione del rumore di sottofondo

Una volta identificati i pixels in movimento per ogni fotogramma, è necessario raggrupparli in regioni, a cui ci si riferisce come blobs, così che i pixels appartenenti ad un singolo oggetto possano essere raggruppati insieme.

Si ricorre quindi all'Accrescimento delle Regioni, un metodo di segmentazione dell'immagine che determina quali pixels possono essere considerati parte dello stesso elemento.

L'Accrescimento delle Regioni si basa su tecniche quali l'erosione morfologica dell'immagine e filtri di dilatazione, i quali sostituiscono il valore del pixel con il valore massimo o minimo della matrice per delimitare i contorni dell'oggetto e rimuovere i blobs troppo piccoli e quelli considerati rumori di sottofondo.

Talvolta i singoli oggetti in movimento vengono separati in modo errato in due o più sottoregioni per mancanza di connettività tra i pixels che può verificarsi a causa dell'occlusione di altri elementi, come ad esempio un albero, di bruschi cambiamenti di velocità, posizione o orientamento degli oggetti.

Per risolvere questo problema è possibile correggere i parametri dell'algoritmo di video analisi ricorrendo ad un'applicazione connessa alla videocamera per configurare i parametri analitici.

2.3.3 Tracciamento degli oggetti

Dopo aver individuato i singoli oggetti, è necessario stabilire una corrispondenza per ogni fotogramma tra i blobs. Per far questo vengono analizzati i valori cromatici dei pixels e il riferimento cartesiano del centro di ogni oggetto.

Quindi viene calcolata una matrice di corrispondenza. Ogni riga della matrice corrisponde a una traccia dal fotogramma precedente e ogni colonna coincide con un blob del fotogramma corrente. Ogni elemento della matrice è la somma della distanza euclidea definita secondo i valori di posizione (valori di coordinate Y e X) e la distanza euclidea basata sui valori di colore (valori R, G e B) e rappresenta il grado di corrispondenza tra un blob e l'altro come distanza euclidea. Minore è il distanza, migliore è la corrispondenza.

Per ogni blob in movimento nel primo fotogramma, viene creata una nuova traccia. Nei fotogrammi successivi, in base ai valori della matrice di corrispondenza, i blobs del fotogramma corrente vengono ricercati come traccia nel fotogramma precedente. Ai blobs corrispondenti vengono assegnati dei numeri per tracciarli rispetto al fotogramma precedente. Le statistiche della traccia sono aggiornate con la posizione e le informazioni sul colore del blob. Se non viene trovata alcuna corrispondenza, la traccia

viene dichiarata persa. La traccia non viene cancellata ma viene mantenuta in memoria per ancora qualche fotogramma, tipicamente 5, nel caso in cui venga trovata una corrispondenza. Se compare un nuovo blob nel nuovo fotogramma che non può essere abbinato a una traccia esistente, ne viene generata una nuova.

L'elenco delle tracce attive viene mantenuto per ogni fotogramma. Se c'è un legame tra più tracce per un singolo blob, allora viene considerato quello più vecchio. Questo perché la traccia considerata per un numero maggiore di fotogrammi ha una probabilità minore di essere rumore di sottofondo (figura 2.5).

Figura 2.5 –Tracciamento del movimento degli oggetti in un fotogramma

Nel caso di un ambiente senza rumori di sottofondo e una perfetta segmentazione dei blobs, l'operazione di tracciamento è semplice. Tuttavia, il livello di imprecisione nella segmentazione finale può anche essere alto e ciò rende la corrispondenza dei blobs un problema difficile da risolvere.

Di seguito vengono riportati alcuni dei motivi che portano a questo errore.

- I blobs potrebbero non essere rilevati a causa delle loro dimensioni ridotte;
- I blobs possono essere occlusi da un altro oggetto statico o in movimento;
- Alcuni oggetti in movimento, come i rami degli alberi, sebbene siano blobs in movimento, non sono di interesse per la nostra applicazione;
- Alcuni blobs potrebbero venire uniti in modo scorretto;
- Gli oggetti che entrano (o escono) nella scena non hanno blobs corrispondenti nei fotogrammi precedenti (o successivi);
- Un oggetto può cambiare orientamento o essere temporaneamente occluso, portando a un cambiamento nelle sue superfici visibili, variando la sua statistica dei colori e le sue dimensioni.

3 ANALISI DI COMPORTAMENTO BASATA SULLA VIDEO ANALISI

In questo capitolo viene introdotta l'Analisi di Comportamento basata sulla Video Analisi, un insieme di tecnologie *software* e *hardware* utilizzato per monitorare le intenzioni ed il comportamento delle persone.

Nel primo paragrafo viene fornita una descrizione del sistema, di come viene applicato in diversi contesti e quali sono i benefici che offre a chi ne usufruisce.

Il secondo paragrafo approfondisce il suo funzionamento, seguendo il flusso dei dati, a partire dall'algoritmo di video analisi, passando per il modello statistico e di inferenza fino alla trasformazione eseguita dall'ETL.

Viene fatta una breve digressione sull'algoritmo applicato a questo studio.

Successivamente nel terzo paragrafo viene riportato un esempio di applicazione ad un caso reale, mentre nell'ultimo vengono delineati i limiti di questo modello.

3.1 Introduzione all'Analisi di Comportamento

L'Analisi di Comportamento basata sulla Video Analisi consiste in un sistema realizzato da Discovery Reply che sfrutta le funzionalità delle telecamere di video sorveglianza e le combina con quelle della video analisi al fine di estrapolare dati statistici e studiare il comportamento di clienti all'interno di un punto vendita in tempo reale nel rispetto della *privacy*.

Questo strumento viene impiegato per l'identificazione e il conteggio di persone e oggetti all'interno delle aree video sorvegliate, per la classificazione e analisi di caratteristiche quali l'età, il genere e l'etnia delle persone, per monitorare l'entrata, il tempo di permanenza e l'uscita delle persone, per tracciare i percorsi compiuti, per stimare l'affollamento di un'area e mappare le zone calde e le zone fredde.

L'Analisi di Comportamento viene utilizzato per monitorare sia piccole sia grandi aree, porte d'ingresso, corridoi, code e permette di confrontare i dati ricavati da più aree osservate contemporaneamente o dalla stessa area ma in momenti diversi.

Questo sistema fornisce una serie di informazioni quantitative e qualitative che offrono molte applicazioni per scopi commerciali e offre innumerevoli vantaggi specialmente per il settore del *retail* (figura 3.1).

Infatti, combinando le caratteristiche della *video security* con quelle della video analisi per raccogliere dati statistici, quest'ultima può essere utilizzata come uno strumento di marketing per sostituire le impressioni soggettive dello *store manager* con KPI quantitativi e riportare i risultati ottenuti sotto forma di *dashboards* e resoconti utili per la gestione del processo decisionale.

I *Key Performance Indicators* individuati per interpretare i dati raccolti sono:

- Il numero dei visitatori, ovvero la distribuzione nel tempo delle persone che visitano un'area, tenendo in considerazione gli ingressi e le uscite da un varco e la direzione che prendono;
- Il tempo di permanenza dei visitatori, ossia l'intervallo di tempo che intercorre tra l'arrivo di una persona in un'area e il momento in cui imbocca l'uscita;
- La classificazione dei visitatori in base all'età (bambino, giovane, adulto, anziano), al genere (maschio o femmina) e all'etnia (caucasica, africana, asiatica);
- L'attrattività, che viene misurata come il rapporto tra il numero di visitatori che entrano in un'area e il numero di passaggi davanti alla vetrina al fine di misurare l'interesse scaturito guardando la vetrina;
- Le mappe calde, ovvero la rappresentazione grafica dei dati che vengono individuati da fasce di colore per indicare, ad esempio, le zone di maggior passaggio, ricorrendo ai colori più freddi come il verde o il blu per indicare le aree meno visitate e a quelli più caldi per quelle di maggior interesse.

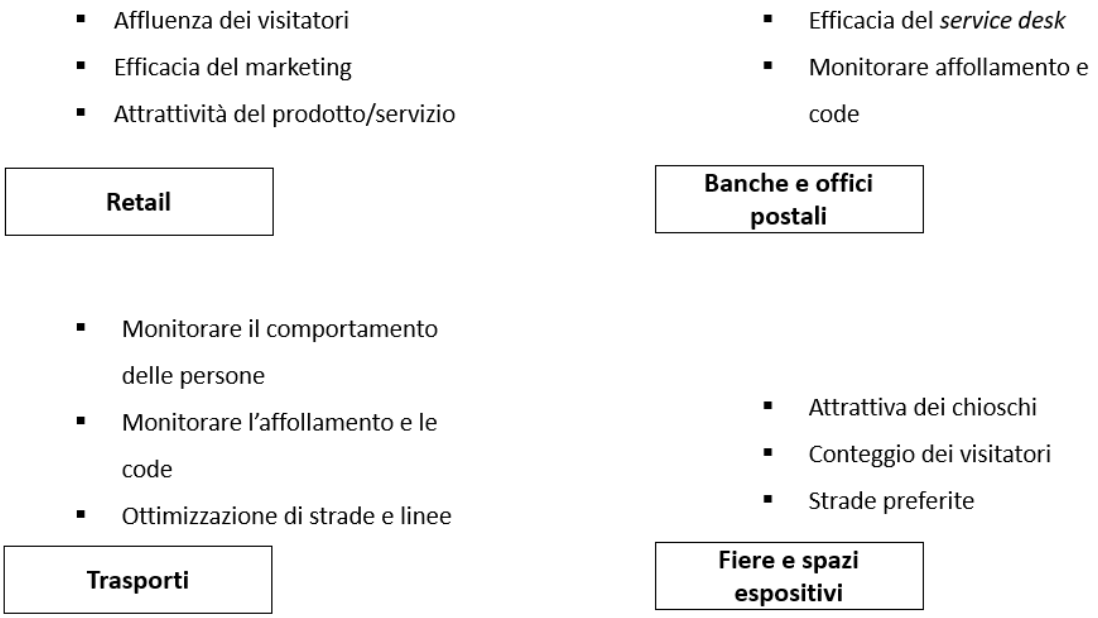

Figura 3.1 – Key Performance Indicator a seconda del settore

Grazie all'Analisi di Comportamento è quindi possibile misurare l'efficacia delle attività di marketing. Ciò permette di sapere quante persone visitano il negozio e quanto sono interessate ai prodotti in vetrina o ai video pubblicitari proiettati sugli schermi, se il target individuato corrisponde con quello effettivo, quali sono le strade percorse di più all'interno di un supermercato e dove la gente si sofferma più a lungo, apportando un contributo sostanziale quando si tratta di mettere in atto una strategia di vendita.

Tuttavia, non ne beneficiano solo attività commerciali come negozi, centri commerciali, fiere, spazi espositivi, ma anche scuole, mense, chiese, aeroporti, stadi, strade e tutti quei luoghi che hanno la necessità di individuare eventi anomali al fine di rendere quelle aree più sicure prevendendo crimini e incidenti, proteggere le persone, monitorare il comportamento delle persone e dei veicoli con l'obiettivo di ottimizzare i percorsi ed evitare il crearsi di code, controllare la partecipazione pubblica per evitare il sovraffollamento e garantire l'efficacia ai *service desks* come nel caso delle banche e degli uffici postali.

3.2 Funzionamento dell'Analisi di Comportamento

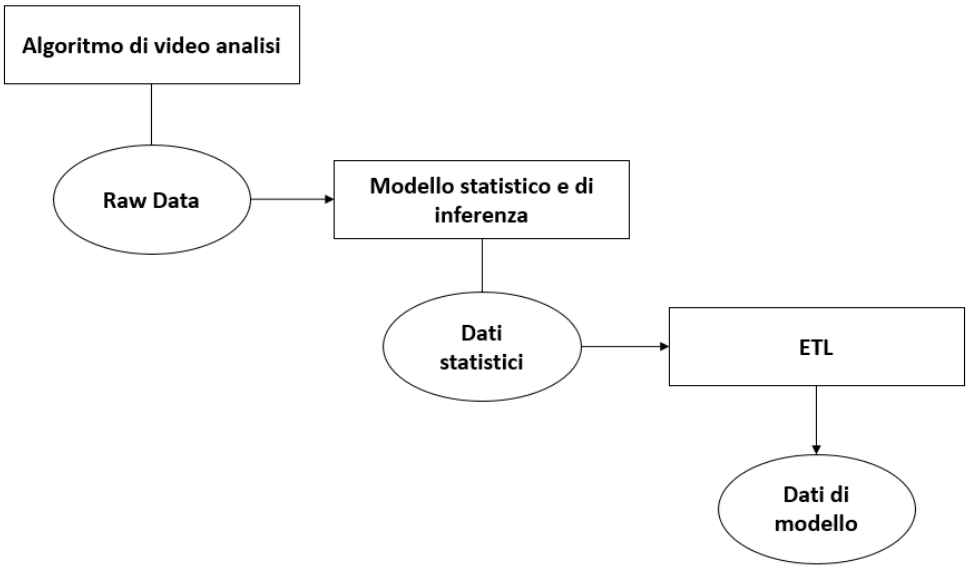

Figura 3.2 – Schema sul Funzionamento dell'Analisi di Comportamento

3.2.1 Algoritmi di video analisi

L'Analisi di Comportamento basata sulla Video Analisi permette l'acquisizione dei dati dalla video analisi applicando gli algoritmi offerti dai seguenti sistemi ai flussi di dati video: *VTrack*, sviluppato da TechnoAware Srl e *ccEngine SDK* fornito da Correlance Srl.

I procedimenti sistematici di calcolo utilizzati sono:

- *VTrack.AreaCounting*: rileva quando le persone entrano in un'area monitorata, il tempo di permanenza e il momento in cui escono;
- *VTrack.Gateflow*: rileva l'attraversamento di un varco virtuale nelle direzioni configurate;
- *VTrack.OccupancyRate*: rileva il tasso di occupazione di un'area (quale zona è occupata, quando è occupata e quanto percentualmente è occupata);
- *Vtrack.Hotzone*: rileva il movimento di soggetti presenti in un'area virtuale ed elabora i valori sommandoli gli uni agli altri (immagine cumulativa) creando così una mappa calda;

- *ccEngine.FaceClassifier*: classifica i volti in base all'età (bambino <14, giovane <30-35, adulto <50-60, anziano > 60), al genere (maschio o femmina) e all'etnia (caucasica, africana e asiatica).

L'algoritmo preso in analisi per il miglioramento dei meccanismi di stima è il *VTrack.Gateflow*.

3.2.1.1 Approfondimento VTrack.Gateflow

L'algoritmo *VTrack.Gateflow* conta i passaggi attraverso un *gate* virtuale e individua la direzione dell'attraversamento.

I singoli *gates* virtuali, ovvero i varchi fisici monitorati singolarmente che forniscono i dati statistici vengono accorpati e considerati come un'unica area logica costituita dalla somma di tutti i varchi.

Ad esempio, le porte di ingresso di un punto vendita monitorate singolarmente sono considerate varchi fisici, mentre il loro insieme un varco virtuale.

I valori di conteggio che interessano il numero di visitatori $Visitors.count_*_{(i)}$, in un intervallo T_i sono dati dall'elaborazione degli eventi di ingresso e uscita registrati negli N *gates* virtuali collegati all'area di riferimento del modello, ai quali vengono assegnati valori identificativi.

Il valore iniziale di $Visitors.count_start_{(i)}$, ovvero del numero di visitatori nel momento in cui viene accesa la videocamera si assume noto e uguale a 0 nel caso di default o uguale al valore di $Visitors.count_end_{(i-1)}$ se esistente.

Le formule applicate dall'algoritmo per calcolare il valore finale del numero dei visitatori sono riportate e descritte qui di seguito.

$$Visitors.count_end_{(i)} = Visitors.count_start_{(i)} + \sum N\ S_{(n)}$$

Dove:

- $Visitors.count_start_{(i)} = Visitors.count_end_{(i-1)}$

- *N* = zona N identificata da id
- $S_{(n)} = \sum_j VTrack.GF_{(j)}.count * P_n$

VTrack.GF$_{(i)}$ corrisponde al segno del valore di rilevazione dell'evento. Può essere positivo e viene indicato con il segno + se una persona attraversa il *gate* per entrare in un'area o negativo e viene indicato con il segno – se ne esce.

Per ogni *gate* virtuale viene anche definita la polarità P_n. Questo parametro agisce come moltiplicatore del valore *VTrack.GF$_{(i)}$.count*, e consente di interpretare correttamente il segno positivo o negativo del singolo contributo che interviene nella sommatoria.

Per il calcolo del numero di visitatori medio *Visitors.count_mean$_{(i)}$* in un intervallo di tempo T_i è importante tenere in considerazione la distribuzione temporale degli eventi di ingresso e uscita. Se consideriamo tutti gli eventi i di ingresso/uscita tali che $i \in T_i$ e indichiamo con t_i l'istante temporale dell'evento, il valore dei visitatori al tempo t_i è dato dalla seguente formula:

$$V(t_i) = V(t_i\text{-}1) + VTrack.GF_i.count * P_n$$

Dove il termine *VTrack.GF$_{(i)}$* è esattamente l'evento i cui ci riferiamo, tale che:

$$t_i V = Track.GF_i.time$$

La formula per calcolare $V(t_i)$ configura quindi una serie temporale che rappresenta l'andamento dei valori nel numero assoluto dei visitatori nell'intervallo di tempo T_i.

Dalla serie individuata si possono quindi estrarre i valori medi, minimi e massimi dei visitatori nell'intervallo T_i.

Il valore medio, in particolare, può essere ricavato come media pesata di ogni valore nella serie rispetto al tempo di persistenza del valore nella serie, cioè al valore:

$$\Delta t_i = t_{i+1} - t_i$$

3.2.2 Modello statistico e di inferenza

Gli algoritmi di video analisi, come i *VTrack.Gateflow*, generano *Raw Data* (o dati grezzi), che rappresentano elementi che non sono ancora stati processati né manualmente né da un *software*.

Il processo di trasformazione da dati grezzi a dati di modello viene compiuto dai Modelli di Analisi Statistica e di Inferenza, che aggregano i R*aw Data* per estrapolare i dati statistici e descrivono come cambiano nel tempo.

I Modelli di Analisi Statistica e di Inferenza sono disponibili per modellare correttamente un'osservazione costante di:

- visitatori, ovvero il numero di persone che si recano in un'area e il loro tempo di permanenza;
- ingressi, ovvero il numero di persone che entrano o escono da un varco e la direzione che prendono;
- mappe calde.

I dati statistici estrapolati sono rappresentati come valori discreti nel tempo. L'intervallo di tempo T_i considerato è costante e si assume arbitrariamente della durata di un'ora.

Il modello descritto finora è applicabile solamente nel caso ideale, ovvero nel caso in cui il numero degli ingressi attraverso un varco coincida con quello delle uscite. Tuttavia è praticamente impossibile che si verifichi questa situazione, a causa dei numerosi errori che concorrono ad aumentare la differenza tra i valori di ingresso e uscita.

I motivi per cui il conteggio risulta impreciso sono molteplici. Vengono riportati qui di seguito alcuni esempi.

- Gruppo di persone: in alcuni casi è impossibile conteggiare singolarmente due o più persone affiancate che attraversano uno o entrambi i portali. In questo caso l'errore deriva dal fatto che vengono conteggiati o meno passaggi di quelli realmente avvenuti o il numero preciso di passaggi solo su uno dei due portali.

- Ombre: l'ombra di una persona che cammina vicino ad un varco di ingresso o di uscita può essere conteggiata come doppio passaggio o ingresso/uscita da un varco.
- Condizioni ambientali e metereologiche: le condizioni ambientali influiscono in modo rilevante sul rilevamento delle persone. In condizioni di luminosità esterna variabile (assenza di luce o sovraesposizione), pioggia e/o nebbia può diventare difficile rilevare il passaggio di individui.
- Mancanza di manutenzione: le telecamere richiedono manutenzione regolare, in quanto un sensore sporco potrebbe non essere in grado di individuare lo spostamento di persone.
- Ambientazione interna ed esterna: è necessario che le superfici su cui vengono posizionate le telecamere siano idonee, i soffitti sufficientemente alti e bisogna accertarsi che la pavimentazione e/o il tappetino di ingresso non siano di colore troppo scuro poichè impedirebbero di distinguere il passaggio di una persona.
- Visitatori VS lavoratori: non è possibile riconoscere e quindi stornare dal conteggio i passaggi degli addetti alle vendite o dei responsabili.
- Delimitazione dell'area: la telecamera installata potrebbe non essere in grado di monitorare tutta la superficie desiderata, ma solo parzialmente. Una soluzione potrebbe essere ricorrere a più di una telecamera. Tuttavia, anche in questo caso emergerebbero altre problematiche, come il doppio conteggio delle persone nel caso avvenga la sovrapposizione delle aree osservate o il mancato conteggio di coloro che attraversano le aree di confine se le aree monitorate sono adiacenti. Talvolta non è possibile adottare questa soluzione per una questione di costi.

Queste sono solo alcune tra le tante motivazioni che comportano delle imprecisioni per quanto riguarda il conteggio dei passaggi attraverso un *gate* virtuale.

È necessario quindi considerare un meccanismo correttivo alla serie $V(t_i)$.

Attualmente il sistema per correggere gli errori di conteggio sui singoli varchi prevede l'utilizzo di una soluzione elaborata con tecniche ETL.

3.2.3 Extract, Transform and Load

L'*Extract, Trasform and Load* (o abbreviato ETL) consiste in un processo di estrazione dei dati, trasformazione e inserimento in un sistema di sintesi. Più in dettaglio, si occupa di:

- Estrarre i dati (*input*) da *database* transazionali (ad esempio SAP, sistema ERP, sistema operativo, ...) e convertirli in un solo formato pronto ad essere processato;
- Trasformare i dati selezionando solo quelli di interesse, eliminando i duplicati, applicando dei filtri, dividendo una colonna in colonne multiple, accorpando dati da risorse differenti, spostando righe o colonne o applicando qualsiasi tipo di validazione dei dati.
- Memorizzare i dati (*output*) in un sistema di sintesi.

L'obiettivo di questo sistema è quello di uniformare i dati che provengono da sorgenti differenti in modo che siano pronti per essere elaborati dai *tools* di analisi.

Il parametro di esecuzione che governa l'ETL e che viene applicato in questa circostanza alla serie $V(t_i)$ è il *Correction Model Type*, ovvero un meccanismo che governa il calcolo e che può assumere due valori: *enter_bind*, nel caso vengano privilegiate le entrate ed *exit_bind*, nel caso vengano invece privilegiate le uscite.

Prima di proseguire con la spiegazione, è importante ora introdurre brevemente il concetto di polarità P_n.

Per polarità si intende un parametro configurabile che individua la direzione dell'attraversamento del varco. Quando viene registrato il valore 1 significa che la persona che ha attraversato il *gate* è entrata, mentre quando assume il valore 2 che è uscita.

Gli errori di conteggio sui singoli varchi hanno generalmente una polarità P_n, cioè sbagliano di più in un verso che nell'altro. I singoli contributi sono quindi soggetti ad un errore percentuale di misurazione all'origine. Per inserire questo correttivo di calcolo, lo si considera parte della polarità P_n, non più conteggiata come intero (1; -1) ma con valore *double* che include una percentuale che compensa l'errore (ad esempio, 1,025). Con

questa impostazione tutti i valori del modello e di polarità vanno configurati e trattati come valori *double* agli effetti del calcolo.

Inserire un valore correttivo consente un calcolo più attendibile dei valori di modello quando esiste una valutazione dell'errore basata su una osservazione diretta.

Si prosegue ora la spiegazione sul funzionamento del *Correction Model Type*.

Quando viene applicato l'*enter_bind*, i valori di entrata indicati nel dataset con il numero 1 rimangono invariati e viene applicato un fattore correttivo a quelli di uscita indicati con il numero 2 al fine di portare il valore del Δ a 0. Mentre, quando si ricorre all'*exit_bind*, si applica il fattore correttivo ai valori di entrata.

Viene apportato un esempio al fine di semplificare la comprensione di questo concetto.

Si supponga che l'area logica di un punto vendita sia costituita da due varchi, uno di entrata e uno di uscita. Il negozio apre alle 10 di mattina e chiude alle 6 di pomeriggio. All'orario di apertura t_0, $V(t_0)$ è pari a 0 e ciò significa che non sono presenti persone nel negozio. Alle ore 18, ovvero quando chiude il negozio, $V(t_i)$ dovrebbe essere nuovamente uguale a 0, ma è pari a 10 poichè sono stati registrati 30 ingressi (figura 3.3) e 20 uscite (figura 3.4).

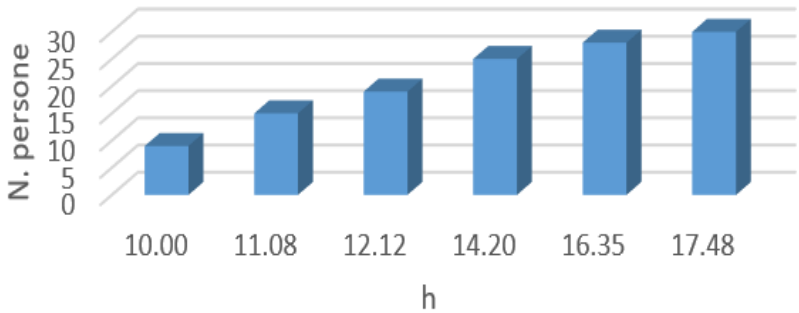

Figura 3.3 – Numero delle persone entrate dal varco

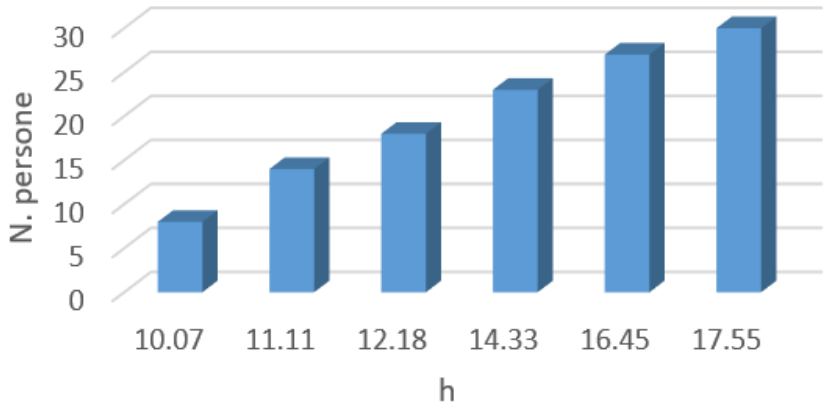

Figura 3.4 – Numero delle persone uscite dal varco

Per correggere questa differenza è necessario applicare il valore di convergenza.

In questo caso si decide di applicare l'*enter_bind*, mandando a convergenza il valore delle uscite, ovvero moltiplicandolo per il fattore di correzione e mantenendo fisso il valore degli ingressi (tabella 3.5 e figura 3.6).

Orario (h)	**Ingressi (n. persone)**	**Uscite non corrette (n. persone)**	**Fattore di correzione**	**Uscite corrette (n. persone)**
10.07	9	8	1,037	8 * 1,037 = 9
11.11	15	14	1,071	14 * 1,071 = 15
12.18	19	18	1,056	18 * 1,056 = 19
14.33	25	23	1,087	23 * 1,087 = 25
16.45	28	27	1,037	27 * 1,037 = 28
17.55	30	20	1,5	20 * 1,5 = 30

Tabella 3.5 – Tabella che mostra come viene applicato il fattore di correzione

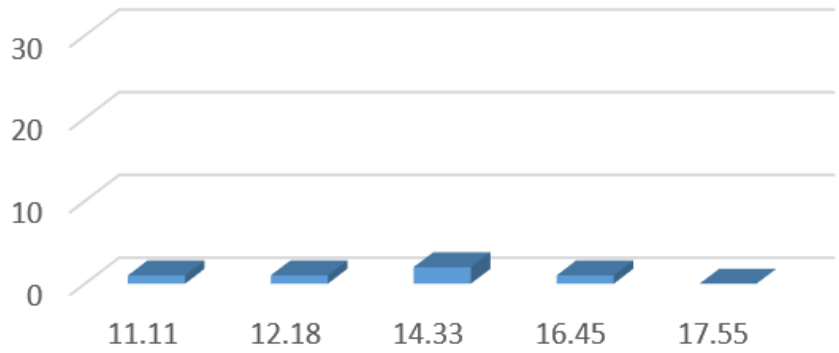

Figura 3.6 – Numero delle persone ancora presenti nel punto vendita considerando gli ingressi e le uscite amplificate dal fattore correttivo

Applicando il fattore correttivo ai valori di uscita, all'orario di chiusura del negozio, V(t_i) è uguale a 0.

È possibile quindi calcolare l'ipotetica cumulativa di persone all'interno dell'area per l'intera giornata. E' anche possibile analizzare i valori ottenuti ottenendo in questo modo la stima dei visitatori su base oraria (figura 3.7).

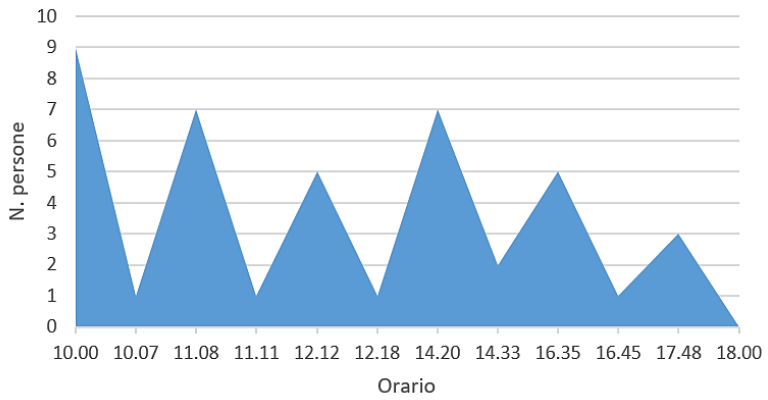

Figura 3.7 – Stima dei visitatori su base oraria

Una volta estratti e corretti i dati statistici ricorrendo all'ETL, si ottengono i dati di modello.

Queste informazioni vengono visualizzate su una *dashboard* accessibile da un *web browser* utilizzando un computer o un tablet. I dati rilevanti sono mostrati sotto forma di istogrammi e grafici a torta per facilitarne la lettura e l'esportazione giornaliera (figura

3.8, figura 3.9, figura 3.10). Queste informazioni vengono archiviate e possono essere analizzate in qualsiasi momento, comparando lo stesso KPI in giorni diversi o durante lo stesso giorno ma di punti vendita differenti.

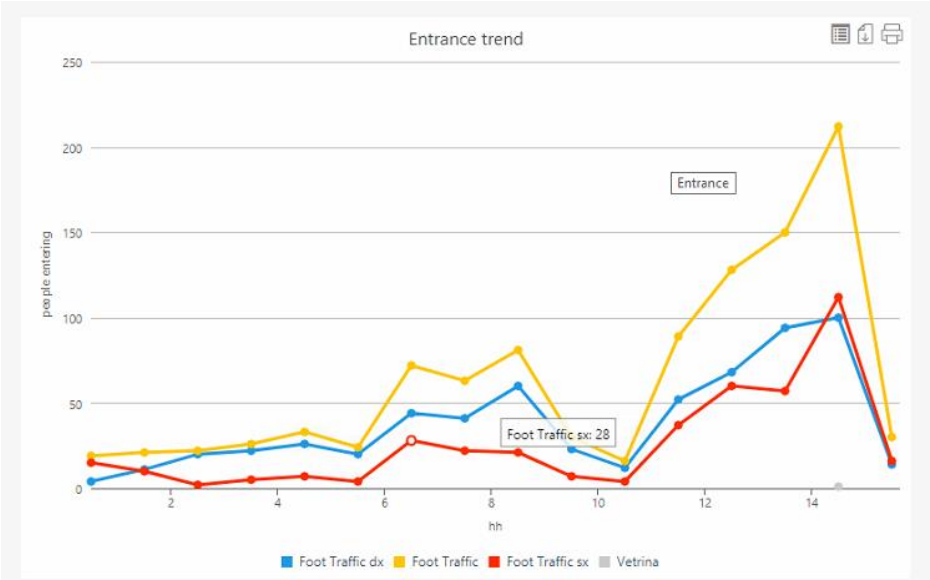

Figura 3.8 – Esempio di schermata della dashboard che rappresenta il numero e la direzione di passaggi in una giornata davanti ad una vetrina

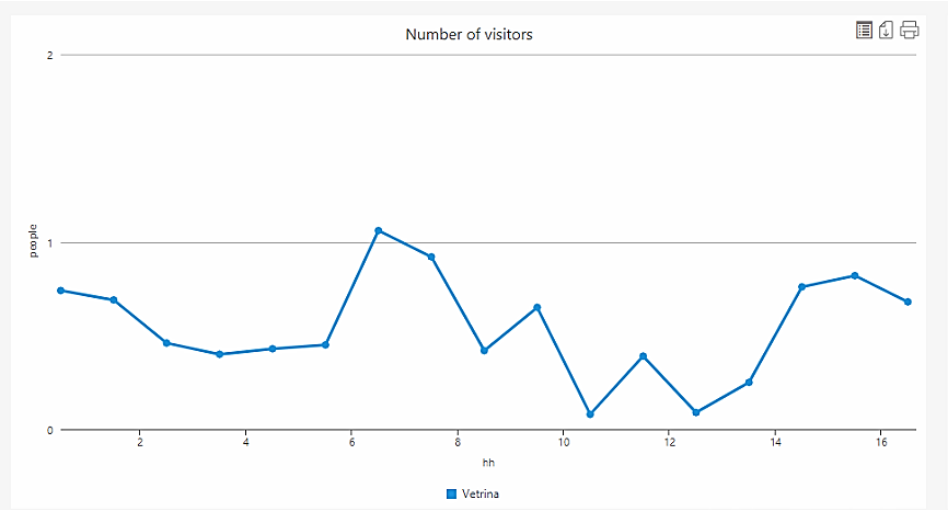

Figura 3.9 – Esempio di schermata della dashboard che rappresenta il numero di visitatori in un'area monitorata in un giorno

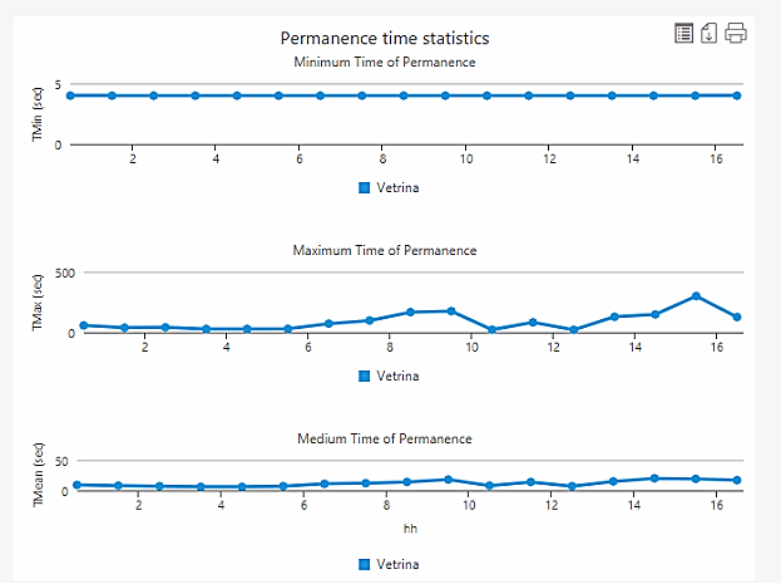

Figura 3.10 – Esempio di schermata della dashboard che rappresenta il tempo di permanenza minimo, massimo e medio dei visitatori

3.3 Applicazione del modello a un caso reale

Al fine di permettere un confronto tra il modello correttivo attuale basato sul parametro di esecuzione che governa l'ETL, il *Correction Model Type* e quello realizzato e descritto nel capitolo successivo, si applica l'*enter_bind* ad un caso reale.

Si considerano le registrazioni avvenute il 1 maggio 2016 in un punto vendita di telefonia a Roma.

Per ogni fascia oraria indicata nella colonna *time [PK] integer* vengono riportate il numero di persone entrate e il numero delle persone uscite da un varco, corrispondenti rispettivamente alle colonne *N persone entrate* e *N persone uscite*, la cui direzione è identificata dai valori 1 (ingresso) e 2 (uscita) nella colonna *direction_id [PK] integer* (vedi immagine).

Per stimare il fattore di correzione, è necessario calcolare il rapporto tra il numero di ingressi e quello delle uscite in valore assoluto.

Ad esempio, per la fascia oraria 9 – 10 di mattina, 73 persone entrano in negozio e 56 ne escono. Il valore di convergenza è pari a:

73 / |-56| = 1,303571

Quindi, moltiplicando il valore corrispondente alle uscite per il valore di convergenza, è possibile correggerne il risultato.

In tabella sono riportati tutti i fattori correttivi per il giorno preso in esame (tabella 3.11).

date [PK] character varying	time [PK] integer	direction_id [PK] integer	N persone entrate	N persone uscite	Fattore di correzione	N persone uscite corretto
01/05/2016	8	1	46	0		
01/05/2016	8	2	46	-34	1,352941	46
01/05/2016	9	1	73	-34		
01/05/2016	9	2	73	-56	1,303571	73
01/05/2016	10	1	106	-56		
01/05/2016	10	2	106	-85	1,247059	106
01/05/2016	11	1	146	-85		
01/05/2016	11	2	146	-113	1,292035	146
01/05/2016	12	1	177	-113		
01/05/2016	12	2	177	-151	1,172185	177
01/05/2016	13	1	192	-151		
01/05/2016	13	2	192	-168	1,142857	192
01/05/2016	14	1	225	-168		
01/05/2016	14	2	225	-211	1,066351	225
01/05/2016	15	1	289	-211		
01/05/2016	15	2	289	-255	1,133333	289
01/05/2016	16	1	299	-255		
01/05/2016	16	2	299	-271	1,103321	299

Tabella 3.11 – Tabella che riporta il conteggio degli ingressi e delle uscite per il 1 maggio 2016

Considerando la fascia oraria presa in esame, non è tuttavia possibile identificare quali delle 17 persone si trovavano realmente in negozio e quali non sono state conteggiate dalle telecamere.

Lo stesso discorso riguarda tutte le altre fasce orarie ad eccezione dell'ultima, in quanto nel caso ideale alla chiusura del punto vendita il numero di ingressi dovrebbe equivalere quello delle uscite. In questa circostanza la differenza tra i due valori può essere imputata all'errore.

Da questo esempio quindi emerge che il *Correction Model Type* presenta numerosi limiti e non è sufficiente per correggere in modo preciso l'errore.

3.4 Limiti del modello

Come è emerso dall'esempio precedente, il *Correction Model Type* presenta dei limiti e non è sufficiente per correggere in modo preciso l'errore.

Uno tra questi consiste nel non calcolare correttamente la differenza tra ingressi ed uscite. Infatti questo delta è dato dalla somma dell'errore dell'algoritmo di video analisi con il numero di persone presenti all'interno del punto vendita che non sono ancora uscite. Applicando il fattore di correzione al termine di ogni ora, vengono quindi alterati i risultati.

Questo parametro potrebbe essere applicato solo all'orario di chiusura del negozio, ma così facendo si perderebbe l'opportunità di ottenere informazioni aggiornate in tempo reale.

Come conseguenza, ad esempio, non sarebbe più possibile monitorare il traffico all'interno di una mensa o una coda in cassa e indicare il momento migliore per usufruire del servizio al fine di minimizzare il tempo di attesa.

È importante tener presente i limiti di questo sistema anche quando si considera un orario di apertura 24/7.

Quando viene installata una telecamera in un ristorante o un negozio aperto 24 ore al giorno, non è possibile sapere quante persone erano già presenti nell'area monitorata all'istante t_0. Il conteggio quindi risulta inesatto.

Supponendo che l'algoritmo di video analisi e la telecamera forniscano dei dati precisi e che non sia necessario tenere in considerazione un modello di correzione, il numero delle persone uscite risulterebbe sempre maggiore di quelle entrate. Verrebbero infatti conteggiate quelle già presenti nell'area al momento dell'installazione di cui però non era stato rilevato l'ingresso. Il flusso potrebbe temporaneamente bilanciarsi, per poi ritornare sempre ad una condizione simile a quella iniziale. Ad esempio, si consideri il caso in cui la telecamera venga messa in funzione in corrispondenza della porta d'ingresso di un punto di ristorazione quando sono già presenti 35 persone. Il sistema viene acceso per la prima volta alle 11 AM.

La tabella 3.12 riporta gli ingressi, le uscite e il numero di persone all'interno dell'area monitorata suddivisi in base alla fascia oraria per il primo giorno di rilevamento degli attraversamenti del varco.

Orario	**Ingressi**	**Uscite**	**N persone realmente nell'area**
10 AM			35
11 AM	10	12	33
12 PM	20	22	31
1 PM	27	34	24
2 PM	23	39	15
3 PM	11	14	12
4 PM	3	8	1
5 PM	2	1	1
6 PM	0	1	0
7 PM	13	0	13
8 PM	24	12	25
9 PM	28	31	22
10 PM	8	28	2
11 PM	5	6	1
Tot.	174	208	

Tabella 3.12 – Numero di ingressi, di uscite e persone presenti nell'area monitorata

Al termine della giornata, il numero di persone entrate è pari a 174, mentre il numero di quelle uscite corrisponde a 208. Considerando che una persona si trova ancora all'interno del ristorante, la differenza tra ingressi e uscite è comunque pari a 34 persone.

Si può notare che, pur essendoci una corrispondenza tra entrate e uscite tra le 5 e le 6, l'errore accumulato inizialmente, continua a persistere.

Quindi, quando viene fatta una nuova installazione, non avendo a disposizione uno storico dei dati per calcolare il delta, non è possibile stornare dal conteggio le persone che erano già presenti nell'area e simulare che al tempo t_0 il luogo monitorato fosse vuoto.

4 STRUMENTO DI ANALISI E CORREZIONE DELL'ERRORE

Per assicurarsi che i dati elaborati dal modello predittivo siano attendibili, è necessario correggere l'errore derivante dalle imprecisioni commesse dall'algoritmo di video analisi. Infatti, durante la rilevazione del passaggio attraverso un varco virtuale, il numero di ingressi e quello delle uscite spesso non coincidono.

In questo capitolo viene descritto il modo in cui questa problematica è stata risolta.

Prima di entrare nel dettaglio della questione, sono state valutate le condizioni iniziali.

Una volta considerati i limiti della situazione di partenza, il primo passo è stato quello di raccogliere i dati grezzi dal *software* utilizzato per la gestione dei dati ricavati dalla video analisi. Successivamente i dati sono stati interpretati e sono stati individuati quelli rilevanti per lo studio.

Quindi è stato creato uno strumento per verificare la distribuzione dell'errore per il periodo di monitoraggio, per testare l'efficacia del modello senza correzione e per correggere il delta dell'errore.

4.1 Presupposti iniziali

Di seguito sono riportate alcune considerazioni sulle condizioni iniziali da cui si è partiti per la realizzazione della soluzione.

Innanzitutto, è importante chiarire che non è stato possibile ottenere alcuna informazione sulle circostanze in cui i dati sono stati rilevati come, ad esempio, le condizioni di luce, le previsioni metereologiche, la tipologia di vetrina, il possibile cambiamento di allestimento all'interno del punto vendita, etc.

Il modello elaborato per correggere l'errore è quindi il risultato di un percorso di ricerca empirica.

Inoltre, si suppone che non sia stato apportato alcun cambiamento ai parametri dell'algoritmo di video analisi che possa aver influenzato il risultato nell'arco temporale durante il quale il traffico è stato monitorato.

4.2 Raccolta dei dati

È stato possibile visualizzare i dati estrapolati dalla video analisi con *pgAdmin III*, uno strumento di progettazione e di gestione di database PostgreSQL che permette di elaborare la struttura dei database e di visualizzare i dati su un foglio di calcolo (figura 4.1).

La tabella presa in esame è stata denominata *inout* e fornisce dati sui passaggi nelle due direzioni attraverso un varco. Funziona come un contatore: mostra quante persone sono entrate e uscite in un intervallo temporale di un'ora.

Gli algoritmi che sono stati applicati per elaborare le informazioni necessarie sono due: *Gateflow* e *GenderAge*. Il primo serve per conteggiare gli attraversamenti, il secondo per classificare le persone in base al genere, all'età e all'etnia.

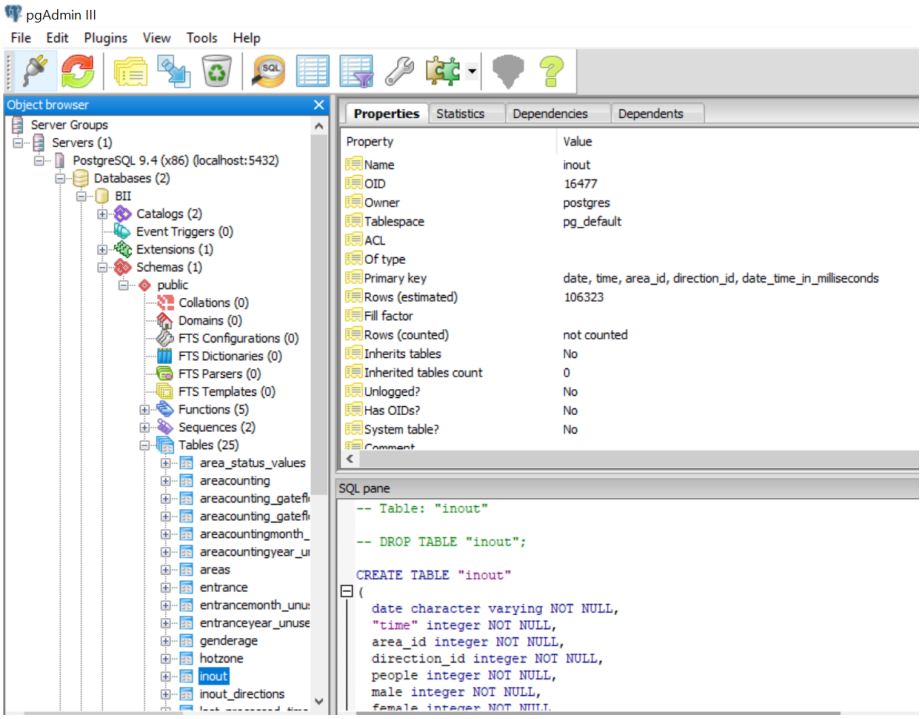

Figura 4.1 – Schermata principale di pgAdmin III

La tabella *inout* si compone delle seguenti voci:

- *Date [PK] character varying*: rappresenta la data in cui è avvenuta la registrazione (sono stati considerati i giorni a partire da aprile 2016, in quanto dati rilevanti).
- *Time [PK] integer*: indica la fascia oraria in cui è avvenuta la registrazione.
- *Area_id [PK] integer*: riporta l'id dell'area monitorata. Questo campo costituisce la chiave esterna con la tabella *Areas* (tabella 4.2).

Area_id [PK] integer	**Name character varying**	**Status integer**
17	Left	3
18	Right	3
19	All	3
20	entranceTest	3
24	Varco Sinistro	1
25	Varco Destro	1
26	Smart TV	1
27	Store	2
29	Varco Destro debug	3
30	Store debug	3

Tabella 4.2 – Tabella Areas

- *Direction_id [PK] integer*: individua le entrate con il numero 1 e le uscite con il 2.
- *People integer*: rappresenta il numero di persone entrate o uscite dal varco.
- *Male integer, Female integer, Child integer, Young integer, Adult integer, Senior integer, Causasian integer, Asian integer, African integer*: dalla sesta alla penultima colonna emerge la classificazione delle persone in base al genere, età ed etnia. In questo caso la tabella non è stata integrata con *Gender_age*, quindi queste colonne non verranno prese in considerazione.
- *Date_time_in_milliseconds [PK] bigint*: rappresenta l'ora tenendo conto del fuso orario.

Tra tutti le aree id prese in analisi dall'algoritmo (tabella 4.3), è stata esaminata l'area id 27, in quanto contenente dati rilevanti e sufficienti per giustificare i risultati del modello applicato per correggere l'errore del delta.

date [PK] character varying	time [PK] integer	area_id [PK] integer	direction_id [PK] integer	people integer	male integer	female integer	child integer	young integer	adult integer	senior integer	causasian integer	asian integer	african integer	date_time_in_millisecond s [PK] bigint
13/04/2016	11	18	1	0	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	19	1	0	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	24	1	3	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	24	2	2	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	25	1	3	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	25	2	2	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	27	1	3	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	11	27	2	2	0	0	0	0	0	0	0	0	0	1460538000000
13/04/2016	12	18	1	0	0	0	0	0	0	0	0	0	0	1460541600000
13/04/2016	12	19	1	0	0	0	0	0	0	0	0	0	0	1460541600000
13/04/2016	12	24	1	8	0	0	0	0	0	0	0	0	0	1460541600000
13/04/2016	12	24	2	3	0	0	0	0	0	0	0	0	0	1460541600000
13/04/2016	12	25	1	8	0	0	0	0	0	0	0	0	0	1460541600000
13/04/2016	12	25	2	3	0	0	0	0	0	0	0	0	0	1460541600000
13/04/2016	12	27	1	8	0	0	0	0	0	0	0	0	0	1460541600000

Tabella 4.3 – Estratto dalla Tabella InOut

4.3 Interpretazione dei dati

Il foglio di calcolo *Area id 27* è stato estrapolato dalla tabella *InOut* ordinando i dati per giorno/mese/anno, fascia oraria e id dell'area monitorata ed estraendo quelli appartenenti all'id 27 (tabella 4.4).

È stato ricavato un foglio di calcolo per l'area id 27 composto da 15197 righe.

date [PK] character varying	time [PK] integer	area_id [PK] integer	direction_id [PK] integer	people integer
13/04/2016	11	27	1	3
13/04/2016	11	27	2	2
13/04/2016	12	27	1	8
13/04/2016	12	27	2	3
13/04/2016	13	27	1	2
13/04/2016	13	27	2	2
13/04/2016	14	27	1	3
13/04/2016	14	27	2	3
13/04/2016	15	27	1	6
13/04/2016	15	27	2	7

Tabella 4.4 – Estratto dalla Tabella InOut che contiene solo l'area 27

A questa tabella è stata aggiunta una colonna denominata *conversione*, utilizzata per attribuire il segno positivo ai valori corrispondenti alle entrate e quello negativo ai valori di uscita. È stata applicata la funzione:

IF (direction_id [PK] integer = 1; + people integer; – people integer)

Il significato di questa operazione è il seguente. Se la casella selezionata parte della colonna *direction_id [PK] integer* è pari a 1, allora al valore corrispondente della colonna *people integer* viene assegnato il segno positivo, altrimenti quello negativo.

Il passo successivo è stato quello di creare un'altra colonna, denominata *sommatoria persone in negozio (giorno)*, che nel caso specifico della prima riga riporti semplicemente il valore di *conversione*, mentre dalla seconda riga in poi applichi la seguente funzione:

$$= IF\ (date\ [PK]\ character\ varying_{i} = date\ [PK]\ character\ varying_{i-1};\ conversione_{i} + sommatoria\ persone\ in\ negozio\ (giorno)_{i-1};\ conversione_{i}\ )$$

Gli ultimi dati ricavati al fine di avere una visione completa sulla situazione descritta dalla tabella *inout* sono: il numero di persone che entra e il numero di persone che esce per ogni fascia oraria.

Nel foglio di calcolo *area id 17* è stata aggiunta una riga all'inizio dove è stata riportata la data del primo giorno preso in considerazione per facilitare il conto e altre due colonne, *N persone entrate* e *N persone uscite*.

Alla prima riga di *N persone entrate* e *N persone uscite* è stato aggiunto manualmente il valore di 0 per facilitare il calcolo.

Al fine di contare il numero delle persone che entrano ed escono, sono state applicate le formule descritte qui di seguito.

Per il numero di persone entrate:

$$= IF\ (area_id\ [PK]\ integer_{i} = area_id\ [PK]\ integer_{i-1}\ ;\ IF\ (direction_id\ [PK]\ integer_{i} = 1;\ conversione_{i} + N\ persone\ entrate_{i-1}\ ;\ N\ persone\ entrate_{i-1});\ IF\ (\ area_id\ [PK]\ integer_{i} =$$

area_id [PK] integer $_{i+1}$; IF(direction_id [PK] integer $_i$ = 1 ; conversione $_I$ + $ N persone entrate $_{i-1}$; $ N persone entrate $_{i-1}$)))

Per il numero di persone uscite:

= IF (area_id [PK] integer $_i$ = area_id [PK] integer $_{i-1}$; IF (direction_id [PK] integer $_i$ = 2; conversione $_i$ + N persone uscite $_{i-1}$; N persone uscite $_{i-1}$); IF (area_id [PK] integer $_i$ = area_id [PK] integer $_{i+1}$; IF(direction_id [PK] integer $_i$ = 2 ; conversione $_i$ + $ N persone uscite $_{i-1}$; $ N persone uscite $_{i-1}$)))

Dopo aver ricavato tutti i dati necessari all'analisi, sono stati selezionati i dati rilevanti.

Sono stati considerati il numero di persone entrate, il numero di persone uscite e la differenza tra i due valori per l'ultima ora lavorativa di ogni giorno il cui orario di chiusura del punto vendita è maggiore o uguale alle 19.

Il procedimento applicato è descritto qui di seguito.

In primo luogo è stato necessario individuare quando la data della riga selezionata differiva da quella della riga successiva. È stata quindi applicata la seguente condizione alla colonna *sommatoria persone in negozio (giorno)* per evidenziare la cella contenente il risultato ottenuto a fine giornata ogni qual volta le date differiscono:

$ date [PK] character varying $_{i-1}$ <> date [PK] $ character varying $_i$

Una formattazione simile è stata applicata alla colonna *time [PK] integer,* al fine di individuare le celle contenenti un orario di chiusura pari o superiore alle 19.

Cell Value > 18

Per ottenere i dati rilevanti al calcolo, la tabella è stata filtrata prima in base alla colonna *sommatoria persone in negozio (giorno)* e poi a *time [PK] integer*, mantenendo in entrambi i casi solo i valori evidenziati.

4.4 Analisi dell'andamento dell'errore

Nel foglio di calcolo *area id 27_grafico* sono state riportate le seguenti colonne: *date [PK] character varying*, *time [PK] integer*, *people integer*, *sommatoria persone in negozio (giorno)*.

La prima e l'ultima sono state utilizzate per analizzare l'andamento dell'errore per tutto il periodo fornito, creando il seguente grafico (figura 4.5):

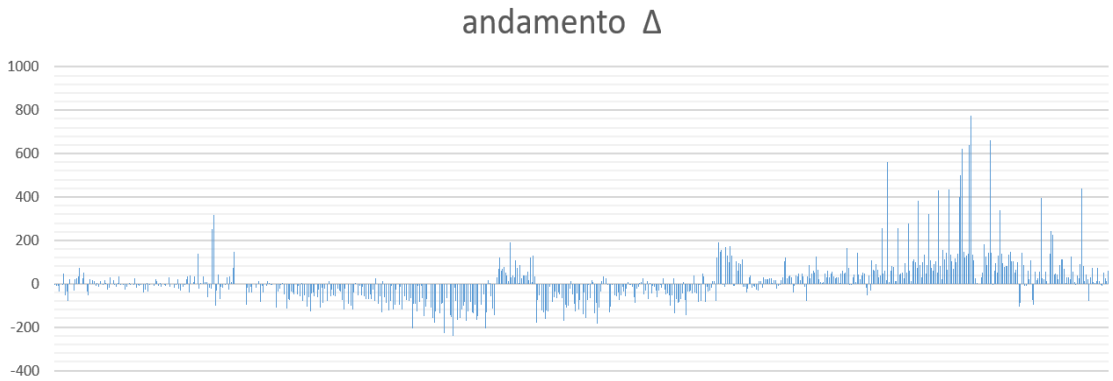

Figura 4.5 – Grafico che delinea l'andamento dell'errore

Grazie a questa rappresentazione dei dati è possibile studiare il comportamento e la variazione nel tempo dell'errore, ovvero il valore che descrive quanto si discostano gli ingressi dalle uscite.

Questo strumento può essere sfruttato per comprendere quanto i dati ricavati grazie alla video analisi siano attendibili, di quanto possano sbagliare e in quali periodi sono subentrati dei fattori che hanno influito positivamente o negativamente sull'andamento dell'errore.

Oltre al comportamento dell'errore nel tempo, è stato possibile studiare la distribuzione dei dati dell'algoritmo.

4.5 Distribuzione dei dati

In primo luogo sono stati inseriti su un foglio di calcolo i dati da analizzare, ovvero quelli corrispondenti alla colonna denominata *sommatoria persone in negozio (giorno).*

Poi sono stati calcolati due punti utili per l'analisi: il valore minimo e quello massimo.

Successivamente è stato necessario decidere in quante classi suddividere l'insieme dei dati per illustrarne la distribuzione.

Ad esempio, se i dati forniti sono 500-600, si possono considerare 20 classi. Più si vuole andare nel dettaglio, più il numero delle classi aumenta.

È stata quindi calcolata l'ampiezza dell'intervallo considerando la differenza tra il valore massimo e quello minimo e il rapporto tra il risultato ottenuto ed il numero di classi.

= (max-min) / numero classi

Sulla base di questi calcoli, è stato possibile creare una tabella che serve per costruire l'istogramma della distribuzione dei dati. La tabella si compone delle seguenti colonne:

- Intervallo: indica l'intervallo $_i$ che contiene i dati;
- V min = valore minimo considerato nell'intervallo $_i$;
- V max = valore massimo considerato nell'intervallo $_i$;
- Frequenza = numero di valori che ricadono nell'intervallo $_i$.

  La frequenza è stata calcolata come:

  = COUNTIFS (Tutti i valori della colonna Dati; ">=" & Vmin $_i$;
  Tutti i valori della colonna Dati; "<" & Vmax $_i$)

  Ad eccezione dell'ultima riga, dove viene considerata solo la prima condizione dell'COUNTIFS).
- *%Frequenza* = rapporto tra la frequenza e il numero totale di valori raccolti;
- *Densità* = rapporto tra la frequenza percentuale e l'ampiezza degli intervalli.

La tabella risultante (tabella 4.6) è la seguente:

Intervallo	V_{min}	V_{max}	Frequenza	%Frequenza	Densità
1	-241	-190,15	4	0,006088280	0,0001197302
2	-190,15	-139,3	18	0,027397260	0,0005387858
3	-139,3	-88,45	55	0,083713851	0,0016462901
4	-88,45	-37,6	107	0,162861492	0,0032027825
5	-37,6	13,25	192	0,292237443	0,0057470490
6	13,25	64,1	161	0,245053272	0,0048191401
7	64,1	114,95	59	0,089802131	0,0017660203
8	114,95	165,8	35	0,053272451	0,0010476391
9	165,8	216,65	5	0,007610350	0,0001496627
10	216,65	267,5	5	0,007610350	0,0001496627
11	267,5	318,35	2	0,003044140	0,0000598651
12	318,35	369,2	2	0,003044140	0,0000598651
13	369,2	420,05	3	0,004566210	0,0000897976
14	420,05	470,9	3	0,004566210	0,0000897976
15	470,9	521,75	1	0,001522070	0,0000299325
16	521,75	572,6	1	0,001522070	0,0000299325
17	572,6	623,45	1	0,001522070	0,0000299325
18	623,45	674,3	2	0,003044140	0,0000598651
19	674,3	725,15	0	0,000000000	0,0000000000
20	725,15	776	1	0,001522070	0,0000299325

Tabella 4.6 – Tabella per costruire l'istogramma con 20 classi

Considerando i valori delle colonne *Intervallo* e *Densità*, è stato ricavato un istogramma (figura 4.7) che rappresenta la distribuzione dell'algoritmo di video analisi.

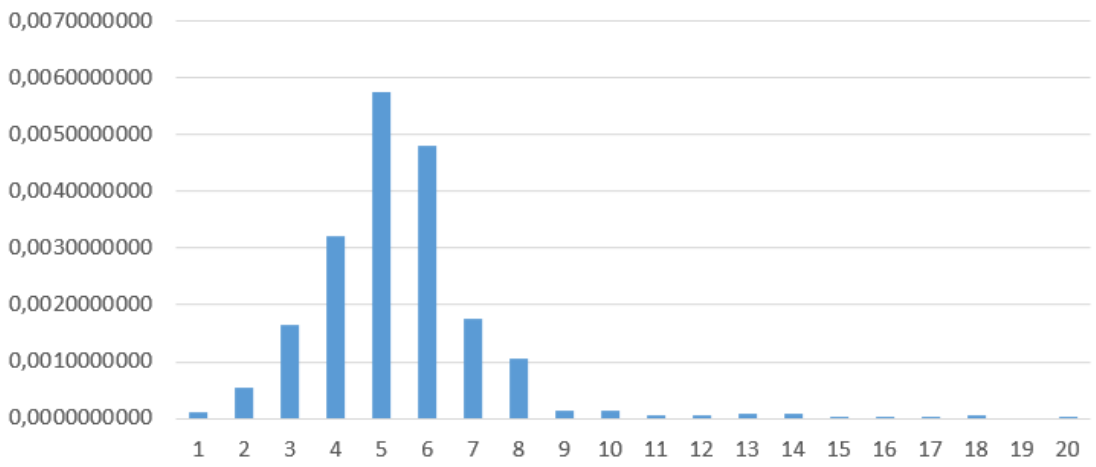

Figura 4.7 – Istogramma della distribuzione dei dati considerando 20 classi

Sull'asse delle ascisse è rappresentato il numero di intervalli, mentre sull'asse delle ordinate è rappresentata la densità, ovvero la percentuale di valori che appartengono ad un intervallo.

Grazie all'andamento, alla posizione, all'altezza e alla dispersione di questo grafico, è possibile giudicare l'efficacia dell'algoritmo di video analisi.

Qualora si volesse ottenere un grafico di precisione maggiore, è possibile suddividere i dati in più classi.

Se ne considerino ora 40 (tabella 4.8).

Intervallo	Vmin	Vmax	Frequenza	%Frequenza	Densità
1	-241	-215,575	2	0,003044140	0,000119730188
2	-215,575	-190,15	2	0,003044140	0,000119730188
3	-190,15	-164,725	7	0,010654490	0,000419055658
4	-164,725	-139,3	11	0,016742770	0,000658516034
5	-139,3	-113,875	26	0,039573820	0,001556492444
6	-113,875	-88,45	29	0,044140030	0,001736087726
7	-88,45	-63,025	49	0,074581431	0,002933389607
8	-63,025	-37,6	58	0,088280061	0,003472175453
9	-37,6	-12,175	81	0,123287671	0,004849072615

10	-12,175	13,25	111	0,168949772	0,006645025435
11	13,25	38,675	99	0,150684932	0,005926644307
12	38,675	64,1	62	0,094368341	0,003711635829
13	64,1	89,525	30	0,045662100	0,001795952820
14	89,525	114,95	29	0,044140030	0,001736087726
15	114,95	140,375	23	0,035007610	0,001376897162
16	140,375	165,8	12	0,018264840	0,000718381128
17	165,8	191,225	5	0,007610350	0,000299325470
18	191,225	216,65	0	0,000000000	0,000000000000
19	216,65	242,075	1	0,001522070	0,000059865094
20	242,075	267,5	4	0,006088280	0,000239460376
21	267,5	292,925	1	0,001522070	0,000059865094
22	292,925	318,35	1	0,001522070	0,000059865094
23	318,35	343,775	2	0,003044140	0,000119730188
24	343,775	369,2	0	0,000000000	0,000000000000
25	369,2	394,625	2	0,003044140	0,000119730188
26	394,625	420,05	1	0,001522070	0,000059865094
27	420,05	445,475	3	0,004566210	0,000179595282
28	445,475	470,9	0	0,000000000	0,000000000000
29	470,9	496,325	0	0,000000000	0,000000000000
30	496,325	521,75	1	0,001522070	0,000059865094
31	521,75	547,175	0	0,000000000	0,000000000000
32	547,175	572,6	1	0,001522070	0,000059865094
33	572,6	598,025	0	0,000000000	0,000000000000
34	598,025	623,45	1	0,001522070	0,000059865094
35	623,45	648,875	1	0,001522070	0,000059865094
36	648,875	674,3	1	0,001522070	0,000059865094
37	674,3	699,725	0	0,000000000	0,000000000000
38	699,725	725,15	0	0,000000000	0,000000000000
39	725,15	750,575	0	0,000000000	0,000000000000
40	750,575	776	1	0,001522070	0,000059865094

Tabella 4.8 – Tabella per costruire l'istogramma con 40 classi

Dividendo i dati in 40 classi si ottiene un istogramma (figura 4.9) più dettagliato.

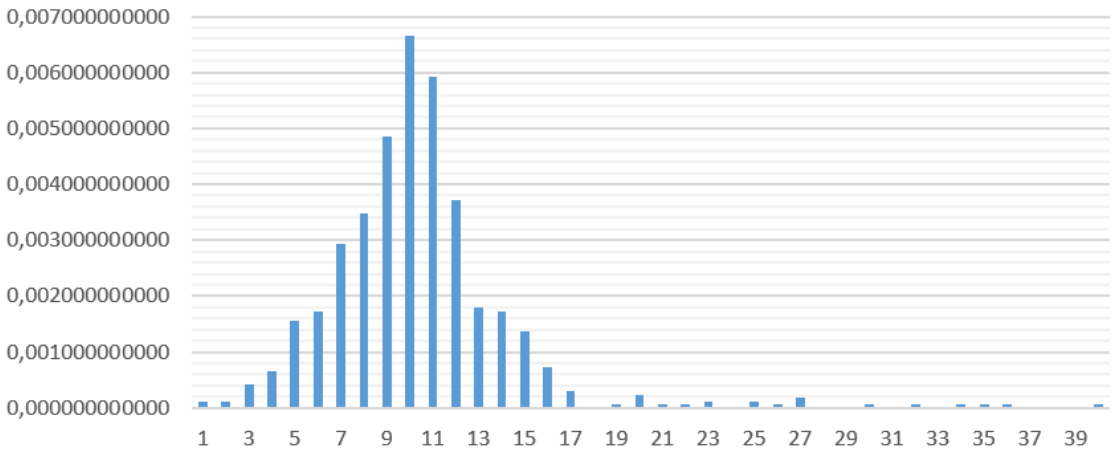

Figura 4.9 – Istogramma della distribuzione dei dati considerando 40 classi

Di seguito vengono fatte alcune considerazioni utili per comprendere la distribuzione dei dati in questo contesto.

a. Se l'istogramma tende più a sinistra, significa che compie un errore nel conteggio delle entrate (figura 4.10), nel caso in cui tenda a destra delle uscite (figura 4.11). Il caso ideale si ha quando si trova in corrispondenza del valore medio degli intervalli, ovvero nel centro.

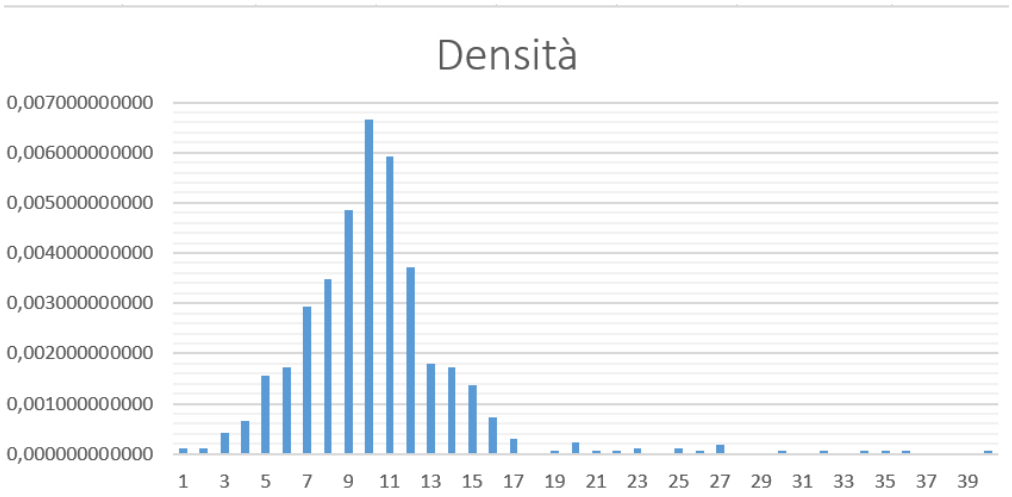

Figura 4.10 – Istogramma che descrive un errore nel conteggio delle entrate

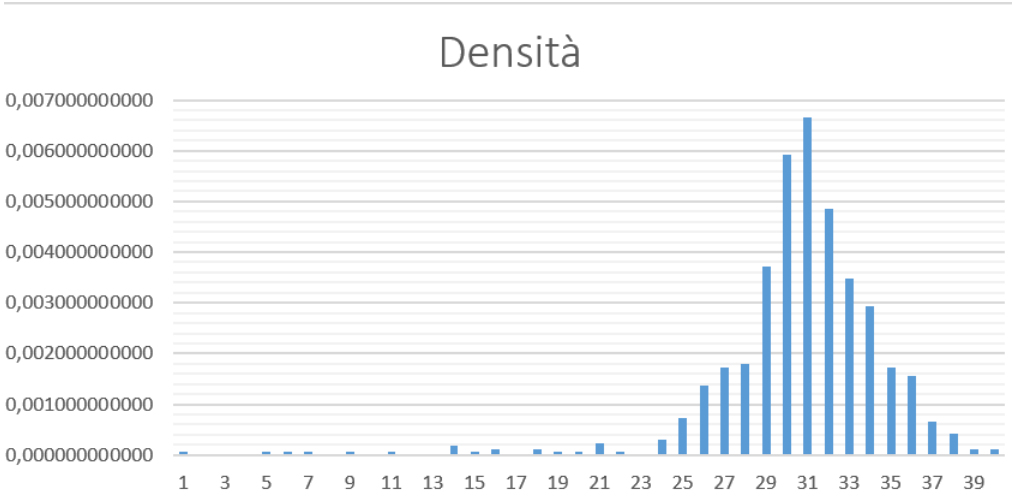

Figura 4.11 – Istogramma che descrive un errore nel conteggio delle uscite

b. Se l'istogramma tende ad una distribuzione normale, ovvero se le variabili tendono a concentrarsi attorno al valore medio, allora l'errore degli ingressi e delle uscite si compensa (figura 4.12).

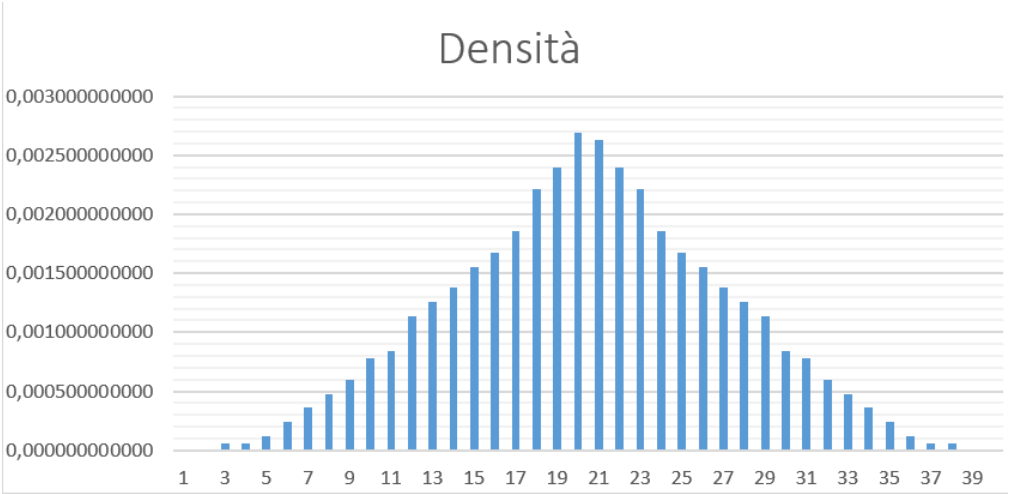

Figura 4.12 – Istogramma che mostra come l'errore tra gli ingressi e le uscite si compensa

c. Se l'istogramma tende verso l'alto si stringe e l'intervallo in cui ricade lo zero è quello contenente più valori (ovvero se la media e la mediana tendono a coincidere), allora si avvicina al caso ideale (figura 4.13). In caso contrario i dati assumono una distribuzione via via più casuale (figura 4.14).

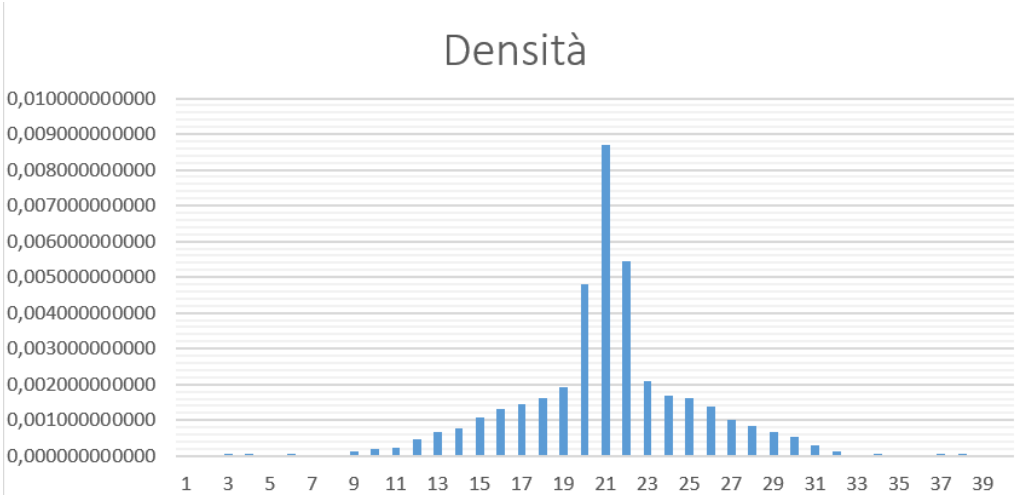

Figura 4.13 – Istogramma che si avvicina al caso ideale

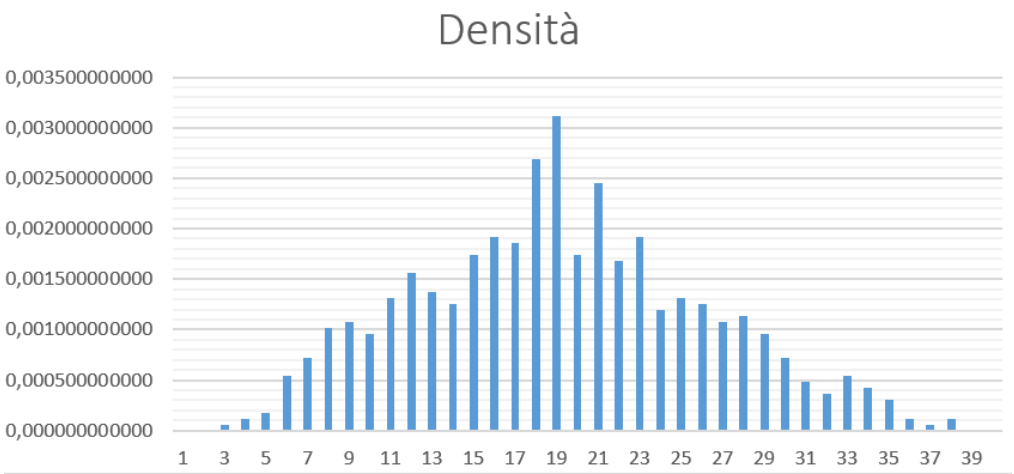

Figura 4.14 – Istogramma descrive una distribuzione casuale

4.6 Modello di correzione

Dopo aver effettuato un'analisi dell'andamento dell'errore e della distribuzione dei dati, è stato possibile elaborare uno strumento per la correzione dell'errore.

Questo modello è stato applicato ad un caso pratico, quello del conteggio dei passaggi attraverso un varco nel mese di maggio, per capire in che modo viene trasformato il dato al fine di renderlo più attendibile.

In primo luogo sono stati riportati sul foglio di calcolo *Maggio16* i dati estratti da *Area id 27* riguardanti il mese di maggio 2016: *date [PK] character varying*, *time [PK] integer*, *area_id [PK] integer*, *direction_id [PK] integer*, *people integer*, *conversione*, *sommatoria persone in negozio (giorno)*, *Δ fine giornata*, *N persone entrate*, *N persone uscite* (tabella 4.15).

date [PK] character varying	time [PK] integer	area_id [PK] integer	direction_id [PK] integer	people integer		conversione	sommatoria persone in negozio (giorno)		Δ fine giornata	N persone entrate	N persone uscite
01/05/2016	8	27	1	46		46	46		46	46	0
01/05/2016	8	27	2	34		-34	12		12	46	-34
01/05/2016	9	27	1	27		27	39		39	73	-34
01/05/2016	9	27	2	22		-22	17		17	73	-56
01/05/2016	10	27	1	33		33	50		50	106	-56
01/05/2016	10	27	2	29		-29	21		21	106	-85
01/05/2016	11	27	1	40		40	61		61	146	-85
01/05/2016	11	27	2	28		-28	33		33	146	-113
01/05/2016	12	27	1	31		31	64		64	177	-113
01/05/2016	12	27	2	38		-38	26		26	177	-151
01/05/2016	13	27	1	15		15	41		41	192	-151
01/05/2016	13	27	2	17		-17	24		24	192	-168
01/05/2016	14	27	1	33		33	57		57	225	-168
01/05/2016	14	27	2	43		-43	14		14	225	-211
01/05/2016	15	27	1	64		64	78		78	289	-211
01/05/2016	15	27	2	44		-44	34		34	289	-255
01/05/2016	16	27	1	10		10	44		44	299	-255
01/05/2016	16	27	2	16		-16	28		28	299	-271

Tabella 4.15 – Estratto dalla tabella Maggio16 per il giorno 1 maggio 2016

Successivamente i dati sono stati filtrati ricavando l'ultima riga per ogni giorno e considerando solo i giorni in cui il punto vendita è stato chiuso dopo le 19 comprese al fine di ottenere solo informazioni rilevanti.

Quindi è stato creato un altro foglio di calcolo, chiamato *Maggio 16_CalcoloDelta e Correz,* dove sono stati riportati i dati selezionati.

Per correggere l'errore, è stato innanzitutto calcolato il valore medio in valore assoluto dell'errore per ogni giorno applicando la seguente formula:

$$\textit{Media } \Delta \textit{ mensile} = ABS\,(AVERAGE\,(\textit{sommatoria persone in negozio (giorno)}_{i} : \textit{sommatoria persone in negozio (giorno)}_{i+n-1}))$$

dove *n* rappresenta il numero di valori appartenenti alla colonna.

Quindi è stato ricalcolato e corretto il numero di persone che hanno attraversato il varco ogni giorno:

= CEILING.MATH ((SUM (N persone entrate $_i$; ABS (N persone uscite $_i$)) + $ media Δ mensile) / 2)

Ricorrendo a questa correzione è possibile eliminare l'errore, affinché il numero di entrate corrisponda a quello delle uscite e che il valore corretto sia compreso tra il valore medio delle entrate e il valore medio delle uscite per ogni mese.

Di seguito viene riportato un esempio applicando le formule sopra descritte al caso del mese di maggio (figura 4.16).

date [PK] character varying	time [PK] integer	people integer		conversione	sommatoria persone in negozio (giorno)		Δ fine giornata	N persone entrate	N persone uscite
02/05/2016	19	3		-3	23		23	227	-204
03/05/2016	20	2		-2	52		52	201	-149
05/05/2016	19	8		-8	-36		36	153	-189
06/05/2016	20	3		-3	-55		55	219	-274
07/05/2016	19	5		-5	21		21	125	-104
09/05/2016	19	4		-4	18		18	172	-154
10/05/2016	19	5		-5	11		11	149	-138
11/05/2016	19	5		-5	-10		10	137	-147
12/05/2016	19	4		-4	4		4	141	-122
13/05/2016	19	8		-8	-15		15	124	-139
14/05/2016	20	2		-2	11		11	138	-127
16/05/2016	19	11		-11	-2		2	172	-174
17/05/2016	19	8		-8	16		16	156	-140
18/05/2016	19	7		-7	1		1	134	-120
19/05/2016	19	15		-15	-26		26	126	-152
20/05/2016	19	10		-10	-20		20	159	-179
21/05/2016	20	1		1	28		28	142	-114
23/05/2016	20	2		2	15		15	134	-102
24/05/2016	20	2		-2	-1		1	144	-145
25/05/2016	19	9		-9	-15		15	140	-155
26/05/2016	19	13		-13	5		5	150	-145
27/05/2016	19	6		-6	33		33	173	-140
28/05/2016	19	5		-5	-1		1	127	-128
30/05/2016	19	13		-13	-5		5	180	-185
31/05/2016	19	14		-14	-29		29	157	-186

Tabella 4.16 – Tabella con i dati del mese di maggio 2016

Il primo passaggio consiste nel calcolare il valore della media dell'errore mensilmente.

$$Media\ \Delta\ mensile = (23 + 52 + (– 36) + (– 55) + 21 + 18 + 11 + (– 10) + 4 + (– 15) + 11 + (– 2) + 16 + 1 + (– 26) + (– 20) + 28 + 15 + (– 1) + (– 15) + 5 + 33 + (– 1) + (– 5) + (– 29)) / 25 = 0.92$$

Per correggere, ad esempio, il valore degli attraversamenti effettuati il 2 maggio 2016, è necessario sommare il numero di persone entrate, con il numero di quelle uscite in valore assoluto e il valore appena ottenuto. Si divide il risultato per due, affinchè si distribuisca in modo uniforme tra gli ingressi e le uscite.

$$N\ persone\ corretto = (227 + ABS (-204) + 0.92) / 2 = 216$$

Per accertarsi che il risultato sia affidabile, è possibile verificare che la media del numero di persone corretto di tutto il mese sia compresa tra il valore medio del numero di persone entrate e quello delle persone uscite in valore assoluto.

- Valore medio N persone entrate = 155,2
- Valore assoluto del valore medio N persone uscite = 152,48
- Valore medio N persone corretto = 154,56

$$152,48 < 154,56 < 155,2$$

Correggendo il numero di persone per il resto del mese si ottengono i seguenti risultati (tabella 4.17):

Δ fine giornata	23	52	36	55	21	18	11	10	4	15	11	2	16	1	26	20	28	15	1	15	5	33	1	5	29
N persone entrate	227	201	153	219	125	172	149	137	141	124	138	172	156	134	126	159	142	134	144	140	150	173	127	180	157
N persone uscite	-204	-149	-189	-274	-104	-154	-138	-147	-122	-139	-127	-174	-140	-120	-152	-179	-114	-102	-145	-155	-145	-140	-128	-185	-186
N persone corretto	216	176	172	247	115	164	144	143	132	132	133	174	149	128	140	170	129	119	145	148	148	157	128	183	172

Tabella 4.17 – Tabella che illustra il risultato ottenuto dopo aver applicato la correzione

Dall'osservazione dell'esito, si conclude che il risultato ottenuto fornisce un valore più preciso del numero di persone che hanno attraversato il varco ogni giorno.

4.7 Considerazioni sullo strumento realizzato

Le metodologie finora descritte possono essere utilizzate per analizzare l'andamento dell'errore nel tempo, la sua distribuzione al fine di verificare quanto il modello sia efficace e se sia necessario applicare delle correzioni all'algoritmo di video analisi e ai parametri della telecamera per migliorarlo.

La possibilità di correggere il valore corrispondente al numero di persone che hanno visitato il punto vendita fornisce invece l'opportunità di applicare il fattore di correzione dell'ultimo mese analizzato a quello attuale, in modo da compensare l'errore giorno per giorno in tempo reale.

Monitorandone l'andamento, si può anche individuare il momento in cui è necessario considerare un nuovo fattore di correzione per pareggiare il numero di ingressi con quello delle uscite o se è possibile mantenere quello attuale.

5 MODELLO PREDITTIVO

In questo capitolo viene delineato l'obiettivo ultimo di questa tesi, ovvero la realizzazione di un modello in grado di fornire una previsione futura sul numero di ingressi in un punto vendita, e come è stato raggiunto.

Nel primo paragrafo viene approfondita la tecnologia utilizzata per il raggiungimento di questo obiettivo: il *Machine Learning*. La piattaforma selezionata per l'elaborazione del modello è *Azure Machine Learning Studio*.

Nel secondo paragrafo viene fornita una spiegazione su come questo progetto sia stato applicato ad un caso pratico.

Il terzo paragrafo tratta della raccolta dei dati e riporta la descrizione delle tre tabelle utilizzate: quella estratta dalla video analisi che indica il numero di persone che sono entrate nel punto vendita, quella contenente i dati relativi ai fattori che influenzano il flusso di visite e quella che riporta i dati del meteo di Roma.

Di seguito, nel quarto paragrafo viene spiegato come è stato progettato il modello, descrivendo i moduli forniti dalla piattaforma per la preparazione, la combinazione, la suddivisione dei dati e l'algoritmo di previsione adoperato.

L'ultimo paragrafo si concentra sulla lettura e interpretazione dei dati ottenuti, sul grado di precisione del risultato e su un esempio di anomalia.

5.1 Machine Learning

5.1.1 Introduzione al Machine Learning

"Si dice che un programma apprenda dall'esperienza E in relazione ad alcune classi di compiti T e alla misura di prestazione P, se le sue prestazioni nei compiti in T, misurate con P, migliorano con l'esperienza E".

Questa è la definizione ufficiale e più moderna di *Machine Learning* elaborata da Tom M. Mitchell che spiega come un programma può migliorare la *performance* di un determinato compito basandosi sull'esperienza passata e imparando da essa.

Riformulando questo concetto in modo più esplicito, si può dire che con *Machine Learning* (o apprendimento automatico) si intende l'abilità di un computer di apprendere in modo autonomo senza essere stato programmato per svolgere determinati compiti. Diversamente dagli esseri umani, i computer non imparano dall'esperienza di vita quotidiana, ma estraendo informazioni dai *dataset* che gli vengono forniti. Una volta raccolto un numero sufficiente di dati, la macchina formula dei modelli e delle regole che li descrivono. Quando la si pone di fronte ad un nuovo problema che rientra in una delle casistiche individuate, lo confronta con quelli precedenti ed elabora una soluzione basandosi su come quella tipologia di problema è stata risolta in passato.

Un sistema di *Machine Learning* si compone di tre elementi principali. È possibile individuare un modello, ovvero il sistema che esegue le predizioni o identificazioni, dei parametri, ossia i segnali o fattori usati dal modello per prendere delle decisioni ed un sistema di apprendimento che corregge il valore dei parametri osservando le differenze tra il risultato previsto e quello effettivo.

In base alla loro tipologia e alla combinazione, è possibile suddividere l'apprendimento automatico in tre categorie:

- Apprendimento supervisionato, ovvero un approccio utilizzato quando il target o il risultato della variabile dipendente è conosciuto a priori ed è predetto da un modello ricorrendo ad una vasta gamma di fattori di predizione che rappresentano le variabili indipendenti;
- Apprendimento non supervisionato, che viene impiegato quando si è incerti su quale dovrebbe essere l'effetto delle variabili sul risultato;
- Apprendimento per rinforzo, che rappresenta l'apprendimento continuo di un computer dall'esperienza passata al fine di prendere decisioni sempre più accurate e specifiche.

L'apprendimento automatico viene sfruttato per risolvere tre tipologie di problemi.

- La prima è la classificazione, che implica l'assegnazione di un elemento ad una categoria rispetto ad un'altra. Ad esempio, può individuare se una email deve essere catalogata come spam o posta ricevuta.
- La seconda tipologia è il raggruppamento, che consiste nell'ordinare i dati in gruppi che presentano caratteristiche comuni, utilizzato principalmente nel marketing per segmentare i propri clienti in gruppi omogenei e individuare mercati potenziali.
- Infine, la terza e ultima tipologia è quella della regressione, che prevede il valore di un elemento mettendo in relazione il suo valore attuale con alcune delle condizioni che influenzano il risultato.

Il modello elaborato in questa tesi di laurea rientra nella prima categoria dell'apprendimento supervisionato, in quanto il suo compito è quello di individuare una relazione tra il numero di visitatori di un negozio e le condizioni in cui questo valore è stato estrapolato, come le condizioni metereologiche, il giorno della settimana, il video promozionale in vetrina, etc...

L'algoritmo utilizzato è di conseguenza una regressione. La piattaforma utilizzata per sviluppare il modello è *Azure Machine Learning Studio*.

5.1.2 Azure Machine Learning Studio

Azure Machine Learning Studio è una soluzione in *cloud* di Microsoft che combina le metodologie e gli strumenti del *Data Mining*, *Data Science* e *Machine Learning*.

Differisce dalla maggior parte degli strumenti di *Data Science* perché non presuppone la conoscenza di uno o più linguaggi di programmazione e lo studio teorico delle tecniche per estrarre informazioni da grandi moli di dati.

Si tratta di una piattaforma collaborativa che viene utilizzata per creare, testare e illustrare modelli analitici predittivi sui dati che gli vengono forniti. I modelli realizzati

vengono pubblicati come servizi web che possono essere successivamente utilizzati per applicazioni personalizzate o da strumenti di *Business Intelligence*.

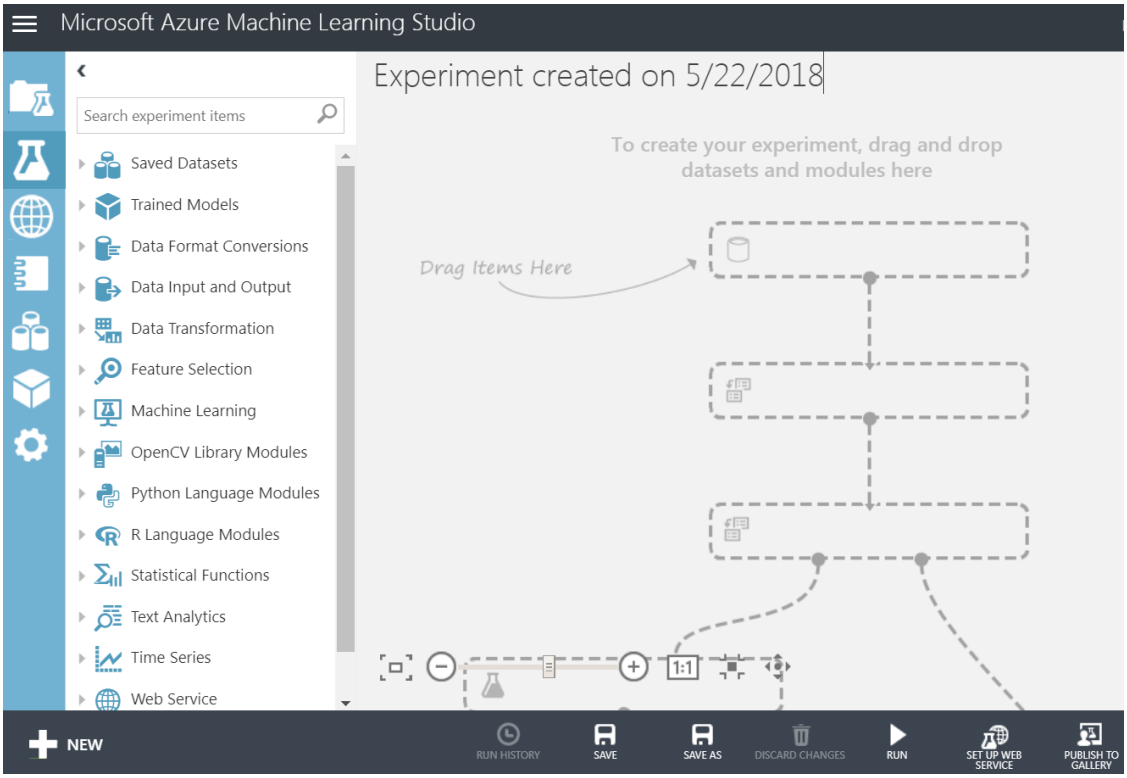

Figura 5.1 – Interfaccia grafica di Azure Machine Learning Studio

Azure Machine Learning Studio offre un set di dati e moduli di analisi da trascinare su un canvas interattivo (figura 5.1), che possono essere collegati tra loro per eseguire un esperimento che viene testato sulla piattaforma stessa. Ha un'interfaccia visuale e assomiglia più a *Microsoft Visio* rispetto agli strumenti tradizionali di *Machine Learning*.

Un esperimento può essere modificato un numero infinito di volte e per testarlo basta semplicemente programmarlo di nuovo. Quando il risultato ottenuto è soddisfacente, è possibile convertire l'esperimento di *training* in esperimento predittivo e pubblicarlo come servizio web affinchè la soluzione possa essere utilizzata anche da terzi.

Pur non implicando competenze di programmazione, *Azure Machine Learning Studio* permette di personalizzare i moduli ricorrendo a linguaggi quali R, Python o SQL.

5.1.2.1 Criteri di scelta della piattaforma

Le ragioni per cui è stata selezionata questa piattaforma per le sperimentazioni sulla predizione del numero di visitatori sono molteplici.

Innanzitutto l'ambiente grafico di *Azure Machine Learning Studio* è molto intuitivo e grazie ad una guida molto esaustiva e a numerosi tutorials è possibile imparare ad utilizzare questa tecnologia autonomamente e in tempi brevi.

Il servizio per analisi di *Big Data* è in *cloud* e di conseguenza non richiede l'utilizzo di un computer molto potente poichè viene processato in uno dei centri di elaborazione dati di Microsoft.

Si tratta di una piattaforma gratuita e il costo di archiviazione dei dati è molto basso. Questo ha permesso lo studio di un elevato numero di dati appartenenti ad una o più tabelle contemporaneamente e la formulazione di numerose versioni del modello con l'obiettivo di arrivare alla soluzione migliore senza dover fare alcun investimento sulla ricerca.

In aggiunta, le integrazioni possibili con infrastrutture IT sono numerose. È possibile integrare la piattaforma con *Azure SQL, HDInsight/Hadoop/Spark, Azure Data Factory, Microsoft Excel*, etc. Ciò ha reso possibile sfruttare i dati estrapolati dai sistemi di *video intelligence* e quelli presenti su altri *database* per effettuarne l'analisi.

Questa piattaforma offre anche la possibilità di applicare più algoritmi allo stesso esperimento e confrontarne i risultati contemporaneamente.

Infine, trattandosi di uno strumento collaborativo, è possibile lavorare sullo stesso modello con altri utenti e, una volta ultimato, condividerlo. In questo modo questo lavoro potrà essere utilizzato e rielaborato da chiunque ne abbia bisogno al fine di adattarlo alle proprie necessità.

5.2 Scenario

Per svolgere questo studio, sono stati presi in analisi i dati ricavati da un caso reale, nello specifico quelli estrapolati dal monitoraggio di un negozio di telefonia a Roma da aprile 2016 ad aprile 2018.

Al fine di verificare l'efficienza del modello, è stato ipotizzato di compiere l'esperimento negli ultimi giorni di gennaio 2017, avendo a disposizione le registrazioni degli ultimi quattro mesi e volendo prevedere il numero di visitatori per il mese di febbraio 2017.

In primo luogo, l'attuazione di questo progetto prevede l'installazione di una telecamera in prossimità dell'ingresso, con particolare riguardo alle condizioni di luce e all'inclinazione di ripresa della telecamera.

Quindi, è necessario applicare un algoritmo di video analisi, nel caso di questo studio quello del *Gateflow* descritto in precedenza, al fine di estrapolare le informazioni sull'attraversamento del varco.

Contemporaneamente vengono raccolti i dati riguardanti alle condizioni deterministiche e aleatorie in cui avviene la registrazione, ovvero la variazione dei fattori che influiscono sul numero di visitatori del punto vendita e le condizioni meteo della città.

Analizzando la correlazione tra i dati estrapolati dalla video analisi e questi ultimi, è possibile elaborare una previsione dell'affluenza di persone e osservare come il flusso varia al variare delle condizioni iniziali.

5.3 Raccolta dati

Le tabelle prese in analisi da cui sono state estratte le informazioni per effettuare la predizione degli ingressi sono *Inout*, *Negozio* e *Meteo_roma*.

5.3.1 Inout

La tabella *Inout* (tabella 5.2) è quella precedentemente utilizzata per la correzione dell'errore dell'algoritmo di video analisi nel conteggio degli ingressi e delle uscite.

date [PK] character varying	time [PK] integer	area_id [PK] integer	direction_id [PK] integer	people integer	giorno settimana	Mese	conversione	sommatoria persone in negozio (giorno)	N persone entrate	N persone uscite	N persone corretto
13/09/2016	20	27	2	2	3	9	-2	-37	349	-386	387
14/09/2016	19	27	2	28	4	9	-28	-25	291	-316	323
15/09/2016	19	27	2	16	5	9	-16	-3	286	-275	300
16/09/2016	20	27	2	2	6	9	-2	4	249	-245	266
17/09/2016	20	27	2	3	7	9	-3	-50	332	-382	376
19/09/2016	19	27	2	39	2	9	-39	-114	368	-482	444
20/09/2016	19	27	2	27	3	9	-27	-72	318	-390	373
21/09/2016	20	27	2	2	4	9	-2	-75	273	-348	330
22/09/2016	19	27	2	9	5	9	-9	-36	332	-368	369
23/09/2016	19	27	2	20	6	9	-20	-39	306	-345	345
24/09/2016	20	27	2	3	7	9	-3	-47	269	-316	312
26/09/2016	20	27	2	2	2	9	-2	-47	367	-414	410
27/09/2016	19	27	2	25	3	9	-25	-8	327	-335	350
28/09/2016	19	27	2	40	4	9	-40	-56	341	-397	388
29/09/2016	20	27	2	2	5	9	-2	2	275	-273	293
30/09/2016	20	27	2	2	6	9	-2	-78	373	-451	431
01/10/2016	20	27	2	3	7	10	-3	-54	329	-383	385
03/10/2016	19	27	2	27	2	10	-27	-104	341	-445	422

Tabella 5.2 – Tabella Inout prima di essere filtrata

I campi originali estratti dal sistema e quelli ricavati correggere l'errore sono i seguenti:

- *date [PK] character varying;*
- *time [PK] integer;*
- *area_id [PK] integer;*
- *direction_id [PK] integer;*
- *people integer;*
- *conversione;*
- *sommatoria persone in negozio (giorno);*
- *N persone entrate;*
- *N persone uscite.*

Dopo aver applicato la correzione, è stato ricalcolato il numero di ingressi per ogni giorno e i valori sono stati riportati in una colonna denominata *N persone corretto*.

In seguito sono stati ricavati il giorno della settimana ed il mese con le funzioni WEEKDAY e MONTH su Excel.

Il campo *giorno settimana* contiene i valori da 1 a 7 corrispondenti al giorno della settimana, considerando che la domenica corrisponde al numero 1, il lunedì al 2, etc. Questa colonna, insieme a quella *mese* serviranno per essere rielaborate dal modello di *Machine Learning* al fine di ricavare delle informazioni che influenzano la previsione.

I dati individuati come rilevanti studiati dal modello sono i seguenti:

- *date [PK] character varying*: indica il giorno in cui è avvenuta la rilevazione dei dati. Dopo avere analizzato i dati con lo strumento di monitoraggio dell'andamento dell'errore illustrato precedentemente, è stato deciso di considerare un arco temporale corrispondente al periodo compreso tra il 13 settembre 2016 e il 13 febbraio 2017 in quanto i risultati sono considerati affidabili e la tendenza regolare e uniforme.
- *giorno settimana:* questi valori sono stati ricavati per essere categorizzati dal modello di *Machine Learning* come giorno feriale o giorno festivo. Si è notato infatti un cambiamento sostanziale nel numero di visitatori nei due periodi distinti. Nel weekend il numero di visitatori ha la tendenza a ridursi.
- *mese*: come per il giorno della settimana, questi valori sono stati ricavati per essere suddivisi in base alla stagione, dal momento che, assieme alle promozioni periodiche, si tratta di un fattore che influenza il numero di visitatori del punto vendita.
- *N persone corretto*: rappresenta il numero di persone che sono entrate e uscite da un negozio ogni giorno dopo essere stato corretto secondo i criteri descritti nel capitolo sulla correzione dell'errore.

5.3.2 Negozio

La tabella *NegozioTelefonia* (tabella 5.3) contiene informazioni riguardanti la tipologia di pubblicità che compare sullo schermo in vetrina, l'addetto alle vendite e la categoria in cui ricade lo spot televisivo in relazione alla data.

Data	Video Vetrina	Commesso	Spot televisivo
13/09/2016	1	Susanna	Standard
14/09/2016	1	Alessandro	Short
15/09/2016	3	Alessandro	Short
16/09/2016	3	Alessandro	Short
17/09/2016	1	Susanna	Standard
19/09/2016	1	Susanna	Super promo
20/09/2016	1	Susanna	Standard
21/09/2016	1	Alessandro	Short
22/09/2016	1	Susanna	Standard
23/09/2016	1	Alessandro	Short
24/09/2016	3	Alessandro	Short
26/09/2016	1	Susanna	Standard
27/09/2016	1	Claudia	Standard
28/09/2016	1	Susanna	Standard
29/09/2016	3	Alessandro	Short
30/09/2016	1	Susanna	Super promo
01/10/2016	1	Susanna	Standard
03/10/2016	1	Susanna	Super promo

Tabella 5.3 – Estratto dalla tabella che indica la combinazione di video proiettato in vetrina, addetto vendite e spot televisivo

I campi della tabella sono:

- *data*: questa informazione è rilevante dal momento in cui identifica il giorno in cui si è verificata una combinazione di condizioni e viene utilizzata come chiave primaria per essere correlata alla tabella *Inout*.
- *Video vetrina*: indica la tipologia di video visualizzato sullo schermo posto in vetrina. La tipologia dello spot pubblicitario può essere: promozionale (corrispondente al numero 1 in tabella), *story telling* (corrispondente al numero 2) o emozionale (corrispondente al numero 3).
- *Addetto vendite:* precisa quale commesso ha lavorato in quel giorno. Questo dato è rilevante in quanto i clienti abituali preferiscono parlare con un operatore che conosca le loro abitudini e li sappia supportare.
- *Spot televisivo:* distingue tre categorie di spot: quello *short* che dura solo qualche secondo*, quello standard* e il *super promo* che rappresenta le offerte che si protraggono per periodi molto limitati.

5.3.3 Meteo

I dati sul meteo della città di Roma (tabella 5.4) sono stati scaricati dall'archivio storico del sito www.ilmeteo.it.

Giorno	T Media	T min	T max	Precip	Umid	Vento Max	Raffica	Fenomeni
01/04/2016	20 °C	10 °C	22 °C	-	49%	20 km/h	-	Nessuno
02/04/2016	18 °C	9 °C	24 °C	-	63%	13 km/h	-	Nessuno
03/04/2016	19 °C	9 °C	24 °C	-	60%	13 km/h	-	Nessuno
04/04/2016	20 °C	9 °C	24 °C	-	55%	17 km/h	-	Nessuno
05/04/2016	20 °C	9 °C	25 °C	-	58%	15 km/h	-	Nessuno
06/04/2016	21 °C	9 °C	27 °C	-	53%	19 km/h	-	Nessuno
07/04/2016	21 °C	10 °C	27 °C	-	54%	17 km/h	-	Nessuno
08/04/2016	18 °C	14 °C	20 °C	-	69%	11 km/h	-	Nessuno
09/04/2016	15 °C	14 °C	16 °C	n/d	63%	20 km/h	-	Pioggia
10/04/2016	16 °C	7 °C	20 °C	-	63%	19 km/h	-	Nessuno
11/04/2016	17 °C	7 °C	21 °C	-	61%	19 km/h	-	Nessuno
12/04/2016								
13/04/2016	21 °C	14 °C	25 °C	-	47%	41 km/h	50 km/h	Nessuno
14/04/2016	21 °C	15 °C	24 °C	-	51%	15 km/h	-	Nessuno
15/04/2016	18 °C	7 °C	22 °C	-	54%	19 km/h	-	Nessuno
16/04/2016	20 °C	9 °C	24 °C	-	46%	39 km/h	46 km/h	Nessuno
17/04/2016	23 °C	12 °C	26 °C	-	35%	24 km/h	39 km/h	Nessuno
18/04/2016	20 °C	13 °C	23 °C	-	55%	13 km/h	-	Nessuno
19/04/2016	18 °C	10 °C	22 °C	-	58%	26 km/h	-	Nessuno
20/04/2016	20 °C	12 °C	25 °C	-	41%	20 km/h	-	Nessuno
21/04/2016	18 °C	8 °C	24 °C	-	52%	17 km/h	-	Nessuno

Tabella 5.4 – Estratto sul meteo di Roma

I campi della tabella sono:

- *Giorno*: sono state considerate le condizioni metereologiche comprese tra il 13 settembre 2016 e il 13 febbraio 2017;
- *Temperatura media, massima e minima;*
- *Precipitazioni;*
- *Umidità;*
- *Velocità massima del vento;*
- *Raffica;*
- *Fenomeni*: suddivide le condizioni metereologiche in quattro categorie. Quando compare il termine "Nessuno" si indica un clima sereno, mentre le altre condizioni sono espresse attraverso l'uso di simboli che, prima di essere elaborati, sono stati convertiti in parole.

Le immagini e simboleggiano una giornata di “Pioggia”, di “Nebbia” o di “Neve”.

Le colonne considerate rilevanti per l’analisi sono:

- *Giorno*: indica la data ed è utilizzata come chiave primaria per mettere in relazione questa tabella con le altre due;
- *Fenomeni*: le condizioni metereologiche sono considerate importanti in questo contesto in quanto possono influenzare il comportamento delle persone. Analizzando i dati è possibile infatti distinguere un aumento di visitatori nei giorni in cui il clima è sereno.

5.4 Struttura del modello

In questo paragrafo vengono illustrati i passaggi che hanno portato alla realizzazione del modello predittivo, descrivendo i moduli utilizzati, la preparazione dei dati e l’algoritmo applicato.

La rappresentazione grafica in figura (figura 5.5) mostra come è strutturato il modello sul canvas di Azure Machine Learning Studio.

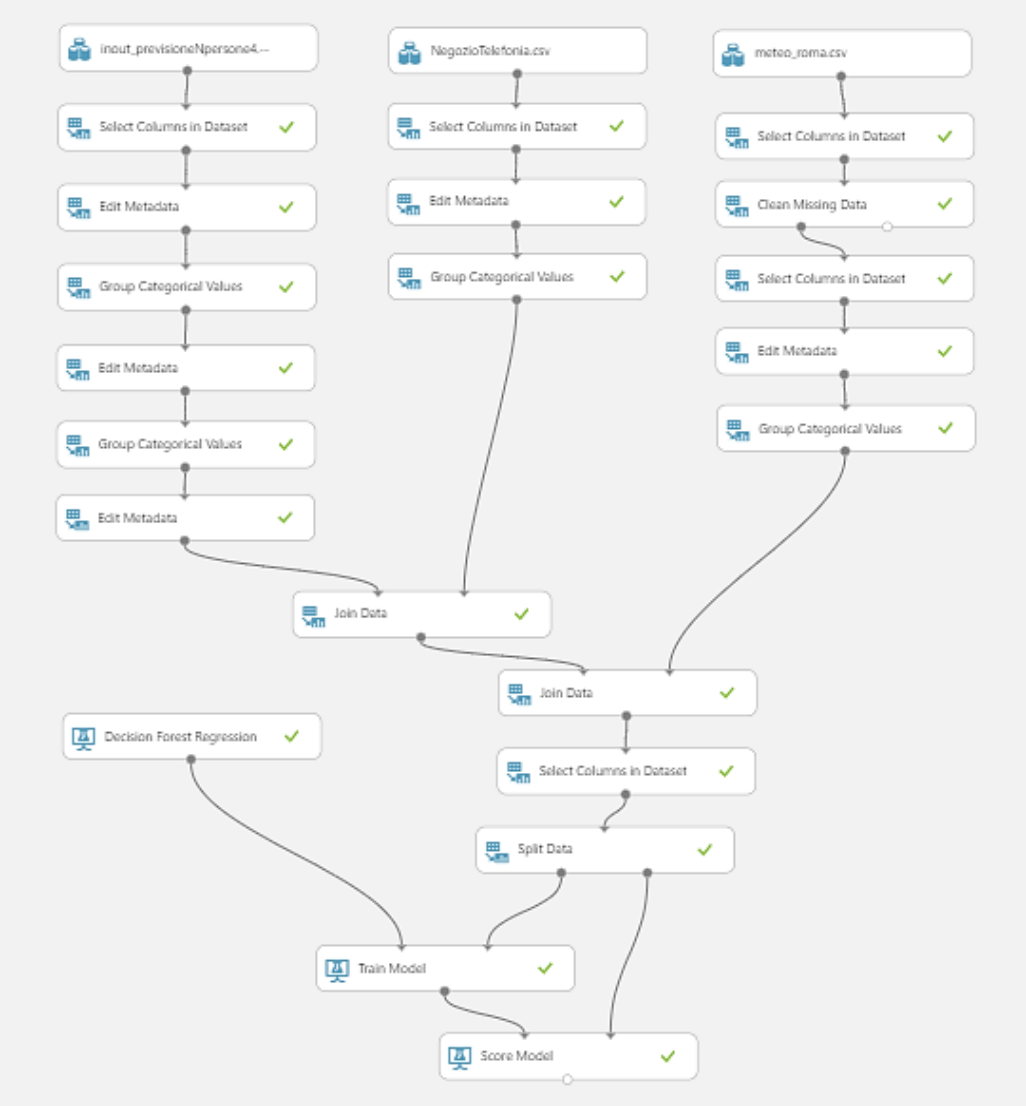

Figura 5.5 – Modello creato su Azure Machine Learning Studio

5.4.1 Conversione files

Prima di essere elaborate dal modello di *Machine Learning*, tutte le tabelle devono essere convertite in file csv. Il simbolo che identifica le colonne è convertito con un editor testuale in virgole in modo che *Azure Machine Learning Studio* sia in grado di distinguerle.

5.4.2 Preparazione dei dati

In primo luogo è necessario importare le tre tabelle come datasets e trascinarle nel canvas, il piano di lavoro della piattaforma.

Prima di essere analizzati dall'algoritmo, i dati devono essere ordinati, eseguendo tre passaggi fondamentali:

- Selezione delle colonne da esaminare;
- Trasformazione dei dati in informazioni rilevanti.
- Pulizia delle righe con dati mancanti;

5.4.2.1 Selezione delle colonne da esaminare

Per selezionare i campi contenenti elementi significativi, si ricorre al modulo *Select Columns in Dataset*.

Questo modulo permette di creare un sottoinsieme di campi includendo o escludendo le colonne. È possibile distinguerle in base al nome, all'indice della colonna, ovvero al modo in cui sono ordinate all'interno della tabella, alla tipologia del dato (stringa, valore numerico, intero o valore mobile) o alla combinazione di questi criteri.

Ad esempio, nel caso della tabella *inout_previsione*, sono state utilizzate le impostazioni in figura per selezionare i campi: *date [PK] character varying, N persone corretto, giorno settimana, mese* (figura 5.6).

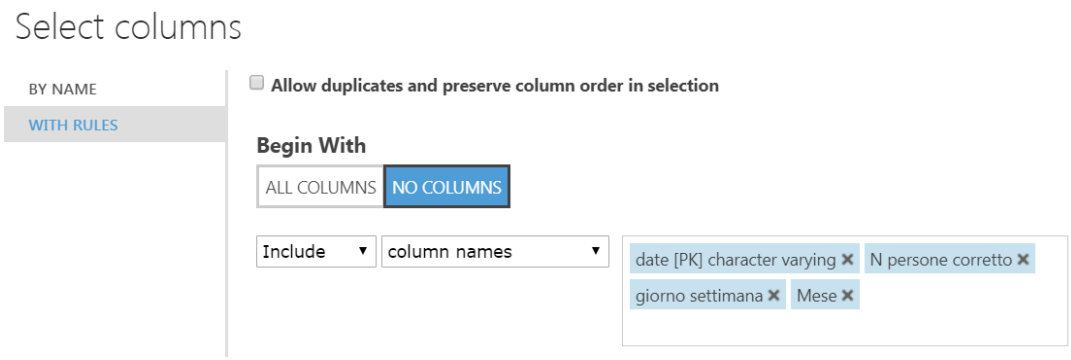

Figura 5.6 – Impostazioni del modulo Select Columns in Dataset per la tabella inout_previsione

5.4.2.2 Trasformazione dei dati in informazioni rilevanti

Per indicare quali colonne devono essere categorizzate o unificate, per trasformare valori contenenti una data o un orario in numerico, per rinominare una colonna si ricorre alla combinazione dei moduli *Edit Metadata* e *Group Categorical Values*.

Per il modulo *Edit Metadata* è necessario individuare il campo o il gruppo di campi su cui si vuole lavorare, come nel caso del modulo descritto precedentemente.

Poi, qualora fosse necessario trasformare la tipologia di dato, è possibile convertirla in stringa, valore numerico intero o mobile, variabile booleana, temporale o data. Ad esempio, se il dataset di partenza contiene valori numerici riportati come testo, è possibile convertirli prima di utilizzarli per operazioni matematiche.

Nel caso in cui i valori contenuti nella colonna siano ripetitivi, possono essere trasformati in categorie.

Il modulo *Group Categorical Values* permette di raggruppare i valori in categorie e rinominarli.

Dopo aver selezionato le colonne, è possibile decidere se il nuovo valore sostituirà o affiancherà quello originale. Poi si considera il numero di livelli per indicare il numero di classi in cui vengono raggruppati i valori, di cui uno di default per valori soggetti a errore.

Un esempio di categorizzazione dei campi è quella eseguita per la colonna *giorno settimana* della tabella *inout_previsione*. In questo caso i giorni della settimana sono stati convertiti in “feriale” o “festivo” poiché si è osservato che il numero di visitatori ha un trend negativo durante il weekend.

La colonna *giorno settimana* è stata impostata come una stringa e sono stati specificati che i valori devono essere trattati come categorie e non devono essere utilizzati per calcoli matematici. Quindi sono stati considerati otto livelli, uno per ogni giorno della settimana più uno di default. I giorni dal lunedì (2) al venerdì (6) sono stati trasformati in “feriale”, mentre il sabato (7) e la domenica (1) in “festivo” (figura 5.7).

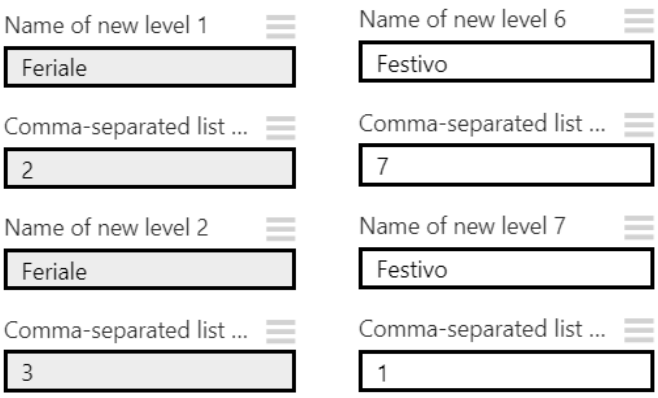

Figura 5.7 – Impostazioni del modulo Group Categorical Values: trasformazione del lunedì e del martedì in feriale, del sabato e della domenica in festivo

Un esempio di rinominazione della colonna è invece quella del campo *mese* sempre appartenente alla medesima tabella.

Per rendere coerente il nome della colonna con il contenuto, è stato convertito in *stagione*, mantenendo la tipologia del dato e la categorizzazione invariate (figura 5.8).

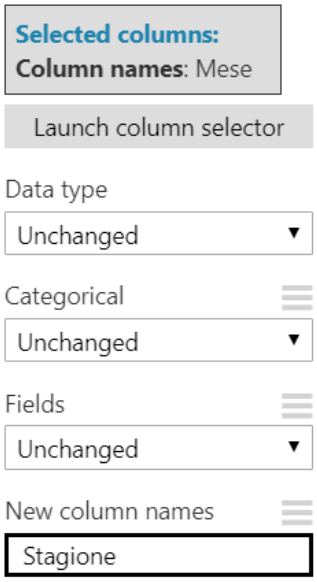

Figura 5.8 – Impostazioni del modulo Edit Metadata per cambiare il nome alla colonna mese in stagione

5.4.2.3 Pulizia delle righe con dati mancanti

Talvolta è necessario dedurre il valore o eliminare quelle righe che non contengono un numero sufficiente di informazioni per prevenire problemi causati dalla mancanza di dati durante il training di un modello. A questo scopo si utilizza il modulo *Clean Missing Data* (figura 5.9).

In primo luogo è necessario selezionare le colonne su cui si vuole lavorare. Poi si specifica il minimo ed il massimo numero di valori mancanti affinché l'operazione venga eseguita. Infine si seleziona la modalità con cui avverrà la correzione. È possibile ripulire le tabelle applicando diverse operazioni:

- Eliminare righe o colonne con valori mancanti;
- Sostituire i valori mancanti con un valore simbolico, la media, la mediana, la moda o valori inferenziali basati su metodi statistici.

Questo modulo è stato applicato, ad esempio, alla tabella *meteo_roma*, con l'obiettivo di eliminare tutte le righe qualora mancasse anche solo un valore.

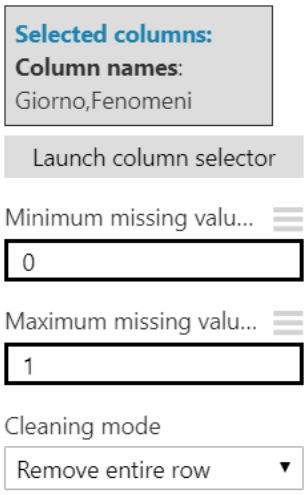

Figura 5.9 – Impostazioni del modulo Clean Missing data per eliminare le righe con elementi mancanti della tabella meteo_roma

5.4.3 Combinazione dei dati

Dopo aver organizzato i dati delle tre tabelle, è necessario combinarle tra loro utilizzando il modulo *Join Data*.

Per unire le tabelle con la funzione *join*, è necessario che una colonna per tabella rappresenti la chiave primaria, ovvero l'elemento che permette di individuare una relazione univoca tra le righe delle due colonne. Una volta individuate le chiavi primarie, si specifica quale tipologia di *join* deve essere impiegato.

Nel caso di questo modello, è stato utilizzato un *inner join*, che restituisce i valori delle righe combinate solo quando le colonne impostate come chiave primaria corrispondono.

La chiave della tabella *inout_previsione* è rappresentata dalla colonna *date [PK] character varying,* la chiave di *NegozioTelefonia* è *Data*, quella di *meteo_roma* è *Giorno*. Tutte le colonne indicano la data come giorno/mese/anno in cui si sono verificate determinate condizioni.

5.4.4 Suddivisione dei dati per il training e la predizione

Una volta raccolte in un'unica tabella, prima che vengano elaborate dall'algoritmo, le informazioni devono essere suddivise in due gruppi distinti. Il primo viene utilizzato per l'apprendimento, mentre il secondo per l'applicazione del modello. A questo proposito si ricorre al modulo *Split Data***.**

Con questo modulo è possibile impostare il modo in cui i dati vengono divisi e la percentuale di dati per ogni categoria.

Per quanto riguarda il modello di predizione, è stato deciso di dividere la tabella in base alle righe e di sfruttare il primo 75% dei dati per l'apprendimento e il restante 25% per l'applicazione del modello.

5.4.5 Algoritmo di previsione

L'algoritmo utilizzato in questo studio è un regressore di foresta decisionale, ovvero un algoritmo di regressione che classifica i dati creando molteplici alberi decisionali e considerando la media di ogni distinta combinazione al fine di ottenere un risultato attendibile.

Tuttavia, a differenza degli alberi decisionali che suddividono i dati in base a criteri precisi, l'algoritmo di foresta decisionale seleziona casualmente le caratteristiche estrapolate per costruire gli alberi decisionali.

Su *Azure Machine Learning Studio*, è possibile utilizzare il modulo denominato *Decision Forest Regression* per elaborare i dati.

Innanzitutto, è necessario decidere se l'analisi si deve basare sul metodo del *bagging* o sulla replicazione.

Con *bagging* (o aggregazione di *bootstrap*), si intende l'elaborazione di una distribuzione gaussiana come risultato della combinazione di più algoritmi di apprendimento al fine di ridurre la varianza del risultato e fare una previsione più accurata.

La replicazione invece prevede l'apprendimento sugli stessi dati di partenza per ogni albero decisionale. La determinazione di quale predicato sia utilizzato per dividere ciascun nodo dell'albero è casuale.

Successivamente è necessario definire la metodologia di apprendimento, che si può basare sull'inserimento di un unico valore per ogni parametro (numero di alberi decisionali, massima profondità per ogni albero, numero di divisioni randomiche per ogni nodo, dimensione minima del campione per ogni foglia) nel caso in cui si abbia la certezza di come si vuole configurare il modello, altrimenti su un *range* di valori se non si è sicuri dei parametri migliori per trovare la soluzione ottimale.

Relativamente a questo modello le impostazioni sono illustrate in figura (figura 5.10).

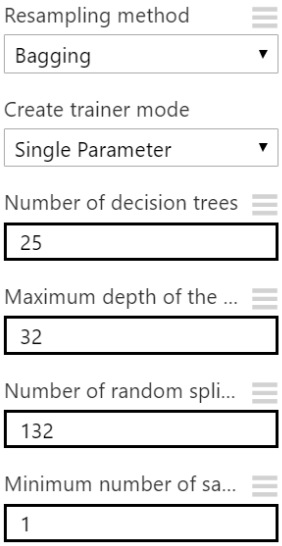

Figura 5.10 – Impostazioni del modulo Decision Forest Regression

Dopo aver inserito il modulo corrispondente all'algoritmo che si vuole adoperare, è necessario ricorrere al modulo *Train Model* per addestrare il modello.

Si seleziona quindi l'elemento che si vuole prevedere sulla base delle combinazioni dei dati forniti, nel caso di questo studio il numero di persone corretto, e si combina con il modulo dell'algoritmo (*Decision Forest Regression*) e con quello che suddivide i dati per l'apprendimento e la predizione (*Split Data*).

Quello che ne risulta è un numero preciso di alberi (figura 5.11) corrispondente a quello indicato come valore del parametro nel modulo precedente.

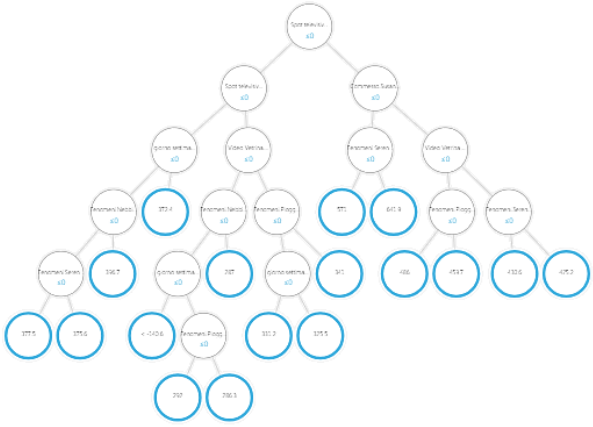

Figura 5.11 – Esempio di albero creato dal regressore di foresta decisionale

5.5 Lettura e interpretazione dei dati

Una volta che il modello è stato sottoposto ad apprendimento, si estrapolano i risultati utilizzando il modulo *Score Model*.

Nel caso degli algoritmi di classificazione, questo modulo fornisce la probabilità di ottenere un determinato risultato e la classe in cui rientra, mentre, per la classificazione di immagini, un valore booleano.

Per gli algoritmi di regressione invece, genera una previsione sotto forma di valore numerico, come illustrato in figura 5.12.

giorno settimana	N persone corretto	Video Vetrina	Commesso	Spot televisivo	Fenomeni	Scored Label Mean
Feriale	499	Story Telling	Susanna	Super promo	Sereno	491.178526
Feriale	481	Story Telling	Susanna	Super promo	Pioggia	461.620833
Feriale	394	Promozionale	Susanna	Standard	Sereno	385.795753
Feriale	454	Story Telling	Susanna	Super promo	Sereno	491.178526
Festivo	323	Promozionale	Alessandro	Short	Sereno	329.839286
Feriale	396	Promozionale	Susanna	Standard	Sereno	385.795753
Feriale	415	Promozionale	Susanna	Standard	Sereno	385.795753

Figura 5.12 – Estratto dalla tabella creata automaticamente dal modulo Score Model che mostra i risultati di predizione

L'esito della predizione per il mese di febbraio 2017 compare nella colonna *Scored Label Mean*. È stata mantenuta la colonna *N persone corretto* corrispondente al valore reale di visitatori nel punto vendita come *benchmarking* con lo scopo di verificare l'efficacia del modello.

Per verificare l'accuratezza del risultato, sono stati confrontati il valore atteso e il valore ottenuto (figura 5.13).

Si nota che il modello realizzato con il *Machine Learning* riesce a predire il numero di persone che si recheranno nel negozio con un grado di precisione superiore al 90% nel 70% dei casi.

Ad esempio, il quinto giorno di un mese invernale (in questo caso febbraio), in condizioni di bel tempo, durante il weekend, proiettando uno spot promozionale in vetrina, con l'addetto vendite di nome Alessandro presente in negozio e una maggioranza di spot appartenenti alla tipologia *short* in televisione, il modello prevede un'affluenza di 330 persone, contro le 323 realmente entrate in negozio.

Quindi, il risultato ottenuto ha un grado di precisione del 98%.

Giorno settimana	N persone reale	Video Vetrina	Commesso	Spot Televisivo	Fenomeni	N persone predetto	Δ	Percentuale di errore	Grado di precisione
Feriale	499	Story Telling	Susanna	Super Promo	Sereno	494,79112	4,208879	1%	99%
Feriale	481	Story Telling	Susanna	Super Promo	Pioggia	459,97833	21,02167	4%	96%
Feriale	394	Promozionale	Susanna	Standard	Sereno	395,84520	1,845198	0%	100%
Feriale	454	Story Telling	Susanna	Super Promo	Sereno	494,79112	40,79112	9%	91%
Festivo	323	Promozionale	Alessandro	Short	Sereno	330,84124	7,841238	2%	98%
Feriale	396	Promozionale	Susanna	Standard	Sereno	395,84520	0,154802	0%	100%
Feriale	415	Promozionale	Susanna	Standard	Sereno	395,84520	19,1548	5%	95%
Feriale	434	Promozionale	Susanna	Super Promo	Sereno	431,21933	2,780667	1%	99%
Feriale	369	Promozionale	Susanna	Standard	Sereno	395,84520	26,8452	7%	93%
Feriale	328	Promozionale	Alessandro	Short	Pioggia	339,45948	11,45948	3%	97%
Festivo	393	Promozionale	Susanna	Standard	Sereno	379,81353	13,18647	3%	97%
Feriale	424	Promozionale	Susanna	Super Promo	Sereno	431,21933	7,219333	2%	98%
Feriale	289	Emozionale	Alessandro	Short	Sereno	250,10100	38,899	13%	87%

Tabella 5.13 – Estratto dalla tabella utilizzata per calcolare il grado di precisione dei risultati di predizione

È possibile consultare la tabella in versione integrale nell'appendice. Per facilitare la lettura dei 30 risultati di previsione, i dati sono stati suddivisi in 5 categorie in base al grado di precisione (tabella 5.14).

Categoria in cui rientra il grado di precisione	Numero di previsioni	Numero di previsioni (percentuale)
0 – 50%	2	7%
51 – 70%	3	10%
71 – 90%	4	13%

91 – 95%	6	20%
96 – 110%	15	50%

Tabella 5.14 – Numero di previsioni per categoria

Nel 70% dei casi il modello è stato in grado di prevedere il numero di visitatori futuri con una precisione pari o superiore al 91% (figura 5.15). Quindi è possibile affermare che i parametri che sono stati individuati per essere analizzati dal modello di *Machine Learning* sono significativi.

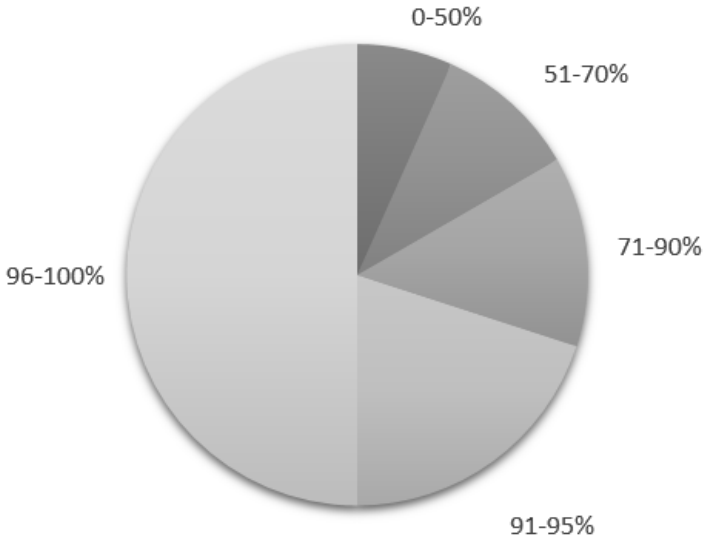

Figura 5.15 – Grafico che rappresenta la suddivisione dei risultati in base alla categoria corrispondente

Essendo i risultati della previsione presentati in ordine cronologico, è stato possibile studiare l'andamento del grado di precisione nel tempo (figura 5.16).

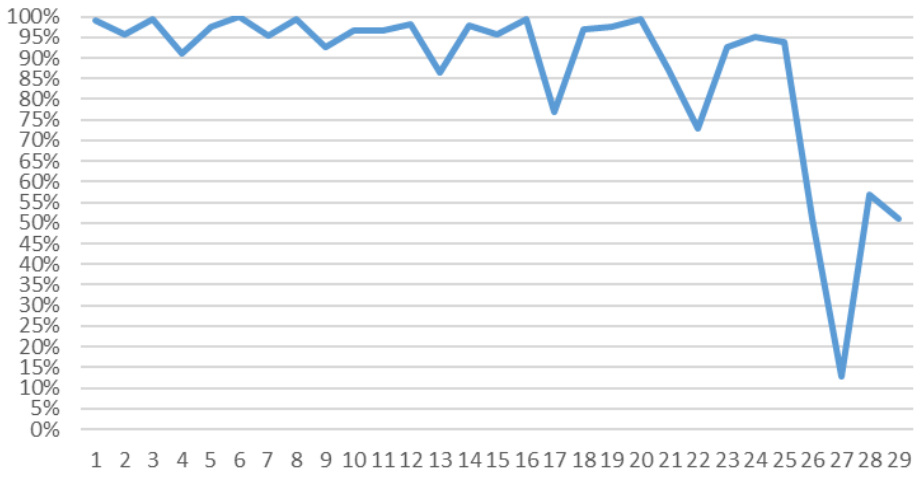

Figura 5.16 – Andamento del grado di precisione nel mese di febbraio 2017

Si nota che, nella prima metà del mese, gli output del modello si sono avvicinati molto a quelli reali, mentre nella seconda metà gli episodi di scostamento sono stati più frequenti. Ciò significa che in quelle occasioni i parametri considerati come input non sono stati sempre sufficienti a prevedere con precisione il traffico di persone giornaliero e che un altro o più fattori non individuati o una condizione anomala, hanno alterato il risultato.

Infatti, a meno che non vengano forniti al modello un'elevata quantità di dati e un maggior numero di parametri, non è possibile prevedere con precisione il flusso di visitatori.

Prendendo in esame i risultati con un grado di precisione inferiore alla media, è stato possibile individuare i fattori più ricorrenti quando il risultato è meno accurato.

La combinazione della tipologia *short* dello spot televisivo con la proiezione di un video emozionale in vetrina e Claudia come addetta presente in negozio corrisponde a due giorni con un grado di precisione basso e ad un'anomalia (figura 5.17).

Giorno settimana	N persone reale	Video Vetrina	Commesso	Spot Televisivo	Fenomeni	N persone predetto	Δ	Percentuale di errore	Grado di precisione
Festivo	15	Emozionale	Claudia	Short	Pioggia	149,80927	134,80927	899%	-799%
Festivo	113	Emozionale	Claudia	Short	Sereno	211,64800	98,648	87%	13%
Feriale	151	Emozionale	Claudia	Short	Sereno	224,83267	73,832667	49%	51%

Tabella 5.17 – Estratto della tabella che illustra i risultati della predizione con un grado di precisione basso

Il modello è in grado di rilevare che la combinazione di determinate condizioni comporta un numero di visitatori minore rispetto a quello medio totale, ma non è in grado di prevedere con esattezza un valore così basso.

Quindi, si può affermare che le configurazioni dei parametri analizzati dal modello in alcune circostanze non sono sempre sufficienti a descrivere il fenomeno.

A differenza dei due casi per i quali non sono stati considerati un numero sufficiente di fattori, è quasi impossibile individuare per quale motivo si verifica un'anomalia.

In questo studio ne è stata rilevata una, quella corrispondente alla previsione che ha un grado di precisione pari a -799% (figura 5.18).

Festivo	15	Emozionale	Claudia	Short	Pioggia	174.60119

Tabella 5.18 – Estratto dalla tabella creata automaticamente dal modulo Score Model che mostra i risultati di predizione

Si può supporre che la ragione di questo risultato possa essere la difficoltà nell'accedere alla strada in cui si trova il negozio, condizioni metereologiche particolarmente avverse o l'impossibilità di rilevare i passaggi se la visuale della telecamera è coperta. Tuttavia in mancanza di ulteriori informazioni o di casi simili verificatisi in passato, il modello fornisce il valore medio delle persone nelle condizioni considerate, che in questo caso risulta essere molto superiore rispetto a quello reale.

In queste circostanze, calcolare il grado di precisione è privo di significato, in quanto il risultato è difforme e quindi anomalo, come mostra la tabella 5.19.

Giorno settimana	N persone reale	Video Vetrina	Commesso	Spot Televisivo	Fenomeni	N persone predetto	Δ	Percentuale di errore	Grado di precisione
Festivo	15	Emozionale	Claudia	Short	Pioggia	149,80927	134,8093	899%	-799%

Tabella 5.19 – Estratto dalla tabella utilizzata per calcolare il grado di precisione dei risultati di predizione

6 CONCLUSIONE

Il modello previsionale realizzato consiste in uno strumento che può risultare utile alle piccole e medie imprese, come negozi di abbigliamento, telefonia o saloni di bellezza, per prendere delle decisioni strategiche dopo aver analizzato la tendenza delle visite passate e stimato quelle future.

Può essere utiizzato per individuare quale sia la strategia di marketing più efficace per promuovere la propria attività in relazione al mercato in cui si opera, al fine di attirare un numero sempre più alto di clienti potenziali e coinvolgere maggiormente quelli attuali.

Osservando la relazione che intercorre tra i differenti parametri, è possibile capire se i cambiamenti applicati nel corso del tempo al negozio sono realmente efficaci o influenzano negativamente il traffico di persone.

Ad esempio, privilegiare la combinazione di una tipologia di vetrina e di uno spot televisivo potrebbe attrarre un target più ampio di passanti.

La conseguenza diretta nell'aumento dei visitatori, infatti, consiste in un numero maggiore di opportunità di vendita e quindi nell'aumento del fatturato del negozio.

6.1 Comparazione tra il modello di correzione attuale e quello elaborato

Per raggiungere questo obiettivo è stato necessario in primo luogo formulare un nuovo modello per correggere gli errori commessi dall'algoritmo di video analisi.

Di seguito vengono riportati degli esempi per confrontare i risultati ottenuti dal modello correttivo attuale rispetto a quelli del modello elaborato.

6.1.1 Correzione giornaliera con confronto per fascia oraria

Gli estratti dalle tabelle 6.1 e 6.2, alle quali sono state applicate le correzioni, rappresentano il numero di persone che si sono recate nel punto vendita nelle ultime

5 ore del 2 maggio 2016. È possibile consultare le tabelle in versione integrale nell'appendice.

Nel primo caso, è stato applicato il *Correction Model Type*, il parametro di esecuzione che governa l'ETL, decidendo a priori di privilegiare le entrate.

Si ricorda quindi che ogni qual volta si ricorre ad un *enter_bind*, vengono corretti i valori delle uscite.

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
02/05/2016	15	17	28	178	-150	
02/05/2016	15	17	11	178	-167	178
02/05/2016	16	10	21	188	-167	
02/05/2016	16	10	11	188	-177	188
02/05/2016	17	9	20	197	-177	
02/05/2016	17	7	13	197	-184	197
02/05/2016	18	27	40	224	-184	
02/05/2016	18	17	23	224	-201	224
02/05/2016	19	3	26	227	-201	
02/05/2016	19	3	23	227	-204	227

Tabella 6.1 – Estratto della tabella che include il modello di correzione attuale applicato alle fasce orarie (2 maggio 2016)

Nel secondo caso invece viene applicato il modello di correzione ideato nel corso di questo studio e finalizzato a ottenere un input più preciso per il modello predittivo.

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
02/05/2016	15	17	28	178	-150	
02/05/2016	15	17	11	178	-167	173
02/05/2016	16	10	21	188	-167	
02/05/2016	16	10	11	188	-177	183
02/05/2016	17	9	20	197	-177	
02/05/2016	17	7	13	197	-184	191
02/05/2016	18	27	40	224	-184	
02/05/2016	18	17	23	224	-201	213
02/05/2016	19	3	26	227	-201	
02/05/2016	19	3	23	227	-204	216

Tabella 6.2 – Estratto della tabella che include il modello di correzione realizzato applicato alle fasce orarie (2 maggio 2016)

Il secondo modello risulta più preciso rispetto al primo in quanto tiene in considerazione non solo l'errore dell'algoritmo di video analisi, ma anche il numero di persone che potrebbero essere ancora presenti nel punto vendita tra una fascia oraria e l'altra.

6.1.2 Correzione mensile con confronto per giorno

Gli estratti dalle tabelle 6.3 e 6.4, alle quali sono state applicate le correzioni, rappresentano invece il flusso di persone che si sono recate nel punto vendita negli ultimi sette giorni considerati per il mese di maggio 2016. È possibile consultare le tabelle in versione integrale nell'appendice.

In questo caso il *Correction Model Type* e il modello di correzione realizzato per questa tesi sono stati applicati al numero totale di visitatori per giorno.

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
24/05/2016	20	2	1	144	-145	144
25/05/2016	19	9	15	140	-155	140
26/05/2016	19	13	5	150	-145	150
27/05/2016	19	6	33	173	-140	173
28/05/2016	19	5	1	127	-128	127
30/05/2016	19	13	5	180	-185	180
31/05/2016	19	14	29	157	-186	157

Tabella 6.3 – Estratto della tabella che include il modello di correzione attuale applicato ai giorni del mese di febbraio 2017

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
24/05/2016	20	2	1	144	-145	145
25/05/2016	19	9	15	140	-155	148
26/05/2016	19	13	5	150	-145	148
27/05/2016	19	6	33	173	-140	157
28/05/2016	19	5	1	127	-128	128
30/05/2016	19	13	5	180	-185	183
31/05/2016	19	14	29	157	-186	172

Tabella 6.4 – Estratto della tabella che include il modello di correzione realizzato applicato ai giorni del mese di febbraio 2017

Mentre il primo modello si limita a compensare i valori calcolando il fattore correttivo giorno per giorno e applicandolo alle entrate o alle uscite, il secondo considera l'andamento del delta in un arco temporale pari ad un mese e lo distribuisce in maniera uniforme.

È preferibile quindi ricorrere al secondo modello di correzione in quanto ripartisce equamente l'errore accumulato nel tempo.

6.2 Prerequisiti del modello

Assicurarsi che i valori forniti al modello predittivo siano il più accurati possibile non è l'unico prerequisito per ottenere un risultato attendibile.

Prima di elaborare un modello, è importante individuare gli elementi che sono stati considerati per pianificare una strategia effettiva di marketing. L'allestimento della vetrina, l'ambientazione del punto vendita, il messaggio pubblicitario, gli stimoli sensoriali, la formazione del personale a contatto con il cliente rappresentano tutti fattori che possono influenzare il flusso di visitatori.

In seguito, bisogna identificare la tipologia di configurazione che ogni elemento può assumere in relazione alla strategia definita. Ad esempio, nel caso di uno spot pubblicitario, è possibile distinguere diverse categorie, quali quella emozionale, comparativa, narrativa, *call to action*, etc.

Una volta determinati gli elementi e le loro configurazioni, è necessario studiare la correlazione che questi parametri hanno con il risultato che si vuole ottenere e individuare gli indicatori chiave di prestazione per misurare il raggiungimento dell'obiettivo.

Quindi, nel caso specifico di questo elaborato, prima ancora di definire i parametri del modello correlati al flusso di visitatori, bisogna definire la strategia alla base del piano marketing finalizzato ad attrarre un numero maggiore di persone nel punto vendita.

6.3 Applicazioni del modello

6.3.1 Ambito strategico

Il modello predittivo per la stima del flusso di visitatori è uno strumento che può essere sfruttato nel medio-lungo termine.

In primo luogo può essere utilizzato per individuare i fattori che creano *engagement* al fine di delineare una strategia marketing basata sugli elementi che attraggono maggiormente le persone.

Monitorare il cambiamento nel numero di ingressi, al variare della configurazione dei parametri, fornisce infatti delle indicazioni su che cosa influenza il flusso ed il comportamento dei visitatori.

Ad esempio, l'affluenza di persone potrebbe risultare maggiore nei giorni durante i quali viene trasmessa sugli schermi in vetrina una determinata tipologia di pubblicità rispetto ad un'altra. Oppure, la combinazione tra una tipologia di allestimento della vetrina e lo spot pubblicitario mandato in onda in quel periodo potrebbe influenzare positivamente il numero di clienti.

Nel lungo periodo, tuttavia l'interesse delle persone può focalizzarsi su un altro tipo di contenuto o un altro elemento di maggiore attrattiva. Poichè l'obiettivo di ogni imprenditore è quello di riuscire a prevedere l'andamento delle preferenze dei consumatori, è importante rimanere allineati con le richieste del mercato per adattare la propria strategia alle nuove esigenze.

Questo strumento può aiutare a rimanere aggiornati su quali siano i fattori che più attirano i consumatori e influenzano il loro comportamento d'acquisto.

6.3.2 Ambito gestionale

Un'altra applicazione del modello predittivo consiste nello sfruttare l'andamento delle visite e lo studio dei fattori di *engagement* per la formazione del nuovo personale.

Si consideri come esempio la figura dello *store manager*. La pianificazione dell'allestimento di una nuova vetrina o di nuove proposte promozionali generalmente sono frutto dell'esperienza pregressa. Nel momento in cui viene deciso di assegnare il ruolo ad una persona che non conosce il mercato di riferimento del brand e che quindi non ha maturato un'esperienza sufficiente, tale persona non ha a disposizione gli strumenti per valutare una corretta strategia.

Lo studio delle correlazioni tra i fattori di *engagement* effettuato grazie al modello predittivo consente al nuovo assunto di elaborare il nuovo piano marketing.

7 BIBLIOGRAFIA

7.1 Video Analisi

[1] Camastra Francesco, Vinciarelli Alessandro, Machine Learning for Audio, Image and Video Analysis - Theory and Applications, Springer-Verlag London, 2015

[2] Verma Brijesh, Zhang Ligang, Stockwell David, Roadside Video Data Analysis - Deep Learning, Springer-Verlag London, 2017

[3] Manjunath Narayana, *Automatic Tracking of Moving Objects in Video for Surveillance Applications,* Department of Electrical Engineering and Computer Science, University of Kansas, 2008, http://vis-www.cs.umass.edu/~narayana/NarayanaMastersThesis.pdf

[4] Domenico Bloisi, Luca Iocchi, Riccardo Leone, Roberta Pigliacampo, *Modellazione del background, segmentazione, tracking e proiezione sul sistema di riferimento GIS globale,* Dipartimento di Informatica e Sistemistica Università di Roma "La Sapienza", 2007, http://www.dis.uniroma1.it/~iocchi/ARGOS/docs/A3.1.pdf

[5] Tiziana D'Orazio, Grazia Cicirelli, Paolo Spagnolo, Tommaso Gramegna, Arcangelo Distante, *Robot mobile autonomo per la sorveglianza di ambienti strutturati,* Istituto di Studi sui Sistemi Intelligenti per l'Automazione, 2006, http://automazione-plus.it/wp-content/uploads/sites/3/2009/06/20060401018_11.pdf

7.2 Machine Learning

[6] Jeff Barnes, *Microsoft Azure Essentials: Azure Machine Learning,* Microsoft Press, 2015

[7] Kevin Patrick Murphy, *Machine Learning: a Probabilistic Perspective,* The MIT Press, 2013

[8] Shai Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning: From Theory to Algorithms,* Cambridge University Press, 2014

[9] Jason Brownlee, *How to Think About Machine Learning,* Machine Learning Mastery, 2018, https://machinelearningmastery.com/what-is-machine-learning/

[10] Danny Sullivan, *How Machine Learning Works, As Explained By Google,* Martechtoday, 2015, https://martechtoday.com/how-machine-learning-works-150366

[11] Simone Favaro, *Forging The Future,* Simone Favaro's Blog, 2017, http://www.simonefavaro.it/2017/04/07/introduzione-machine-learning/

[12] Marco Andolfi, *Retail, come portare il digitale nel negozio e incrementare vendite ed engagement,* Digital4, 2017, https://www.digital4.biz/marketing/ecommerce/retail-come-portare-il-digitale-nel-negozio-e-incrementare-vendite-ed-engagement/

[13] Heather Saphiro, Josée Martens, Jason H, Willam Anton Rohm, Gary Ericson, *Azure Machine Learning Studio,* Microsoft, 2018, https://docs.microsoft.com/en-us/azure/machine-learning/studio/

8 APPENDICE

8.1 Tabella creata dal modulo Score Model di Azure Machine Learning Studio

giorno settimana	N persone corretto	Video Vetrina	Commesso	Spot televisivo	Fenomeni	Scored Label Mean	Scored Label Standard Deviation
Feriale	499	Story Telling	Susanna	Super promo	Sereno	491.178526	32.129199
Feriale	481	Story Telling	Susanna	Super promo	Pioggia	461.620833	8.755965
Feriale	394	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	454	Story Telling	Susanna	Super promo	Sereno	491.178526	32.129199
Festivo	323	Promozionale	Alessandro	Short	Sereno	329.839286	7.176506
Feriale	396	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	415	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	434	Promozionale	Susanna	Super promo	Sereno	429.877381	8.165138
Feriale	369	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	328	Promozionale	Alessandro	Short	Pioggia	337.847768	7.446437
Festivo	393	Promozionale	Susanna	Standard	Sereno	381.324702	16.088468
Feriale	424	Promozionale	Susanna	Super promo	Sereno	429.877381	8.165138
Feriale	289	Emozionale	Alessandro	Short	Sereno	289.55188	46.031138
Feriale	405	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	414	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	429	Promozionale	Susanna	Super promo	Sereno	429.877381	8.165138
Festivo	311	Emozionale	Alessandro	Short	Sereno	246.475	101.278392
Feriale	329	Promozionale	Alessandro	Short	Pioggia	337.847768	7.446437
Feriale	387	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Feriale	339	Promozionale	Alessandro	Short	Nebbia	333.614782	9.076156
Feriale	351	Promozionale	Susanna	Standard	Sereno	385.795753	17.280321
Festivo	319	Emozionale	Alessandro	Short	Pioggia	256.625	69.234488
Festivo	15	Emozionale	Claudia	Short	Pioggia	174.60119	112.750687
Feriale	268	Emozionale	Alessandro	Short	Pioggia	283.837054	29.130342
Feriale	322	Promozionale	Alessandro	Short	Sereno	333.980407	9.689711
Feriale	419	Promozionale	Susanna	Standard	Pioggia	380.541369	17.379665
Feriale	151	Emozionale	Claudia	Short	Sereno	271.120721	97.060351
Festivo	113	Emozionale	Claudia	Short	Sereno	227.954167	126.143082
Festivo	104	Emozionale	Claudia	Short	Sereno	155.240422	123.126013
Feriale	151	Emozionale	Claudia	Short	Sereno	271.120721	97.060351

8.2 Tabella utilizzata per calcolare il grado di precisione dei risultati di predizione

Giorno settimana	N persone reale	Video Vetrina	Commesso	Spot Televisivo	Fenomeni	N persone predetto	Δ	Percentuale di errore	Grado di precisione
Festivo	15	Emozionale	Claudia	Short	Pioggia	149,80927	134,80927	899%	-799%
Festivo	113	Emozionale	Claudia	Short	Sereno	211,64800	98,648	87%	13%
Feriale	151	Emozionale	Claudia	Short	Sereno	224,83267	73,832667	49%	51%
Feriale	151	Emozionale	Claudia	Short	Sereno	224,83267	73,832667	49%	51%
Festivo	104	Emozionale	Claudia	Short	Sereno	148,93200	44,932	43%	57%
Festivo	319	Emozionale	Alessandro	Short	Pioggia	232,56545	86,434554	27%	73%
Festivo	311	Emozionale	Alessandro	Short	Sereno	239,04133	71,958667	23%	77%
Feriale	289	Emozionale	Alessandro	Short	Sereno	250,10100	38,899	13%	87%
Feriale	351	Promozionale	Susanna	Standard	Sereno	395,84520	44,845198	13%	87%
Feriale	454	Story Telling	Susanna	Super Promo	Sereno	494,79112	40,791121	9%	91%
Feriale	268	Emozionale	Alessandro	Short	Pioggia	248,35743	19,64257	7%	93%
Feriale	369	Promozionale	Susanna	Standard	Sereno	395,84520	26,845198	7%	93%
Feriale	419	Promozionale	Susanna	Standard	Pioggia	392,83634	26,163659	6%	94%
Feriale	322	Promozionale	Alessandro	Short	Sereno	337,68240	15,682397	5%	95%
Feriale	415	Promozionale	Susanna	Standard	Sereno	395,84520	19,154802	5%	95%
Feriale	414	Promozionale	Susanna	Standard	Sereno	395,84520	18,154802	4%	96%
Feriale	481	Story Telling	Susanna	Super Promo	Pioggia	459,97833	21,021667	4%	96%
Feriale	328	Promozionale	Alessandro	Short	Pioggia	339,45948	11,459476	3%	97%
Festivo	393	Promozionale	Susanna	Standard	Sereno	379,81353	13,186469	3%	97%
Feriale	329	Promozionale	Alessandro	Short	Pioggia	339,45948	10,459476	3%	97%
Festivo	323	Promozionale	Alessandro	Short	Sereno	330,84124	7,841238	2%	98%
Feriale	387	Promozionale	Susanna	Standard	Sereno	395,84520	8,845198	2%	98%
Feriale	405	Promozionale	Susanna	Standard	Sereno	395,84520	9,154802	2%	98%
Feriale	424	Promozionale	Susanna	Super Promo	Sereno	431,21933	7,219333	2%	98%
Feriale	499	Story Telling	Susanna	Super Promo	Sereno	494,79112	4,208879	1%	99%
Feriale	434	Promozionale	Susanna	Super Promo	Sereno	431,21933	2,780667	1%	99%
Feriale	339	Promozionale	Alessandro	Short	Nebbia	337,19640	1,803603	1%	99%
Feriale	429	Promozionale	Susanna	Super Promo	Sereno	431,21933	2,219333	1%	99%
Feriale	394	Promozionale	Susanna	Standard	Sereno	395,84520	1,845198	0%	100%
Feriale	396	Promozionale	Susanna	Standard	Sereno	395,84520	0,154802	0%	100%

8.3 Tabella che include il modello di correzione attuale applicato ad ogni fascia oraria del 2 maggio 2016

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
02/05/2016	9	13	13	13	0	
02/05/2016	9	8	5	13	-8	13
02/05/2016	10	21	26	34	-8	
02/05/2016	10	12	14	34	-20	34
02/05/2016	11	37	51	71	-20	
02/05/2016	11	44	7	71	-64	71
02/05/2016	12	45	52	116	-64	
02/05/2016	12	43	9	116	-107	116
02/05/2016	13	34	43	150	-107	
02/05/2016	13	26	17	150	-133	150
02/05/2016	14	11	28	161	-133	
02/05/2016	14	17	11	161	-150	161
02/05/2016	15	17	28	178	-150	
02/05/2016	15	17	11	178	-167	178
02/05/2016	16	10	21	188	-167	
02/05/2016	16	10	11	188	-177	188
02/05/2016	17	9	20	197	-177	
02/05/2016	17	7	13	197	-184	197
02/05/2016	18	27	40	224	-184	
02/05/2016	18	17	23	224	-201	224
02/05/2016	19	3	26	227	-201	
02/05/2016	19	3	23	227	-204	227

8.4 Tabella che include il modello di correzione realizzato applicato ad ogni fascia oraria del 2 maggio 2016

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
02/05/2016	9	8	5	13	0	
02/05/2016	9	8	5	13	-8	11
02/05/2016	10	21	26	34	-8	
02/05/2016	10	12	14	34	-20	28
02/05/2016	11	37	51	71	-20	
02/05/2016	11	44	7	71	-64	68
02/05/2016	12	45	52	116	-64	
02/05/2016	12	43	9	116	-107	112
02/05/2016	13	34	43	150	-107	
02/05/2016	13	26	17	150	-133	142
02/05/2016	14	11	28	161	-133	
02/05/2016	14	17	11	161	-150	156
02/05/2016	15	17	28	178	-150	
02/05/2016	15	17	11	178	-167	173
02/05/2016	16	10	21	188	-167	
02/05/2016	16	10	11	188	-177	183
02/05/2016	17	9	20	197	-177	
02/05/2016	17	7	13	197	-184	191
02/05/2016	18	27	40	224	-184	
02/05/2016	18	17	23	224	-201	213
02/05/2016	19	3	26	227	-201	
02/05/2016	19	3	23	227	-204	216

8.5 Tabella che include il modello di correzione attuale applicato mensilmente (febbraio 2017)

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
01/05/2016	16	16	28	299	-271	299
02/05/2016	19	3	23	227	-204	227
03/05/2016	20	2	52	201	-149	201
04/05/2016	17	32	60	384	-444	384
05/05/2016	19	8	36	153	-189	153
06/05/2016	20	3	55	219	-274	219
07/05/2016	19	5	21	125	-104	125
09/05/2016	19	4	18	172	-154	172
10/05/2016	19	5	11	149	-138	149
11/05/2016	19	5	10	137	-147	137
12/05/2016	19	4	4	141	-122	141
13/05/2016	19	8	15	124	-139	124
14/05/2016	20	2	11	138	-127	138
16/05/2016	19	11	2	172	-174	172
17/05/2016	19	8	16	156	-140	156
18/05/2016	19	7	1	134	-120	134
19/05/2016	19	15	26	126	-152	126
20/05/2016	19	10	20	159	-179	159
21/05/2016	20	1	28	142	-114	142
23/05/2016	20	2	15	134	-102	134
24/05/2016	20	2	1	144	-145	144
25/05/2016	19	9	15	140	-155	140
26/05/2016	19	13	5	150	-145	150
27/05/2016	19	6	33	173	-140	173
28/05/2016	19	5	1	127	-128	127
30/05/2016	19	13	5	180	-185	180
31/05/2016	19	14	29	157	-186	157

8.6 Tabella che include il modello di correzione realizzato applicato mensilmente (febbraio 2017)

date [PK] character varying	time [PK] integer	people integer	Δ fine giornata	N persone entrate	N persone uscite	N persone corretto
02/05/2016	19	3	23	227	-204	216
03/05/2016	20	2	52	201	-149	176
05/05/2016	19	8	36	153	-189	172
06/05/2016	20	3	55	219	-274	247
07/05/2016	19	5	21	125	-104	115
09/05/2016	19	4	18	172	-154	164
10/05/2016	19	5	11	149	-138	144
11/05/2016	19	5	10	137	-147	143
12/05/2016	19	4	4	141	-122	132
13/05/2016	19	8	15	124	-139	132
14/05/2016	20	2	11	138	-127	133
16/05/2016	19	11	2	172	-174	174
17/05/2016	19	8	16	156	-140	149
18/05/2016	19	7	1	134	-120	128
19/05/2016	19	15	26	126	-152	140
20/05/2016	19	10	20	159	-179	170
21/05/2016	20	1	28	142	-114	129
23/05/2016	20	2	15	134	-102	119
24/05/2016	20	2	1	144	-145	145
25/05/2016	19	9	15	140	-155	148
26/05/2016	19	13	5	150	-145	148
27/05/2016	19	6	33	173	-140	157
28/05/2016	19	5	1	127	-128	128
30/05/2016	19	13	5	180	-185	183
31/05/2016	19	14	29	157	-186	172