UNIVERSITÀ DI PISA

Corso di Laurea Magistrale in Data Science and Business Informatics

Us Accidents: A Countrywide Traffic Accident Dataset

Camilla Andreazzoli
Sara Hakim
Alessandro Mastrorilli
Davide Ricci
Davide Vecchi

ANNO ACCADEMICO 2023-2024

Indice

1 Introduzione

Il dataset che è stato analizzato, dal titolo "Us Accidents: A Countrywide Traffic Accident Dataset", descrive gli incidenti automobilistici avvenuti in 49 stati degli USA con un arco temporale che va dal 2016 al 2022. Il dataset contiene circa 7.7 milioni di record ed ha una dimensione di circa 3.06 GB. Per questo motivo è stato necessario ridurne le dimensioni restringendosi solo ai dati del 2022 e suddividendo i record in 4 dataset distinti, facendo riferimento alla posizione geografica accorpando i principali stati dell'est, dell'ovest, del nord e del sud.

2 Data Understanding & Preparation

I dataset presentano un numero di record variabile da un minimo di 90000 ad un massimo di 430000 circa, e 46 features. Per prima cosa è stato deciso di droppare 15 colonne in quanto non utili per il nostro studio. Sono state generate le statistiche descrittive per ciascuna feature ed i tipi di dato, che sono rispettivamente numerico (intero e float), timestamp, categorico ed infine booleano. E' stato verificato il numero di valori distinti assunti da ciascun attributo per eliminare eventuali colonne costanti; per tutti i dataset è risultato costante *Turning_ Loop*, mentre per il solo dataset del nord anche *Bump*. A questo punto si è verificata la presenza di valori mancanti. In particolare, ben 10 colonne per nord e sud ed 11 per est ed ovest, coincidono con quelle dei due dataset precedenti ma con l'ulteriore presenza di *City*.
Sono state poi estratte nuove features di interesse anche per i task successivi, partendo dai dati in nostro possesso. Le colonne *day_ of_ the_ week, month, hour* sono state estratte a partire da *Start_ Time*, ed a loro volta *workingDay_ weekend, season* da *day_ of_ the_ week, month*. Inoltre, poichè *Weather_ Condition* presentava quasi 70 valori differenti e molto specifici, è stato discretizzato riducendo i valori assunti dalla variabile a 5, ovvero Fair, Foggy, Cloudy, Snow e Rain.
Giunti a questo punto, sono state visualizzate le distribuzioni delle variabili tramite istogrammi e barplot. In linea generale risulta un evidente sbilanciamento per quasi tutte le features categoriche e booleane. Tra le varie distribuzioni analizzate, ci sono stati risultati interessanti nella distribuzione delle ore relative agli incidenti, differenziando rispetto alla feature *Working_ Weekend*: le due curve mostrano due andamenti molto diversi, infatti per quanto riguarda i giorni feriali c'è un'evidente distribuzione bimodale, con picchi rispettivamente all'incirca alle 7 a.m ed alle 15-16 del pomeriggio; in quella del weekend è possibile riscontrare una sorta di distribuzione negative skwed, con una maggiore frequenza a partire da metà mattinata e con picco a metà pomeriggio. Questi risultati sono presentati in figura 2.1.

È stato osservata la frequenza degli incidenti evidenziando rispetto alla feature *Weather_ Condition*: in questo caso è possibile notare che la maggior parte degli incidenti avveiene quando il tempo risulta normale o nuvoloso; questo potrebbe essere indicativo del fatto che, quando il tempo è stabile, le persone tendono ad uscire maggiormente, con conseguente aumento del numero di automobili presenti per strada. Inoltre, nei momenti più cruciali come pioggia, neve o nebbia, le persone tendenzialmente circolano ad una velocità ridotta, facendo maggiore attenzione.
Sono stati poi osservati gli scatter plot delle varie features. Per i dataset dell'est e dell'ovest sono stati individuati degli errori per la feature pressure, mentre per il solo dataset dell'est anche per temperature. Infatti assumono dei valori che non sono scientificamente ammissibili, ad esempio è stata registrata una temperatura di circa 80 gradi Celsius. In particolare, per il primo dataset sono risultati sette errori, per il secondo cinquanta. Si è quindi deciso di eliminare tali record dai dataset.

Fatto ciò, si è passati alla parte di fillaggio dei missing values.

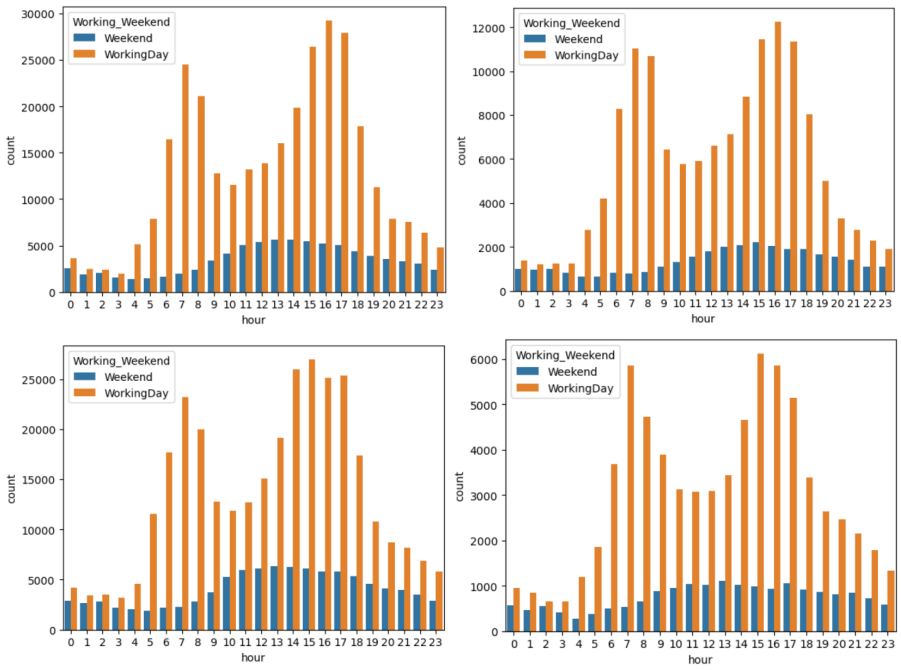

Figura 2.1: Distribuzione delle ore relative agli incidenti, differenziando rispetto alla feature *Working_ Weekend*. Nell'ordine: est, sud, ovest e nord

Gli incidenti accaduti in zone di confine tra stati e contee hanno città mancante e anche utilizzando librerie esterne la città non viene individuata. Visto l'esiguo numero di valori mancanti, è stata deciso di assegnare manualmente la città corretta. Per la discriminazione tra giorno e notte, sono stati assegnati valori a *Astronimical_ Twilight* sulla base della fascia oraria (ottenuta da *Start_ Time*), considerando che varia in base alla *season*. Per le altre variabili che presentano missing values sono state utilizzate la media mobile o la moda a seconda del caso in analisi, raggruppando in base agli attributi più aderenti e significativi e partendo da una granularità più fine, per poi ridurre la granularità se necessario. Si prenda ad esempio il caso di *Weather_ Condition*: per questa in prima battuta si è deciso di assegnare *Rain* nel caso in cui per quell'incidente ci fossero *Precipitation_ in* diverse da 0 e se la *Temperature_ F* fosse al di sopra di 32 Farenheint, se al di sotto invece *Snow*. Dopo di che è stata applicata la logica sopra descritta sulla base della moda, la condizione metereologica più frequente in una certa città in un determinato mese. Dato che però questa granularità è risultata molto fine perchè ci sono mesi in cui in determinate città non ci sono incidenti, si è considerato la *County*. Sulla base di simili ragionamenti sono stati riempiti i valori mancanti degli altri attributi.
A questo punto è stato individuato che i dataset presentavano alcuni duplicati che sono stati rimossi prima di procedere con le successive analisi. Per l'analisi della correlazione ci si è serviti della matrice di correlazione calcolata con l'indice di Pearson. Da questa è emerso che, come già era risultato dall'analisi degli scatterplot, per tutti i dataset c'è un'alta correlazione tra *Temperature_ F* e *Wind_ Chill_ F*. Inoltre, è emersa una correlazione negativa nel dataset dell'ovest tra *Humidity_ perc* e le due features *Temperature_ F* e *Wind_ Chill_ F*. Tale correlazione è una naturale conseguenza delle stesse features, visto il loro significato, infatti la percentuale di umidità è collegata alla temperatura dell'ambiente, tuttavia analizzando gli scatterplot non è stat osservato una tendenza così marcata. Inoltre si è ritenuto che, specialmente nel caso di percentuali di umidità molto alte, questa potesse in qualche modo avere avuto un certo impatto sugli incidenti: il manto stradale potrebbe rimanere umido, comportando una minore aderenza. In conclusione l'unica feature che è stata eliminata per tutti i dataset è stata *Wind_ Chill_ F*. Successivamente è stata analizzata la presenza di eventuali outliers: sono stati normalizzati i dati tramite lo StandardScaler e sono state stampate le loro distribuzioni, a partire da questi si

sono decisi dei valori soglia che sono stati utlizzati per la loro individuazione. E' stato quindi individuato e stimato il numero dei record che assumono valori estremi anche con l'ausilio di boxplot. Da qui si è verificato che i valori estremi assunti in particolare su alcune variabili (ad es la temperatura) non sono da considerarsi errori, ma situazioni estreme e reali, come giornate eccezionalmente calde e fredde, per questo abbiamo deciso di non eliminare questi record.

3 Clustering

In questa sezione sono state riportate le analisi effettuate attraverso diversi algoritmi di clustering: k-means, bisecting k-means, k-modes. Questi algoritmi sono stati applicati ai dataset ottenuti nella sezione 1.

3.1 K-means

Per l'applicazione degli algoritmi k-means e bisecting k-means, sono state utilizzate le features numeriche continue. Quest'ultime sono state normalizzate attraverso il MinMaxScaler applicato sul vector assembler, in modo tale da gestire la presenza di outliers, individuati nella fase di Data Understanding. Il DataFrame risultante (clusteringData) contiene le features normalizzate e la colonna "Severity", che è la variabile target. Inizialmente, è stata condotta un'analisi del clustering attraverso l'utilizzo del K-means. Le prestazioni sono state valutate nell'intervallo compreso tra 2 e 29 cluster, e ad ogni iterazione sono stati stampati i valori di SSE e Silhouette. Successivamente, sono stati generati due grafici distinti per valutare le prestazioni del modello. Il primo grafico illustra l'andamento dell'SSE al variare del numero di cluster nell'intervallo da 2 a 29. L'obiettivo è individuare il punto di inversione di pendenza, comunemente noto come "gomito", che suggerisce il numero ottimale di cluster. La ricerca del punto ottimale è stata fatta sia osservando direttamente il grafico che con l'ausilio del KneeLocator. Il secondo grafico, invece, si basa sul metodo "Silhouette" e mostra il coefficiente di Silhouette per ogni numero di cluster nell'intervallo specificato. Questo grafico mira a identificare il numero ideale di cluster considerando il valore massimo del coefficiente di Silhouette, indicativo di una buona separazione tra i cluster. Entrambi i grafici forniscono una guida visuale per la scelta del numero più appropriato di cluster nel contesto dell'algoritmo K-means.

Per ciascuno dei quattro dataset vengono quindi individuati i migliori valori di k. Viene configurato e addestrato un modello K-means con il k migliore. Viene inoltre creata una tabella di contingenza che mostra la distribuzione della severity nei vari cluster. Infine, i risultati vengono visualizzati attraverso un grafico a barre che illustra la distribuzione delle severity nei cluster. Questa procedura viene poi ripetuta per le variabili State, Astronomical Twilight, Weather Condition, day_of_the_week, season e Working_Weekend.

- Per il dataset Nord sono stati esplorati i valori di k con k = 6,7. i risultati migliori si ottengono con k = 7, con Silhouette = 0.52 e SSE = 1955.07.
- per il dataset Sud, sono stati esplorati i valori di k con k = 6,8. i risultati migliori si ottengono con k = 8, Silhouette = 0.37 e SSE = 21333.25.
- per il dataset Est, sono stati esplorati i valori di k con k = 5,7 i risultati migliori si ottengono con k = 7, Silhouette = 0.43 e SSE = 19108.58.
- per il dataset Ovest, sono stati esplorati i valori di k con k = 4,5,9 i risultati migliori si ottengono con k = 9, Silhouette = 0.48 e SSE = 2444.74.

In linea generale non è stato riscontrato niente di particolarmente rilevante sulla composizione dei cluster ottenuti, ad eccezione della configurazione ottenuta per K=6 nel dataset nord,

scegliendo come label rispetto alla quale visualizzare i dati l'attributo Weather_Condition. Si riporta il risultato nella figura 3.1 sottostante. Come si può osservare, nel cluster 1 sono presenti molti dati associati alla condizione meteorologica della neve. Questo suggerisce una correlazione significativa tra la presenza di neve e il numero considerevole di incidenti, un'associazione che potrebbe essere influenzata anche dalla natura del dataset del Nord America, dove nevica più frequentemente.

Infatti , come si evince dalla figura sottostante, i cluster 3,4 e 5 sono in grado di discriminare molto bene situazioni di nuvolosità e di serenità metereologica dato che gli altri valori sono pressochè assenti.

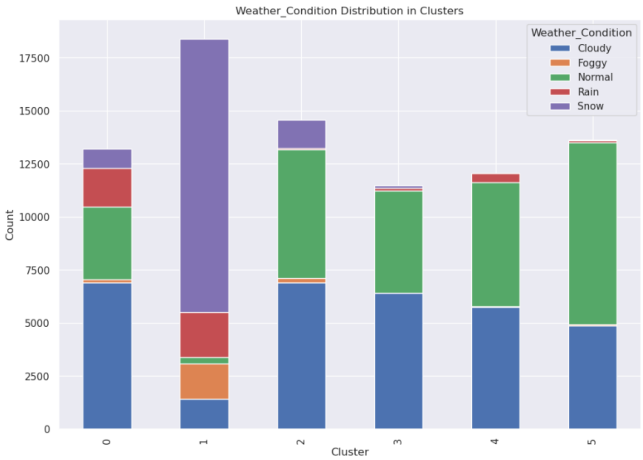

Figura 3.1: Visualizzazione dei cluster rispetto a Weather_Condition

3.3 Bisecting k-means

Il procedimento seguito è analogo a quello del K-means. Si riportano quindi immediatamente i risultati migliori per i dataset.

- Per il dataset Nord è stato testato k = 6, con Silhouette = 0.45 e SSE = 2304.31.
- per il dataset Sud è stato testato k = 8, con Silhouette = 0.38 e SSE = 21649.85.
- per il dataset Est, è stato testato k = 6, con Silhouette = 0.42 e SSE = 20889.94.
- per il dataset Ovest, è stato testato k = 8, con Silhouette = 0.41 e SSE = 3141.06.

  L'analisi ha evidenziato una tendenza complessivamente simile a quella osservata con il K-means, tuttavia con valori con tendenzialmente più bassi per SSE e Silhouette.

3.3 K-modes

In questa sezione vengono presentate le analisi condotte mediante l'algoritmo K-Modes al fine di effettuare il clustering sugli attributi categorici. Per tale scopo, è stata impiegata la libreria esterna "pyspark-distributed-kmodes". Tuttavia, è opportuno evidenziare che l'utilizzo di quest'ultima ha registrato esito negativo durante l'analisi del dataset "West Coast". Di conseguenza, l'analisi si è focalizzata sui restanti tre dataset.

Il workflow seguito è stato il seguente:

- Eliminazione di tutti gli attributi numerici.
- Variazione del parametro k in un intervallo compreso tra 2 e 10, selezionando quello che restituiva i risultati ottimali.

- Analisi dei cluster basandosi sugli attributi: severity, working weekend e season.

Nuovamente, questo esperimento non ha consentito di identificare pattern nascosti all'interno dei vari cluster, poiché non è stato possibile individuare gruppi in grado di discriminare in modo ottimale le diverse classi.

4 Classification

Per ognuno dei quattro casi di studio sono stati definiti modelli di classificazione relativi alla caratteristica Severity e, successivamente, si è cercato di adattare i soliti modelli ad un task di classificazione binaria scegliendo come variabile dipendente Working_Weekend. In particolare sono stati allenati i seguenti modelli facendo opportuno tuning degli iperparametri:

MODELS	PARAMETERS
Decision tree	**max_depth** = [5,10,15,25] **max_bins** = [32,64] **min_inst** = [round(0.005*n),round(0.02*n)] **impurity** = ['entropy', 'gini']
Random forest	**max_depth** = [5,10,15,20]
Multi perceptron layer	**layers** = [n,16,8,4,4], [n,8,4,4], [n,16,16,16,4],[n,8,8,8,4],[n,16,8,4], [n,16,4],[n,8],[n,4] *n = numero di features **maxIter** = 100,200,500 **tol** = 1e-6, 1e-5, 1e-4 **blocksize** = 64,32,16

Figura 4.1: modelli e parametri testati

Per quanto riguarda i parametri del decision tree, sono state settate anche tre configurazioni per i pesi. Queste prendono in considerazione il rapporto tra il numero di record della classe maggioritaria e delle varie classi minoritarie, eventualmente riscalate per opportuni coefficienti nell'intervallo (0,1].

4.1 Severity prediction

Per conseguire il task di classificazione sono state applicate tecniche di OneHotEncoding per le variabili categoriche e successivamente l'intero dataset è stato opportunamento vettorizzato e diviso in due insiemi di train e test con rispettivamente 70 e 30 percento dei dati. Dopo aver notato un estremo sbilanciamento a favore della classe maggioritaria 2 della variabile target, si è deciso di allenare i modelli su 3 insiemi differenti

- Training set no sampling
- Undersampling classe maggioritaria riducendo il numero di record ad un numero uguale alla media dei record delle classi minoritarie.
- Undersampling classe maggioritaria riducendo il numero di record al 30 percento della quantità originale + Oversampling classi minoritarie generando duplicati fino a raggiungere circa lo stesso numero di quelli della classe maggioritaria.

e di confrontare successivamente le rispettive performance sullo stesso insieme di test.

	NORD	SUD	EST	OVEST
DECISION TREE	94% accuracy 55% F1_macro	92% accuracy 56% F1_macro	91% accuracy 53% F1_macro	94% accuracy 58% F1_macro
RANDOM FOREST	93% accuracy 57% F1_macro	92% accuracy 60% F1_macro	89% accuracy 58% F1_macro	95% accuracy 62% F1_macro
ML PERCEPTRON	93% accuracy 33% F1_macro	89% accuracy 27% F1_macro	89% accuracy 29% F1_macro	94% accuracy 24% F1_macro

Figura 4.2: no sampling results

	NORD	SUD	EST	OVEST
DECISION TREE	71% accuracy 53% F1_macro	86% accuracy 56% F1_macro	60% accuracy 50% F1_macro	80% accuracy 49% F1_macro
RANDOM FOREST	64% accuracy 51% F1_macro	83% accuracy 57% F1_macro	42% accuracy 51% F1_macro	84% accuracy 51% F1_macro
ML PERCEPTRON	57% accuracy 58% F1_macro	45% accuracy 28% F1_macro	27% accuracy 22% F1_macro	41% accuracy 20% F1_macro

Figura 4.3: undersampling results

	NORD	SUD	EST	OVEST
DECISION TREE	73% accuracy 52% F1_macro	64% accuracy 51% F1_macro	64% accuracy 52% F1_macro	89% accuracy 52% F1_macro
RANDOM FOREST	88% accuracy 61% F1_macro	55% accuracy 48% F1_macro	69% accuracy 55% F1_macro	65% accuracy 46 % F1_macro
ML PERCEPTRON	47% accuracy 28% F1_macro	28% accuracy 23% F1_macro	30% accuracy 23% F1_macro	NA

Figura 4.4: oversampling and undersampling results

4.2 Working Weekend prediction

La preparazione ai dati per questa analisi di classificazione binaria è la medesima dello studio precedente fatta eccezione per la fase di vettorizzazione in cui è stata aggiunta la variabile Severity ed esclusa la target attuale. Dopo lo step di train e test split non è stato riscontrato un marcato sbilanciamento tra i due valori della variabile dipendente, tuttavia su tre dataset sono state applicate tecniche di oversampling e undersampling per valutare se fosse possibile un miglioramento delle performance. Nel seguito i risultati ottenuti in ognuno dei 4 casi di studio per il caso comune a tutti e 4 i dataset (training set no sampling).

	NORD	SUD	EST	OVEST
DECISION TREE	80% accuracy 52% F1_macro	81% accuracy 53% F1_macro	79% accuracy 50% F1_macro	71% accuracy F1 score 70%
RANDOM FOREST	82% accuracy 61% F1_macro	84% accuracy 65% F1_macro	81% accuracy 56% F1_macro	80% accuracy F1-score 57 %
ML PERCEPTRON	80% accuracy 47% F1_macro	82% accuracy 44% F1_macro	79% accuracy 48% F1_macro	64% accuracy F1 score 55%

Figura 4.5: results of binary classification

4.3 Model Evaluation

4.3.1 Multiclass Evaluation

Per valutare le performance dei vari modelli testati in questo primo esperimento multiclasse la concentrazione è stata rivolta principalmente sulla metrica F1 score. In particolare il MLP, fatta eccezione per il Nord con Undersampling, non ha ottenuto risultati soddisfacenti in nessuna situazione, in quanto nella maggior parte dei casi restituiva predizioni relative alla classe maggioritaria a discapito delle restanti. Essendo obiettivo del nostro caso di studio la definizione di

modelli che riescano a classificare ciascun incidente in base ad ognuno dei 4 valori di severità possibili si può affermare che, per ogni punto cardinale, i migliori modelli sono risultati essere:

- NORD → Random Forest(88 % accuratezza e 61 % F1), Undersampling + Oversampling
- SUD → Random Forest(92 % accuratezza e 60 % F1), No Sampling
- EST → Random Forest(89 % accuratezza e 58 % F1), No Sampling
- OVEST → Random Forest(95 % accuratezza e 62 % F1), No Sampling

Come si può notare dai risultati il RF ha performato decisamente meglio rispetto agli altri ottenendo un coefficente di f1 macro tendente al 60 percento in ognuno dei 4 dataset di test cardinali, tale coefficente in un esperimento con 4 classi è stato considerato più che accettabile, confrontato anche alle performance dei restanti modelli. Tale conclusione è stata definita anche dal fatto che riesce in ogni caso di studio a generare predizioni per ognuna delle 4 possibili classi, come testimoniano anche le seguenti confusion matrix in figura sottostante 4.6.

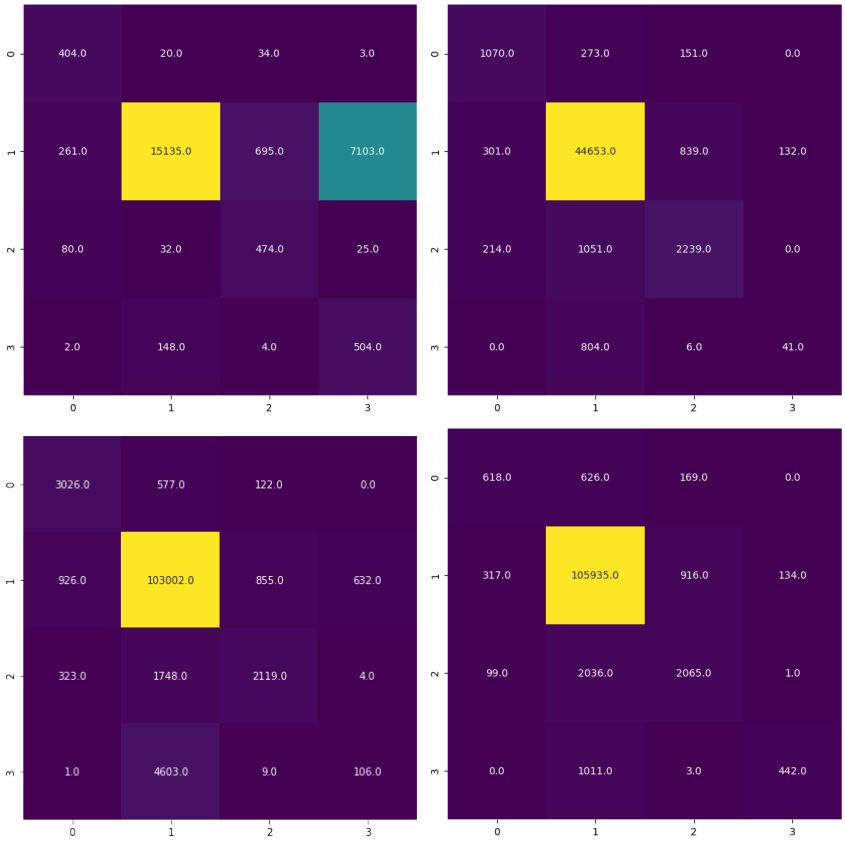

Figura 4.6: Confusion Matrix per Nord,Sud,Est,Ovest

Si è deciso infine di ispezionare quali fossero le variabili che discriminano meglio la target Severity nei vari RF ottenendo il seguente risultato:

- NORD → Distance 0.31, Long 0.14, Lat 0.11, Temperature 0.10
- SUD → Distance 0.57, Hour 0.12, Month 0.10
- EST → Distance 0.35, Lat 0.10, Month 0.09
- OVEST → Distance 0.38, Lat 0.09, Long 0.09

La variabile che discrimina maggiormente la Severity in ognuno dei 4 casi di studio è Distance, che rappresenta la lunghezza della strada nella quale è avvenuto l'incidente. Una strada con una lunghezza maggiore permette di collegare più punti di interesse, e questo si ripercuote direttamente sulla quantità di traffico a cui è esposta. Si potrebbe dunque pensare che, in strade di questo tipo, sarà più probabile che a seguito di un incidente si registrerà un determinato impatto sul traffico.
Altre features che sono risultate importanti sono le coordinate geografiche e temporali di un determinato incidente. Le prime infatti permettono di identificare la città nella quale è avvenuto il fatto, mentre le seconde forniscono informazioni sul momento, che è direttamente collegato al numero di veicoli che transitano la strada.

4.3.2 Binary Evaluation

Per valutare le performance dei vari modelli in questo nuovo scenario di classificazione binaria si è deciso di prendere in considerazione accuratezza e F1 in egual modo. In particolare si è notato un netto miglioramento della rete neurale rispetto al precedente esperimento, soprattutto se definita con 1 oppure nessun layer intermedio, rendendola di fatto un semplice Perceptron. Nel seguito le performance predittive per la variabile Working_Weekend dei migliori modelli testati per i vari casi:

- NORD → Random Forest(82% accuratezza e 61% F1)
- SUD → Random Forest (84% accuratezza e 65% F1)
- EST → Random Forest (81% accuratezza e 56% F1)
- OVEST → Decision Tree (71% accuratezza e 70% F1)

Anche nella predizione della variabile Working Weekend il modello Random Forest risulta essere quello più appropriato restituendo nelle sue migliori configurazioni le seguenti matrici di confusione in figura 4.7 posta all'inizio della pagina seguente.

Le variabili con più importanza nella predizione della target binaria sono risultate essere le seguenti:

- NORD → Hour 0.13, Humidity 0.13, Distance 0.10
- SUD → Hour 0.41, Astronomical Twilight 0.40, Distance 0.10
- EST → Humidity 0.12, Hour 0.10, Lng 0.08
- OVEST → Lng 0.15, Humidity 0.13, Distance 0.13

In particolare la caratteristica relativa all'ora in cui ciascun incidente si è verificato risulta essere la variabile che discrimina meglio se il giorno relativo cade nel weekend oppure no.

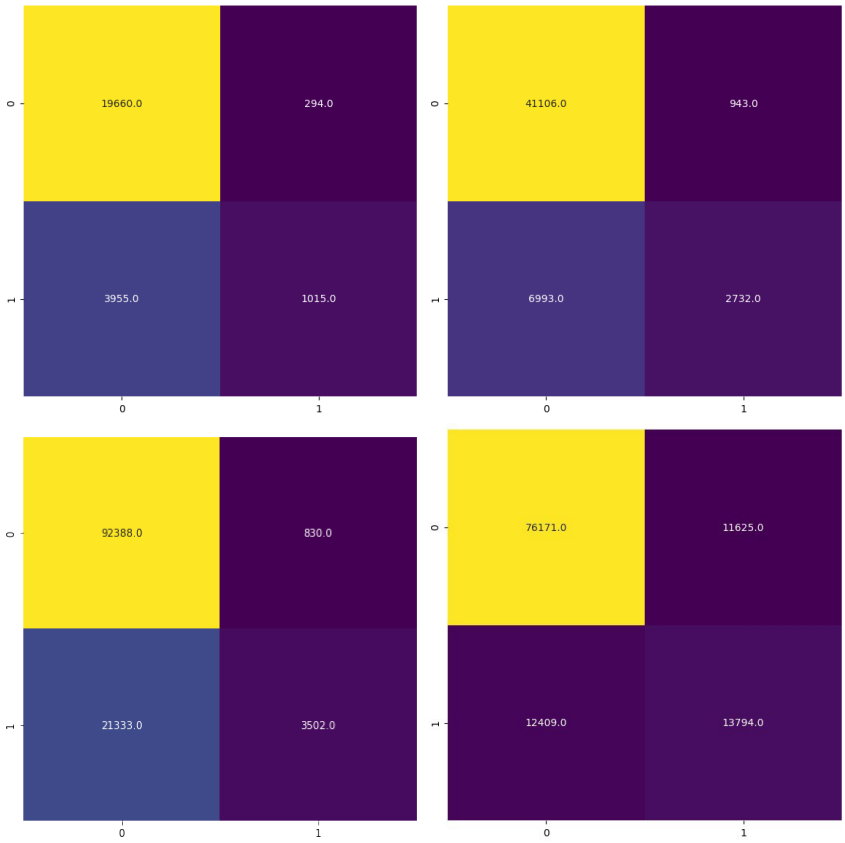

Figura 4.7: Confusion Matrix per Nord,Sud,Est,Ovest

5 Conclusioni

Attraverso la suddivisione del dataset iniziale in quattro dataset distinti in base alle direzioni geografiche (nord, sud, est, ovest), è possibile ottenere una segmentazione più chiara e specifica dei dati.

Al termine degli esperimenti condotti, che hanno come oggetto la feature "severity", per quanto riguarda il clustering, l'analisi dei valori di silhouette nei metodi di clustering K-means e Bisecting K-means ha indicato che, in termini di silhouette e somma degli errori quadratici (SSE), non sono stati ottenuti cluster nettamente distinti. Di questi, il K-means è emerso come il più efficace, consentendo talvolta di identificare pattern significativi come illustrato nella Figura 3.1. Dall'altro lato, l'applicazione del modello K-modes ha prodotto risultati meno soddisfacenti.

Durante l'implementazione di vari modelli di classificazione su ciascun dataset, è emerso che il modello multiclasse non bilanciato ha raggiunto un'accuratezza tendente al 93% per tutti i dataset. Tuttavia, questo risultato era influenzato dalla tendenza del modello a predire costantemente la classe maggioritaria. Per risolvere questo problema, si è proceduto con l'undersampling, che ha comportato un peggioramento delle prestazioni complessive. Successivamente, l'applicazione di una combinazione di undersampling e oversampling ha portato a un miglioramento dell'F1-macro, a discapito di una leggera diminuzione dell'accuratezza.

I risultati più significativi sono stati raggiunti utilizzando il dataset relativo alla regione nord, con un'accuratezza del 82% e un punteggio F1-macro del 61%. Per quanto riguarda la classificazione binaria, il modello Random Forest si è dimostrato il più performante su tutti i dataset. Nel dettaglio, per la regione sud si è ottenuta un'accuratezza dell'84% e un punteggio F1 del 65%, mentre per la regione ovest il Decision Tree ha raggiunto un'accuratezza del 71% e un punteggio F1 del 70%.

Va inoltre osservato che il Decision Tree è stato il modello più performante, specialmente nel dataset relativo alla regione ovest, dove si è ottenuto un F1 score del 70%.

In conclusione, sebbene ci fosse spazio per ulteriori miglioramenti attraverso un ottimale tuning degli iperparametri, tale operazione non è stata praticabile a causa di limitazioni computazionali.