Treball de Fi de Màster

DOBLE MÀSTER UNIVERSITARI EN ENGINYERIA INDUSTRIAL I ENGINYERIA D'ORGANITZACIÓ

Caracterització d'accidents a la ciutat de Barcelona mitjançant clustering

MEMÒRIA

25 d'abril de 2022

Autor: *Oriol Vidal Claramunt*

Director: *Jordi Olivella Nadal*

Convocatòria: *Abril 2022*

ETSEIB

Escola Tècnica Superior

d'Enginyeria Industrial de Barcelona

UPC

Resum

Al present treball es realitza un anàlisi dels accidents de trànsit que han succeït a la ciutat de Barcelona durant els anys 2018 i 2019 mitjançant clustering, tècnica de Machine Learning no supervisat. Els objectius es centren en detectar aquelles situacions comuns en el conjunt d'accidents, partint de les dades públiques de que es disposa i caracteritzant-les mitjançant un conjunt de variables relacionades amb els accidents, l'estat del trànsit i la meteorologia. També es crea un precedent de la utilització de les esmentades tècniques en l'estudi i caracterització dels accidents de trànsit a la cuitat de Barcelona. Fent un anàlisi i adequació de les variables recollides, s'aplica els algorismes seleccionats a dos grups de dades diferents i s'analitza els resultats obtinguts, avaluant la seva validesa i rellevància. Aquests grups són el conjunt de tots els accidents i el subgrup d'accidents més greus. Finalment, es conclou que les situacions identificades són rellevants, de gran interès i que el mètode emprat ha aconseguit capturar les característiques dels accidents estudiats.

Resumen

En el presente trabajo se realiza un análisis de los accidentes de tráfico ocurridos en la ciudad de Barcelona durante los años 2018 y 2019 mediante clustering, técnica de Machine Learning no supervisado. Los objetivos se centran en detectar aquellas situaciones comunes en el conjunto de accidentes, partiendo de los datos públicos de que se dispone y caracterizándolos mediante un conjunto de variables relacionadas con los accidentes, el estado del tráfico y la meteorología. También se crea un precedente de la utilización de las citadas técnicas en el estudio y caracterización de los accidentes de tránsito en la ciudad de Barcelona. Haciendo un análisis y adecuación de las variables recogidas, se aplican los algoritmos seleccionados a dos grupos de datos diferentes y se analizan los resultados obtenidos, evaluando su validez y relevancia. Estos grupos son el conjunto de todos los accidentes y el subgrupo de accidentes más graves. Por último, se concluye que las situaciones identificadas son relevantes, de gran interés y que el método empleado ha logrado capturar las características de los accidentes estudiados.

Abstract

The present work carries out an analysis of the traffic accidents that occurred in the city of Barcelona during the years 2018 and 2019 using clustering an unsupervised Machine Learning technique. The objectives focus on detecting those common situations in the group of studied accidents, based on the public data available and characterizing them through a set of variables in relation to accidents, traffic conditions and weather. A precedent is also created for the use of the aforementioned techniques in the study and characterization of traffic accidents in the city of Barcelona. Making an analysis and adaptation of the variables, the selected algorithms are applied to two different data groups and the results obtained are analysed, evaluating their validity and relevance. These groups are the set of all accidents and the subgroup of the most serious accidents. Finally, it is concluded that the situations identified are relevant, of great interest and that the method used has managed to capture the characteristics of the studied accidents.

Índex de contingut

Índex de taules

Índex de figures

1. Prefaci

A la secció es presenten l'origen i motivació de desenvolupar el projecte i els requeriments imprescindibles per portar-lo a terme.

1.1. Origen i motivació

El projecte neix a partir de la proposta del Professor Jordi Olivella en línia amb l'interès per part de l'alumne en l'anàlisi de dades i les tècniques de clustering. Actualment, l'accidentalitat a la carretera és un tema de gran importància per als organismes responsables a causa de l'impacte generat a les vides de les persones i les conseqüències econòmiques que comporten. S'ha publicat diversos articles adreçant la problemàtica i generant informació rellevant per tal de reduir-los. Entre aquests, es proposa la utilització de mètodes estadístics tradicionals i també de Machine Learning, que els últims anys prenen protagonisme. Per tal de fer un anàlisi dels accidents, es disposa de les dades recollides i publicades per part del servei de dades obertes Open Data BCN, on s'identifiquen característiques descriptives relacionades amb els sinistres viaris a la ciutat de Barcelona. No s'identifica cap estudi fet públic sobre aquestes dades amb mètodes de classificació mitjançant característiques, i així es fa al present projecte. Les principals motivacions personals per la realització d'aquest treball són la possibilitat d'aprendre a treballar i analitzar dades mitjançant una tècnica de Machine Learning com és el clustering, aplicant-la sobre un cas real. Finalment, la possibilitat de provar noves metodologies i aplicar tècniques no utilitzades abans en l'àmbit de la identificació, prevenció i mitigació dels accidents de trànsit també és un al·licient per al desenvolupament del treball.

1.2. Requeriments

Per poder plantejar i executar el projecte és important tenir coneixements sobre la gestió de treballs emmarcats en la ciència de les dades i les tècniques de Machine Learning, alhora que tenir un bagatge o en el seu defecte, formar-se en aquests camps de la ciència. Més concretament, sobre tècniques de clustering de dades, la tècnica principal emprada en aquest projecte, i la manera de plantejar un problema, definir la seva resolució emprant dites tècniques i interpretar-ne la validesa i significació dels resultats. Addicionalment, per tal de fer tot el desenvolupament tècnic i d'exploració de les dades és essencial tenir coneixements sobre els llenguatges de programació més habituals en aquest tipus de treballs, generalment Python o R. Tota la recerca relacionada, recollida, anàlisi i tractament de dades i realització de proves es desenvolupa en un entorn de programació i simulació.

2. Introducció

Amb les millores tecnològiques i de les infraestructures de les últimes dècades, la mobilitat de les persones ha experimentat un creixement exponencial. Amb aquest, han augmentat el nombre d'accidents: a tot el món, gairebé cinquanta milions de persones pateixen lesions cada any com a conseqüència d'un sinistre de trànsit [1]. Si no s'adreça la problemàtica, s'estima que al 2030 sigui la principal font de lesions per a les persones [2] i així s'ha fet per les Nacions Unides, que va definir l'objectiu de reduir les lesions derivades d'aquest tipus d'accidents establint com a fita l'any 2030 [3].

El projecte estudia els accidents de trànsit a la ciutat de Barcelona a partir de les dades recuperables a través del servei de dades obertes Open Data BCN, del propi Ajuntament. Aquest servei de recollida i exposició al públic de les dades és poc freqüent al món. Algunes ciutats també fan públiques algunes mesures estadístiques, com els accidents segons el barri, la zona o la presència de víctimes, però cap amb la profunditat de les característiques úniques de cada accident que es disposa per part de l'Ajuntament de Barcelona. Aquesta informació s'actualitza anualment i es disposa, amb lliure accés, cinc arxius de dades per cada any, relacionats entre ells a través del codi d'accident. Aquests contenen diferents variables que identifiquen les característiques de les persones implicades, ja sigui en forma de conductor, passatger o vianant. També es detallen alguns atributs dels vehicles accidentats, del tipus d'accident, de la causa i de la gravetat d'aquests. Aprofitant aquesta possibilitat, es parteix de les dades per generar informació altament rellevant per a l'autoritat competent, qui podrà interpretar-la i plantejar mesures de contenció i mitigació per fer disminuir el nombre de sinistres a la ciutat.

2.1. Estudis previs

Amb les dades disponibles es poden fer, principalment, tres tipus estudis: els exercicis de previsió de l'ocurrència d'accidents, predicció de la gravetat d'aquests i agrupació segons les seves característiques per la detecció de situacions habituals.

Exercicis de previsió de l'ocurrència

Els primers d'aquests tenen l'objectiu de desenvolupar models de previsió del succés d'un output o situació, com pot ser un accident o una malaltia a través de diversos factors. Amb aquests models, es pot avaluar la probabilitat d'ocurrència d'accidents a partir de les característiques d'un tram de circulació o una zona concretes, predint accidents futurs partint de dades passades. A la següent taula es presenta un resum dels articles revisats que realitzen aquest tipus d'exercicis.

Autors	Article	Objectius	Metodologia
Artime Ríos et al.	*Genetic algorithm based on support vector machines for computer vision syndrome classification in health personnel*	Seleccionar les característiques més rellevants en l'ocurrencia del síndrome CVS. Desenvolupar un model predictiu entre el personal del sector de la salut	Support Vector Machine (SVM) i genetic algorithms
Santos D. et al.	*Machine Learning Approaches to Traffic Accident Analysis and hotspot prediction*	Desenvolupar un model predictiu per a futurs accidents de trànsit emprant dades passades	Decision Tree, Random Forests, Logistic Regression, naive Bayes

Taula 2-1. Revisió de la literatura, exercicis de previsió de l'ocurrència.

Un estudi d'aquesta tipologia és el desenvolupat per Aritme Ríos (2018) [4]. Es busca seleccionar les característiques més rellevants dels subjectes que poden propiciar la ocurrència del síndrome i una vegada identificades, desenvolupar un model de classificació per fer-ne la previsió.

També es fa un estudi enfocat a aquest tipus de previsió a l'article de Santos D. et al. (2021) [5], on es presenta una utilització d'algorismes de Machine Learning per fer prediccions d'incidents en entorns de treball del sector de l'agricultura. Aquests models de previsió es creen a partir de diferents algorismes, com poden ser els arbres de decisió (o Decision Trees), Random Forests, regressions logístiques i classificadors de tipus naive Bayes. També s'utilitza Support Vector Machines entrenat amb el recolzament d'un algorisme genètic.

Estudis de predicció de la gravetat

En segon lloc, es desenvolupen estudis de predicció de la gravetat dels accidents que permeten identificar variables amb efecte sobre aquesta. El resum de la literatura estudiada es presenta a [Taula 2-2].

Un treball estudiat amb aquest objectiu és el Treball Final de Grau d'un alumne de l'escola, Reverter (2021) [6]. Aquest parteix de les mateixes dades sobre els accidents de trànsit a la ciutat i busca fer una classificació segons la seva severitat, identificant els factors i relacions entre variables que contribueixen a la gravetat de l'accident. Es planteja una metodologia amb Random Forests per fer una selecció prèvia de variables i CART per entendre les relacions subjacents entre variables explicatives.

Autors	Article	Objectius	Metodologia
Reverter E.	*Predicting The Severity Of Road Traffic Accidents In The City Of Barcelona.*	Analitzar la gravetat dels accidents i predir-la per a futurs sinistres, identificant factors determinants	Random Forests i Classification And Regression Trees
Chen et al.	*Investigating driver injury severity patterns in rollover crashes using support vector machine models*	Investigar els factors relacionats amb els accidents "rollover". S'identifica variables significatives per identificar patrons de gravetat en els accidents.	Classification and regression tree (CART), Support Vector Machine (SVM)
Chang i Wang	*Analysis of traffic injury severity: An application of non-parametric classification tree techniques*	Establir la relació entre el nivell de gravetat i les característiques relacionades.	Classification and regression tree (CART)

Taula 2-2. Revisió de la literatura, estudis de predicció de la gravetat.

També es presenten estudis on es fa predicció de la gravetat a les següents publicacions. A la primera d'aquestes (Chen et al., 2016) [7] s'investiga els factors que afecten als accidents tipus "rollover", on el vehicle fa el que popularment es coneix com "voltes de campana", i el seu impacte en la gravetat de l'accident. S'estudia diverses variables relacionades amb l'accident, com l'entorn de l'accident, les característiques del/s vehicles, característiques demogràfiques del conductor i patrons de conducta. S'utilitza un model CART (arbre de classificació i regressió) per tal d'identificar variables significatives i algorismes de SVM per tal d'avaluar-ne el rendiment [8].

A la segona de les publicacions, Chang i Wang (2006) [9], es realitza un estudi dels accidents de trànsit a la ciutat de Taipei durant l'any 2001. Es desenvolupa un model CART (Classification And Regression Tree) per establir la relació entre la severitat de l'accident i les característiques del conductor i del vehicle, igual que també les de la via de circulació i aspectes mediambientals. Els resultats indiquen que la variable més important relacionada amb la gravetat de l'accident és el tipus de vehicle. Vianants i conductors de motocicletes i bicicletes són identificats com els que tenen un risc més alt de ser ferits d'entre tots els usuaris de les vies de circulació.

Caracterització

La tercera tipologia dels estudis possible tracta de realitzar diverses agrupacions segons les característiques dels objectes a estudiar, amb l'objectiu de detectar situacions habituals. D'aquestes, se'n pot analitzar el perquè apareixen amb tanta freqüència per després analitzar la millor manera de reduir-los.

Es presenta cinc articles on es fa aquest tipus d'estudi.

Autors	Article	Objectius	Metodologia
Kaplan et al.	*Cyclist–Motorist Crash Patterns in Denmark: A Latent Class Clustering Approach*	Reconèixer patrons en els accidents amb ciclistes i motoristes aplicant tècniques de clustering.	Latent Class Clustering
Depaire i Vanhoof	*Traffic Accident Segmentation by Means of Latent Class Clustering*	Segmentació d'un grup d'accidents en set grups que identifiquen set tipologies diferents.	Latent Class Clustering
Kim i Yukio	*Using a K-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii*	Analitzar patrons espacials en els accidents on hi ha vianants involucrats.	K-means
Shweta et al.	*A Framework for Analyzing Road Accidents Using Machine Learning Paradigms*	Segmentació d'accidents de trànsit i posterior aplicació de mètodes per selecció de variables rellevants	K-means i Feature Selection
Li L. et al.	*Analysis of road traffic fatal accidents using data mining techniques*	S'investiga la relació entre els accidents fatals amb els atributs d'aquests.	Algorisme Apriori, naive Bayes i K-means.

Taula 2-3. Revisió de la literatura, estudis de caracterització dels accidents.

En primer lloc, a l'article de Kaplan et al. (2013) [10] es fa l'anàlisi dels accidents de ciclistes i motoristes a Dinamarca, mitjançant algorismes de Latent Class Clustering, identificant situacions habituals entre les dades estudiades. Aquestes situacions es caracteritzen segons variables majoritàries entre els accidents agrupats al clúster.

A la publicació de Depaire et al. (2008) [11] es proposa segmentar les dades d'accidents, també amb la tècnica Latent Class Clustering i identificar-ne tipus heterogenis. El grup de dades inicial es segmenta en set clústers, que es traslladen a set tipus d'accidents. En segon lloc, es desenvolupa un anàlisi sobre cadascun d'aquests, comparant-los amb l'anàlisi fet sobre el conjunt total de dades.

Un altre mètode d'operació per a l'agrupació d'accidents segons les seves característiques es planteja i desenvolupa a l'article Kim, K. et al. (2007) [12]. Es descriu la tècnica coneguda com K-means en relació a les seves ventatjes i inconvenients per tal d'analitzar patrons espaials en accidents relacionats amb un vianant.

Un segon exemple de l'aplicació dels algorismes K-means és Shweta et al. (2021) [13]. S'aplica l'algorisme K-means de classificació en clústers per tal de fer una segmentació dels

accidents i fer-ne posteriorment, un anàlisi modelant les dades per extreure'n les característiques més rellevants, imatges i patrons amagats utilitzant un algorisme de Machine Learning supervisat per ajudar a crear polítiques per la prevenció d'accidents de trànsit.

A l'article L. LI et al. *(2017)* [14] es presenta un anàlisi basat en els algorismes Apriori, naïve Bayes, i K-means per analitzar el joc de dades i examinar la relació entre els accidents fatals i altres característiques, com el fet que el conductor vagi sota els efectes de l'alcohol, condicions lumíniques, situació de la col·lisió, condicions climàtiques i les condicions de la via per on circula el vehicle.

A continuació, es presenta una revisió de la literatura existent al voltant de la relació entre la densitat de trànsit i l'estat de la meteorologia amb l'ocurrència i gravetat dels accidents.

Densitat del trànsit

Alguns articles mostren un efecte positiu sobre la seguretat viària producte de l'increment de la congestió, mentre que d'altres mostren relacions contràries. Com es suggereix als articles de Albalate et al. (2021) [15][14], Harwood et al. (2013) [16] i Abdel-Aty et al. (2016) [17], entre d'altres, es proposa una relació no linear i quadràtica que confirma el benefici que suposa en matèria d'accidents una ciutat amb una certa densitat de circulació i el detriment d'una ciutat molt fluida, tant en nombre d'accidents com en la gravetat d'aquests. A la següent figura s'identifica la relació no linear i quadràtica proposada.

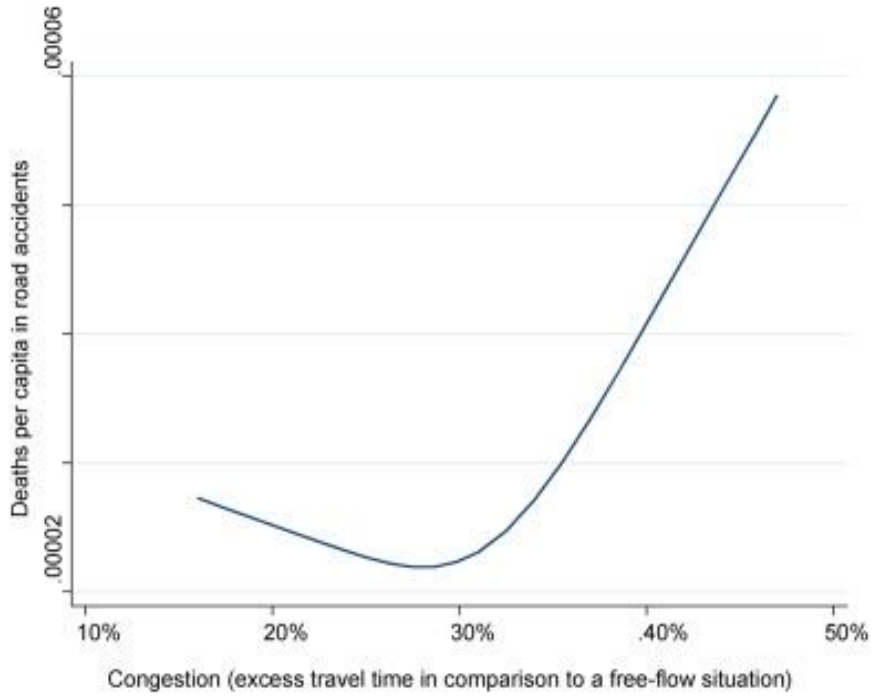


Figura 2-1. Relació entre el nivell de densitat i el nombre d'accidents mortals per càpita, [15].

Es conclou que una situació d'alta congestió, que s'identifica amb un temps de desplaçament igual o superior a 1.3 vegades al temps de desplaçament en flux lliure, implica un increment en el nombre de víctimes fatals, com s'il·lustra a [Figura 2-1].

Un altre article que busca estudiar la relació entre aquests factors és *The relationship between congestion levels and accidents* [18], on s'arriba a uns resultats diferents dels de l'anterior article. S'investiga la relació entre la congestió i els accidents, posant especial èmfasi en l'impacte de la densitat de trànsit en la freqüència d'accidents i la seva gravetat. Les dades estudiades corresponen als accidents de cinc autopistes i cinc carreteres secundàries de l'estat de Maryland, als Estats Units d'Amèrica. S'utilitza mètodes estadístics multivariable i s'arriba a la conclusió que l'impacte depèn del tipus de via de circulació i té una important variació entre les franges de més i menys trànsit.

Els resultats revelen que els accidents en carreteres secundàries tendeixen a disminuir amb l'increment en el volum de congestió. Contràriament, en les autopistes durant les hores de màxima afluència, on hi ha més congestió, augmenten els accidents. També es detalla que en les autopistes, durant els segments de menys trànsit, no sembla tenir un efecte clar sobre el nombre d'accidents que es donen a dites vies de circulació. Es conclou que es tracta d'una relació complexa, donat que els resultats publicats són ambigus i altament dependents dels mètodes utilitzats, les dades i el context.

Condicions meteorològiques

Per acabar, es revisa la possible relació existent dels accidents amb les condicions meteorològiques, per tal d'avaluar si és procedent incloure dades representatives de la meteorologia al present estudi. No són molts els articles disponibles en aquest aspecte. Entre ells, n'hi ha que identifiquen una molt alta correlació entre les condicions meteorològiques i la freqüència i gravetat dels accidents, com són els titulats *Road accidents and rainfall in a large Australian city*, de Keay et al. [19] i *Effects of Rainfall on Vehicle Crashes in Six U.S. States*, de Black et al. [20]. Altres, relacionen l'impacte de precipitacions passades en la freqüencia dels accidents, com l'article *The mixed effects of precipitation on traffic crashes*, de Eisenberg et al. [21]. Finalment, també hi ha articles que conclouen que els efectes del clima no tenen impacte en els accidents i la seva ocurrència, *Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accidents,* Fridstrøm et al. [22].

Un altre d'aquests és l'article que es titula *The association between meteorological factors and road traffic injuries: a case analysis from Shantou city, China* [23], on s'empren diferents tècniques de regressió lineal per investigar la relació entre aquests a la ciutat de Shantou. Es conclou que hi ha un augment d'accidents de trànsit en les estacions d'estiu i hivern. Dels

anàlisis de correlació se'n deriva que els accidents també mantenen una correlació positiva amb la temperatura i les hores de llum durant el dia, mentre que estan negativament correlacionats amb la velocitat del vent. Amb els resultats aportats a l'article, es pot observar que es presenta una correlació positiva entre les variables temperatura i hores de sol contra el nombre de casos d'accidents de trànsit, juntament amb una correlació negativa d'aquests amb la velocitat del vent. No es pot concloure res de la resta de correlacions a causa de no ser estadísticament significatives (P-value>0.05). Tot i que els P-value de les correlacions esmentades reflecteixen unes dades estadísticament significatives, els seus coeficients de correlació indiquen una molt baixa intensitat.

Per al present treball s'escull fer un estudi de caracterització dels accidents de trànsit, partint de les dades disponibles a través del servei de dades obertes Open Data BCN i aplicant tècniques de clustering. Es defineix d'aquesta manera ja que s'ha realitzat molt poques vegades un estudi d'aquest tipus, mai abans sobre la informació dels sinistres a la ciutat de Barcelona. Amb tècniques de clustering, que presenten uns resultats molt bons sobre aquest tipus de dades i es pot generar una informació molt rellevant i d'interès per reconèixer les situacions on es produeixen accidents de forma habitual i repetida a la ciutat.

2.2. Objectius del projecte

A partir de la literatura estudiada, la informació recollida i les especificacions definides per alinear el treball amb les línies de recerca de l'ETSEIB, es defineix l'objectiu d'identificar situacions, característiques i condicions freqüents en els accidents viaris ocorreguts a la ciutat de Barcelona.

La finalitat del projecte és generar informació que ajudi a definir polítiques adequades en la direcció de reduir en nombre i gravetat els accidents de trànsit a la ciutat. Això és possible si s'identifica característiques comuns entre un conjunt rellevant d'accidents que suposi un volum important dels sinistres ocorreguts. Un altre objectiu, de caràcter personal, és desenvolupar un projecte d'anàlisi de dades mitjançant una tècnica de Machine Learning com és el clustering, de gran valor en l'àmbit de l'enginyeria.

Abast

El projecte contempla la definició completa del procés de preparació del joc de dades corresponent als accidents de trànsit a la ciutat de Barcelona entre els anys 2018 i 2019 i l'aplicació d'algorismes de clustering sobre aquest, identificant situacions comuns entre accidents. El disseny dels algorismes emprats o el plantejament de mesures de contenció o mitigació estan fora de l'abast del projecte.

Revisada la divergència de conclusions sobre l'efecte de la densitat de circulació sobre el

nombre d'accidents, es decideix incloure l'estat del trànsit com a variable característica dels accidents a estudiar. A causa de les discrepàncies identificades a la literatura sobre el tema i les característiques de la ciutat, que ocupa poca superfície, situada al costat del mar i amb accidents geogràfics propers, al present treball també s'inclouen dades que reflecteixen la meteorologia.

Identificats els punts de partida, es fa una introducció al mètode seguit per desenvolupar el treball. En primer lloc, es fa una revisió de les tècniques de clustering existents i es selecciona les que encaixen i permeten aconseguir uns resultats en línia amb els objectius del treball. Després, es recullen les dades amb les que es treballa, es fa un anàlisi descriptiu per entendre com es distribueixen els accidents en aquestes i s'avalua la correlació i dependència entre variables, descartant les característiques que introdueixen informació molt semblant.

Quan es disposa d'un joc de dades uniforme i adequat per poder fer l'estudi, es passa a realitzar les proves definides. Com a punt de partida, s'estima el nombre de clústers que generen una millor agrupació de les dades. Tot i que el present treball no busca les millors de les agrupacions, es marca un primer nombre de grups adequat per a la agrupació en clústers. Finalment, es fa l'aplicació dels algorismes de clustering seleccionats a partir del mètode de prova definit i s'obtenen els resultats.

3. Conceptes i metodologia

En aquest capítol de la memòria es descriu la metodologia seguida detallant els procediments d'operació de l'estudi.

3.1. Clustering

S'identifica que una bona tècnica a aplicar a les dades que s'analitzen per tal de caracteritzar-ne situacions és el clustering. Aquest mètode d'agrupament s'emmarca en les tècniques de Machine Learning no supervisat. Al capítol es descriu el clustering i les tècniques seleccionades per aplicar a les dades d'accidents de Barcelona.

Actualment, existeixen dues alternatives per als algorismes de Machine Learning, aquests són els supervisats i els no supervisats [24]. Els algorismes de Machine Learning supervisat són aquells que aprenen a través de dades que han estat preparades per a entrenar al model, és a dir, es presenten exemples dels inputs aportats al model i els outputs desitjats i l'objectiu és aconseguir la regla general que relacioni els inputs amb els outputs. Els valors d'entrada han estat prèviament netejats, de manera que el model treballa amb el mateix tipus de dades estandarditzades i evidentment, correctament etiquetades i classificades. Existeixen els models de xarxes neuronals, models de reducció de dimensions, de classificació i de regressió [25].

Les tècniques de Machine Learning no supervisat fan referència a aquells algorismes en que l'aprenentatge es fa mitjançant dades amb elements no etiquetats, buscant relacions entre aquests. És un procés d'aprenentatge independent, que no treballa amb dades d'entrada que tenen assignada una dada de sortida, és a dir, els valors objectius són desconeguts o no estan identificats [26]. També, les dades que utilitza el model no es troben estructurades, sinó que tenen valors atípics i no han estat netejades.

Com que no es disposa de valors objectiu utilitzables per a construir un model lògic entre entrada i sortida, cal recórrer a diferents tècniques per extreure regles de dades i patrons entre elles per poder entendre millor les dades i trobar un resultat significatiu. Els algorismes de Machine Learning no supervisat es poden classificar en models de reducció de dimensions, de clusterització i d'associació [27]. Els models d'associació busquen definir el nivell de relació entre objectes. Això permet identificar forts lligams entre parelles d'objectes que comparteixen relació entre variables o característiques pròpies.

Els models de clusterització, identificats com una bona opció per realitzar una identificació de situacions entre els accidents analitzats al present treball, agrupen un conjunt d'objectes de tal manera que els del propi grup tenen característiques més similars entre ells que amb

els objectes que pertanyen a altres grups. Aquests models, generalment, es poden classificar en jeràrquics o particionals. Al present treball s'aplica un model de Machine Learning no supervisat, de clusterització i particional. Concretament, els mètodes més comuns són Gaussian Mixture Models (GMM), Hierarchical Agglomerative Clustering (HAC), Density-Based Spatial Clustering of Aplications with Noise (DBSCAN) i K-means, K-modes i K-prototypes.

Algorismes seleccionats

Es seleccionen els tres darrers algorismes, K-means, K-modes i K-prototypes, per ser àmpliament estudiats, amb gran quantitat de literatura que n'avala els resultats i que han estat aplicats amb bons resultats en estudis similars al present.

Els algorismes seleccionats són senzills i fàcilment aplicables, que els fan molt adequats per als objectius d'aquest treball. Consisteixen en tècniques iteratives que busquen dividir un conjunt de dades en K grups o clústers diferents, on cada mostra o punt del joc de dades pertany a un grup. A través de les iteracions, busquen que els punts siguin el més propers possibles en totes les seves dimensions a la resta d'elements del clúster, alhora que s'allunyen dels punts que conformen la resta de grups. K-means només treballa amb variables numèriques, K-modes amb variables categòriques i K-prototypes pot treballar tant amb variables numèriques com categòriques.

L'algorisme K-means assigna els punts a un grup minimitzant l'arrel quadrada de la suma de la diferència de distàncies al quadrat en totes les dimensions entre el punt o accident (X) i el centroide del clúster (Y), el que es coneix com la distància Euclidiana.

$$d(X,Y) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$$

On:

X i Y són els punts a comparar
m és el nombre d'atributs
x_{ik} és el valor de l'atribut *k* de l'element X
x_{jk} és el valor de l'atribut *k* de l'element Y

(*Eq. 3.1*)

S'assigna l'accident al clúster que presenta el valor d'aquesta distància més petita.

També s'utilitza l'algorisme anomenat K-Modes, adaptat per treballar amb variables categòriques. Es diferencia de l'algorisme K-means en que, mentre aquest aplica distàncies Euclidianes per definir la semblança dimensional entre punts, l'algorisme K-Modes utilitza modes per representar els centres dels clústers (els modes actuen com a centroides) i actualitza els modes amb les categories més freqüents al clúster per cada iteració realitzada. La mètrica utilitzada per mesurar la semblança es la distància de Hamming, com es mostra a continuació:

$$d(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \qquad (Eq.\ 3.2)$$

A on: $$\delta(x_j, y_j) = \begin{cases} 0, & x_j = y_j \\ 1, & x_j \neq y_j \end{cases} \qquad (Eq.\ 3.3)$$

Els valors "x_j" i "y_j" de les anteriors equacions són els valors de l'atribut "j" als accidents "X" i "Y". Com major sigui el nombre de discrepàncies de valors categòrics entre "X" i "Y" més diferents s'identifiquen els accidents. Respecte al conjunt de dades categòriques, la moda d'un atribut és 1 o 0, segons el que sigui més comú al grup. El vector de mode minimitza el sumatori de distàncies entre cada objecte del grup i el centre d'aquest.

L'algorisme de K-Modes, com la resta de mètodes de clustering, busca agrupar el joc de dades en un nombre K de clústers. El primer dels passos que es realitza és la definició de centroides, seleccionats inicialment de forma aleatòria. Fet això, agrupa les dades segons la seva distància al centroide més pròxim. Quan ja existeix una primera agrupació en K clústers, es fa una segona computació del centroide segons les mostres que formen el clúster i es revisa, per cada element, el seu centroide més pròxim. A través de les iteracions, es redistribueix els elements si identifica que la distància al centroide d'un clúster veí és més petita que la distància al centroide del propi clúster.

Per acabar, el darrer algorisme seleccionat per fer proves sobre el joc de dades és el K-prototypes. Aquest es descriu a l'article *Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values* [28]. Aquest és una evolució dels algorismes K-means i K-modes, de manera que el procés operatiu i els seus objectius són els mateixos. La diferència rau en el tipus de dades que pot processar: aquest permet generar agrupacions a partir d'un joc de dades mixt, amb variables numèriques i categòriques. Això ho fa a partir d'una mesura de diferència de coincidències per treballar amb les dimensions categòriques i substituint les mitjanes dels clústers per modes. A través d'un mètode basat en la freqüència, actualitza els modes en el procés d'operació per minimitzar la funció de cost de clustering. D'aquesta manera, pot agrupar els mètodes K-means (variables numèriques) i K-modes (variables categòriques) i treballar amb tots dos tipus de variables.

Transformacions sobre les dades

A partir dels tres algorismes de clustering seleccionats, s'executen els programes de càlcul que generen uns resultats. Per fer-ho, cal prèviament fer unes transformacions de dades, exclusives per a l'aplicació de cadascun dels algorismes de clustering. Tot seguit, es descriuen breument els passos previs que cal fer per poder aplicar cadascun dels

algorismes. És important comentar que són proves separades i que, per tant, les transformacions de dades es fan de forma independent i en paral·lel.

L'algorisme K-prototypes treballa amb dades mixtes, de manera que no cal fer transformacions de dades. L'únic pas de processament que cal fer sobre les dades és una estandarditzación sobre les numèriques. Donat que els rangs de dades numèriques generalment no són regulars, és a dir, la diferència entre el valor més baix i més alt que prenen les variables canvia segons aquesta, cal realitzar una transformació de normalització de les dades per evitar causar problemes d'interpretació i de processat quan s'agrupa aquestes dades per caracteritzar un accident. Es transforma els valors que prenen les variables a un interval que es mou entre 0 i 1, escalant els valors proporcionalment entre aquests.

K-means genera agrupacions a partir de variables exclusivament numèriques. Perquè l'algorisme pugui treballar correctament, ha de ser possible calcular distàncies espacials entre els diferents punts a agrupar, de manera que cal que les dimensions d'aquests siguin numèriques. Això planteja el problema que no és possible treballar amb variables categòriques, ja que aquestes no són dimensions numèriques. Cal realitzar la transformació de les dades categòriques en variables binàries, creant *N* binàries, on *N* són les categories que pot prendre una de les variables categòriques. Aquesta binària pren el valor *1* si l'accident presenta la categoria corresponent de la variable transformada en binàries i *0* si no la presenta. D'aquesta manera, s'obté un joc de dades sense variables no numèriques i sobre el que es pot calcular distàncies entre totes les seves dimensions. En acabat, també és recomanat fer una estandarditzación de totes les variables com s'explica a l'anterior paràgraf.

Finalment, per poder generar situacions mitjançant l'algorisme K-modes cal transformar totes les variables a categòriques, de manera que cal fer l'agrupació de les numèriques en diferents categories, segons el nombre d'accidents que presenten aquestes. Generalment, si es disposa d'una gran quantitat de categories en una de les variables, és interessant fer una compactació d'aquestes per aconseguir grups de variables representatius i evitar l'elevada quantitat de categories.

3.2. Etapes

Metodologia d'anàlisi

Es presenta la metodologia dissenyada per dur a terme l'anàlisi de clustering dels accidents de trànsit, distingint entre les diferents etapes definides. Aquesta metodologia es confecciona a partir dels articles i treballs anteriors revisats comentats a la introducció.

Anàlisi de les dades i definició de variables

En aquesta etapa del treball es realitza les següents tasques:

- Definició dels orígens i adequació i transformació de les dades per tal d'aconseguir un joc de dades únic sobre el que aplicar els algorismes de clustering.
- Selecció final de variables utilitzades, mitjançant un anàlisi de correlació entre característiques numèriques i de dependència entre les variables categòriques. En aquest pas es descarta les que presenten una dependència o que aporten informació ja inclosa al joc de dades.
- També es realitza un anàlisi previ al clustering sobre les dades mantingudes.
- Transformacions aplicades a les dades per adaptar-les a les diferents proves a executar i la definició de les variables definitives incloses als resultats, a través de les proves realitzades.

Obtenció dels clústers

En segon lloc, realitzades les proves sobre el joc de dades complet i sobre el subconjunt de dades format pels accidents més greus, es presenta l'obtenció dels clústers d'accidents.

- Descripció de les situacions identificades a través de les agrupacions obtingudes.
- Presentació de resultats intermedis d'interès, ja que caracteritzen segons variables no incloses als resultats finals.

Tot seguit es detalla els procediments seguits per tal de dur a terme la metodologia plantejada.

Una característica que dona més valor al projecte és estudiar el més gran nombre d'accidents possibles, per poder obtenir uns resultats més representatius de la realitat observada. Tot i això, s'opta per incloure els anys a partir del 2018, ja que la densitat de trànsit es va començar a recollir a partir de l'octubre de 2017. Igualment, alguns arxius de dades no disposen de les mateixes variables que els anys posteriors, ja que s'ha anat

afegint dades capturades al llarg del pas del temps. També s'identifica aquesta casuística entre els anys estudiats, ja que de l'origen de dades de descripció de les persones involucrades en accidents, les dades de l'any 2019 contenen la motivació de desplaçament del vianant i del conductor, mentre que les de 2018 no contenen aquesta informació.

També es decideix tancar la finestra d'estudi l'any 2019 ja que el 2020 va aparèixer l'excepcional situació social i econòmica derivada de la pandèmia global del Covid19. Aquesta ha fet aparèixer unes condicions, clarament transitòries, que esbiaixen les dades recollides. Assumint que no s'hauria de repetir de forma habitual en el futur no té justificació ni sentit afegir aquestes dades a l'estudi.

Revisió de les dades

En aquest treball es parteix de la informació recollida al servei de dades obertes Open Data BCN [29], relacionada amb els accidents i els actors implicats. Les dades són registrades per part de la Guardia Urbana en el moment de fer la mediació al lloc de l'accident.

Com es comenta i es discuteix a la introducció, també s'inclou l'estat del trànsit corresponent al moment de produir-se l'accident. Les dades són recollides mitjançant 483 sensors instal·lats sota l'asfalt de trams de circulació de tot tipus, que mesuren variacions del camp magnètic provocades pel pas dels vehicles. També s'utilitzen sensors amb tecnologia infraroja i càmeres amb tractament d'imatge per recollir l'estat de la circulació. Al web Open Data BCN, es recull la informació de cada estació detectora i aquesta es tradueix en l'estat de la circulació del tram. Les dades es recullen en un arxiu corresponent a cada mes de l'any, de manera que per efectuar el present estudi es treballa amb un total de 24 arxius de dades.

Ja que es disposa d'un gran nombre de mostres repartides geogràficament per la ciutat i en tot tipus de vies de circulació, es pot definir l'estat del trànsit de forma general. Certament, no és la manera més encertada d'associar l'estat de circulació amb l'accident, ja que no deixa de ser una mitjana. Tot i això, sí que permet identificar en línies generals l'estat de congestió de la ciutat.

Finalment, també s'incorpora les dades meteorològiques a l'estudi, que s'inclouen tal com es discuteix a la introducció. La font de dades d'aquestes correspon al portal de Dades Obertes de Catalunya, que és una biblioteca d'informació. Concretament, s'utilitzen les dades registrades a l'estació meteorològica del Raval del Servei Meteorològic de Catalunya.

Sobre les dades recollides, s'estudia el sentit i la seva validesa i es descarta aquelles que no són útils. Es duu a terme un anàlisi descriptiu de les variables mantingudes, distingint entre les variables numèriques i categòriques. Es fa a través de taules i gràfics de relació per tal

de veure'n característiques evidents, alhora que amb la distribució del volum d'accidents o la seva gravetat a través de la variable. Amb aquest anàlisi descriptiu preliminar es poden identificar variables categòriques que presenten un volum elevat de valors a prendre i que posteriorment s'avalua la seva possible agrupació en categories més compactes.

Després de realitzar la descripció de les dades, es desenvolupa l'anàlisi de correlació de les variables numèriques. Sobre aquestes s'empren dos mètodes: els anàlisis de correlació de Pearson i de Spearman. L'anàlisi de Pearson avalua la correlació lineal entre dues variables quantitatives i és independent de l'escala de mesura d'aquestes. Pearson permet determinar el grau de relació de dues variables, sempre que aquestes també siguin quantitatives.

El coeficient de correlació de Spearman és una mesura de correlació entre dues variables aleatòries, tant contínues com discretes. El mètode Spearman és menys sensible per als valors més llunyans de l'esperat. Ambdues mesures de correlació oscil·len entre els valors (-1, +1), indicant associacions negatives o positives respectivament. El valor mig, el zero, indica que no hi ha correlació entre variables. La interpretació de les correlacions entre variables sempre és relatiu al camp d'estudi, les dades específiques i els objectius del treball. Al present treball es segueix la guia definida a l'article *Statistical Power Analysis for the Behavioral Sciences,* de J. Cohen [29].

D'aquest, se'n desprèn el següent criteri, on es consideren els següents límits per determinar la intensitat de correlació entre una parella de variables:

	Baix	*Mitjà*	*Alt*
Grau de correlació	*0.10 < x < 0.30*	*0.30 < x < 0.50*	*x ≥ 0.50*
	-0.10 > x > -0.30	*-0.30 > x > -0.50*	*x ≤ -0.50*

Taula 3-1. Criteris de valoració de les correlacions entre variables [29].

Es considera descartar una de les variables de la parella en cas que el grau de correlació entre elles sigui *alt*.

Realitzada la identificació de la correlació entre variables numèriques és interessant identificar aquelles variables categòriques que inclouen informació molt similar al joc de dades. Per fer-ho, es proposa el test de Cramer V, ja que és un test que dona informació sobre com de significant és la dependència entre dues variables categòriques.

Analitzant les parelles de variables juntament amb el P-value corresponent, s'identifica aquelles variables que poden ser descartades del model. De forma general, si el P-Value d'una parella és inferior a 0.05 (significa que hi hauria una possibilitat per sota del 5% d'arribar al resultat de no haver-hi dependència real) es considera que el coeficient és

estadísticament rellevant i es rebutja la hipòtesi nul·la.

Aquest llindar està establert com a conveni de forma transversal a la comunitat acadèmica, tot i que pot ser modificat si l'estudi i els seus requeriments així ho necessiten. En aquest cas, es manté el conveni com a criteri de rellevància de la correlació.

Identificació del número de clústers

Feta la definició del joc de dades a utilitzar i descartades les variables que presenten correlació o dependència, es passa a fer un estudi sobre els accidents i les seves característiques. Per tal d'estudiar un joc de dades i fer una primera aproximació al clustering, habitualment s'empren diferents tècniques d'agrupació, reducció dimensional.

També cal fer una definició del nombre de clústers òptims. Tot i que al present treball no es busca aconseguir la millor agrupació sinó identificar situacions dels accidents de trànsit, és un primer anàlisi que permet entendre la proximitat entre les dades i definir un punt de partida per a l'explotació dels algorismes utilitzats posteriorment. Al present apartat es presenta alguns dels mètodes més comuns.

Factor Analysis of Mixed Data (FAMD)

El primer dels anàlisis es tracta d'un Factor Analysis of Mixed Data, d'ara endavant FAMD. Aquest mètode exploratori, basat en la reducció dimensional mitjançant Principal Components, permet treballar amb tipologies mixtes de dades, tant numèriques com categòriques balancejant la seva influència sobre l'anàlisi. Aquesta reducció de dimensions permet estudiar les similituds entre accidents de forma més fàcil i obrint la porta a una representació gràfica dels resultats, distingint-los segons la seva gravetat i fent-ne la representació en dues dimensions [31].

De representar de manera significativa el joc de dades, aquesta reducció dimensional suposa una major fluïdesa de treball i poder realitzar representacions en poques dimensions dels resultats obtinguts.

Test de Silhouette

Aquest test es refereix a un mètode de validació de l'agrupació de dades en clústers. El present mètode proposa una representació gràfica de la qualitat de l'agrupació de cada objecte del joc de dades. S'utilitza per poder definir el nombre de clústers que minimitzen els errors de classificació, comparant els resultats de les diferents agrupacions. El valor del test defineix en rangs, que es mouen entre -1 i +1. Per cadascun dels objectes en calcula la seva distància al centroide del clúster al que es classifica i la distància al centroide més proper d'entre els que no pertany, en totes les dimensions. Si la nota de l'objecte és negativa, indica

que aquest es troba més proper del segon centroide que no pas del que pertany, indicant una classificació equivocada. Finalment, es realitza una mitjana de les puntuacions de tots els objectes i s'obté la nota de Silhouette per aquella agrupació [32].

Test del colze

Per tal de confirmar els resultats obtinguts mitjançant el test de Silhouette, habitualment es realitza un Elbow Test sobre el joc de dades. Aquest test planteja la representació gràfica del sumatori quadrat de les distàncies entre els punts i el seu centroide, per a un nombre de clústers determinat. Fent el càlcul per a diversos conjunts de clústers, s'identifica a la representació un punt on l'augment del nombre de clústers ja no millora significativament els resultats obtinguts [33].

Selecció de variables

Amb el joc de dades que es deriva dels anàlisi de correlació i dependència i estudiades les dades, es pot passar a realitzar proves aplicant clustering i obtenint uns resultats. En aquest punt, cal realitzar les transformacions de dades que permeten realitzar els càlculs corresponents a cada mètode de clustering seleccionat. Fet això i partint de cadascun dels tres jocs de dades obtinguts en aquesta transformació, es realitzen dues execucions inicials de l'algorisme, agrupant les dades en 6 i 20 clústers. Aquests valors parteixen de la identificació del nombre de clústers realitzada. L'objectiu és poder identificar en les situacions extrem, aquelles variables que prenen el mateix valor en tots els grups. No té sentit mantenir variables que no canvien al llarg dels clústers, ja que aleshores deixa de ser una variable identificadora. D'aquesta manera, si en una iteració es detecta aquesta situació en totes dues agrupacions, es descarta la variable i es torna a fer la classificació per veure'n els canvis. El diagrama d'operació durant la fase de proves és el següent:

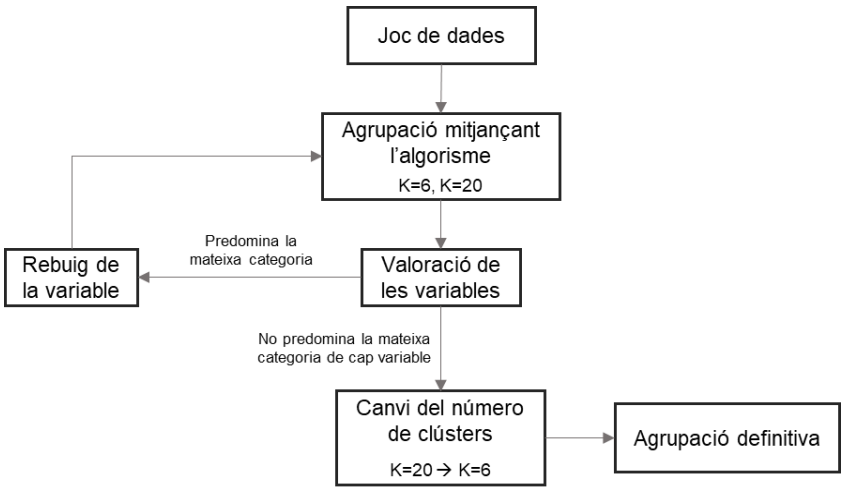


Figura 3-1. Diagrama d'operació per l'execució de proves.

En acabar les iteracions de la prova es disposa del grup de variables definitives que aporten informació sobre els accidents classificats. Donat que l'objectiu del treball no és aconseguir les millors agrupacions del conjunt d'accidents o identificar-ne la causalitat, sinó que el que es busca és identificar situacions que es repeteixen, es fa iteracions sobre els jocs de dades partint de les variables definitives i modificant el nombre de clústers entre 6 i 20 grups.

L'augment del nombre d'agrupacions fa que apareguin accidents mal classificats, però també mostra situacions menys habituals que no es veuen reflectides als clústers òptims, com aquelles que involucren pocs accidents o condicions especials. Per seleccionar el nombre de clústers definitiu es revisen les 15 proves fetes per cada joc de dades i es selecciona aquella que presenta unes agrupacions que no es repeteixin, que tinguin un nombre rellevant d'accidents, que tinguin múltiples categories preses per les variables i que tingui el nombre de clústers més baix possible. Es considera aquests resultats com el producte de l'execució de la prova. Per cadascuna de les tres execucions de proves es revisa el conjunt de resultats no definitius per identificar possibles situacions especials que poder tenir interès en ser comentades.

3.3. Software i llibreries

La definició del mètode de treball va acompanyada per la definició de les eines a utilitzar per tal de realitzar tota la tasca de recull de dades, adequació i transformació d'aquestes, anàlisi descriptiu, correlació i dependència. També l'anàlisi previ al clustering i aplicació del mètode de prova, juntament amb la recuperació i tractament dels resultats obtinguts.

Les anteriors activitats que conformen l'estructura vertebral del projecte es traslladen al llenguatge de programació Python per ser executats a la PaaS (Platform As A Service) Google Colab. Aquesta ofereix uns serveis de computació gratuïts per a l'usuari, amb la possibilitat de compartir el codi en tot moment amb col·laboradors, elaborar-lo en petites franges i mantenir-lo emmagatzemat al núvol, minimitzant el risc de pèrdua. Es pot accedir i recuperar el codi utilitzat per a l'elaboració del present treball a través del següent vincle [34].

Treballant en Python a través de Google Colab, s'ha utilitzat les corresponents llibreries *kmeans*, *kmodes* i *kprototypes* per a poder executar els diferents algorismes utilitzats, K-means, K-modes i K-Prototypes. També és necessari utilitzar la llibreria *NumPy*, que permet realitzar operacions matemàtiques en taules.

4. Anàlisi de les dades i definició de variables

Seguint la metodologia definida, a continuació es desenvolupa la recollida de dades, anàlisi i definició de variables incloses al joc d'accidents.

En primer lloc, es defineixen els orígens de dades i es presenta l'adequació d'aquestes, juntament amb l'anàlisi descriptiu i de correlació i dependència entre variables.

4.1. Orígens de les dades

Tot seguit es mostra una taula amb el detall dels atributs recuperats dels orígens de dades.

En primer lloc, hi ha un conjunt de característiques transversals dels accidents que comparteixen les cinc fonts de dades. Després, es detalla les variables úniques obtingudes de cada origen.

	Font d'origen	Arxiu de dades	Número de variables
Origen 1	Open Data BCN. Servei de dades obertes de l'Ajuntament de Barcelona [29].	Descripció de la causalitat dels accidents gestionats per la Guàrdia Urbana a la ciutat de Barcelona	21
Origen 2		Persones involucrades en accidents gestionats per la Guàrdia Urbana a la ciutat de Barcelona	8
Origen 3		Accidents gestionats per la Guàrdia Urbana a la ciutat de Barcelona segons tipologia	1
Origen 4		Vehicles implicats en accidents gestionats per la Guàrdia Urbana a la ciutat de Barcelona	6
Origen 5		Accidents gestionats per la Guàrdia Urbana a la ciutat de Barcelona	5
Origen 6		Informació sobre l'estat del trànsit als trams de la ciutat de Barcelona	4
Origen 7	Dades meteorològiques de la XEMA. Dades Obertes Catalunya [35].	Característiques meteorològiques de la ciutat de Barcelona	13

Taula 4-1. Característiques recollides dels repositoris d'informació.

Cadascun dels orígens de dades aporta dades úniques sobre l'accident. Els orígens de la informació descrits en aquest apartat són els que es treballen per elaborar l'arxiu únic de dades que conté totes aquelles variables i accidents utilitzats al present treball.

Dades incloses i dades no incloses

Per tal de disposar d'un joc de dades de partida per poder fer l'estudi cal agrupar i treballar les dades recollides, seleccionades d'acord amb els objectius de l'estudi. En aquest punt, els arxius de dades contenen tota la informació recollida, incloent les característiques comuns. A continuació, s'enumeren les dades que es recuperen del total d'arxius i les que no s'utilitzen. En segon lloc, es comenta la tasca d'adequació realitzada sobre les dades, tot comentant les variables mantingudes.

És notable que els arxius utilitzats tenen intervals de temps diferents: els arxius dels accidents estan agrupats per anys, mentre que els arxius del trànsit tenen un interval mensual. Les dades dels arxius orígens de dades que es mantenen per tal de fer-ne transformacions són les següents:

	Nomenclatura	**Descripció**	**Origen**
Dades mantingudes	*Numero_expedient*	Codi de registre de l'accident	Origen 1
	Districte	Nom del districte	
	Barri	Nom del barri	
	Dia_setmana	Abreviació del dia de la setmana	
	Descripcio_tipus_dia	Descriu si el dia és laborable	
	Any	Any quan succeeix l'accident	
	Nom_mes	Mes quan succeeix l'accident	
	Dia_mes	Dia quan succeeix l'accident	
	Hora_dia	Hora quan succeeix l'accident	
	Descripcio_causa_mediata	Motiu pel que es fa la mediació	
	Desc_Tipus_vehicle_implicat	Descriu la tipologia del vehicle	Origen 2
	Descripcio_sexe	Sexe de la víctima descrita	
	Edat	Edat de la víctima descrita	
	Descripcio_tipus_persona	Relació amb l'accident	
	Descripcio_situacio	Lloc on es produeix l'accident	
	Descripcio_victimitzacio	Gravetat de la lesió	
	Tipus_accident	Forma com s'ha produït l'accident	Origen 3
	Descripcio_causa_vianant	Identifica si l'accident és causa del vianant	Origen 4
	Descripcio_color	Color del vehicle implicat	
	Descripcio_carnet	Carnet de conduir del conductor del vehicle implicat	
	Antiguitat_carnet	Antiguitat del carnet de conduir del conductor del vehicle implicat	
	Numero_morts	Nombre de morts en l'accident	Origen 5
	Numero_lesionats_lleus	Nombre de lesionats greus en l'accident	
	Numero_lesionats_greus	Nombre de lesionats lleus en l'accident	
	Numero_victimes	Nombre de víctimes en l'accident	

	Nomenclatura	Descripció	Origen
Dades mantingudes	*Numero_vehicles_implicats*	Nombre de vehicles implicats en l'accident	
	idTram	Tram del que se'n pren la mostra	Origen 6
	data	Data del moment de presa de la mostra	
	estatActual	Estat del trànsit al moment de prendre la mostra	
	01.CODI_ESTACIO	Codi de l'Estació Meteorològica	Origen 7
	02.DATA_LECTURA	Data del registre de la mesura	
	03.TM	Temperatura mitjana [*ºC*]	
	04.TX	Temperatura màxima [*ºC*]	
	05.TN	Temperatura mínima [*ºC*]	
	06.HRM	Humitat relativa mitjana [*%*]	
	07.PPT24H	Precipitació acumulada [*mm*]	
	08.HPA	Pressió atmosfèrica mitjana (valor no corregit a nivell del mar) [*hPa*]	
	09.RS24H	Irradiació solar global [MJ/m^2]	
	10.VVM10	Velocitat del vent a 10 m [*m/s*]	
	11.DVM10	Direcció del vent a 10 m [º]	
	12.VVX10	Ratxa màxima del vent 10 m [*m/s*]	
	13.DVX10	Direcció de la ratxa màxima de vent a 10 m [º]	

Taula 4-2. Dades mantingudes dels orígens de dades.

Les dades dels arxius orígens descartades es mostren a la següent taula. Algunes de les dades repeteixen informació respecte de les mantingudes (*nom_districte*, *descripcio_dia_setmana*, *mes*, *descripció_torn*), d'altres no són transversals durant els dos anys dels que s'estudia dades (*descripcio_motiu_desplaçament_vianant* i *descripcio_motiu_desplaçament_conductor*).

Les restants són dades que, de no aplicar-hi cap transformació o informació complementaria, com que indiquen localitzacions geogràfiques o simplement el nom d'aquesta localització, no són útils (*codi_barri*, *codi_carrer, nom_carrer, num_postal, coordenada_UTM_X, coordenada_UTM_Y, longitud i latitud*).

	Nomenclatura	Descripció
Dades no mantingudes	*Nom_districte*	Nom del districte del lloc de l'accident
	Codi_barri	Codi del barri del lloc de l'accident
	Codi_carrer	Codi del carrer del lloc de l'accident
	Nom_carrer	Nom del carrer del lloc de l'accident
	Num_postal	Nombre postal del lloc de l'accident
	Descripcio_dia_setmana	Nom del dia de la setmana quan succeeix l'accident
	Descripcio_torn	Torn del dia quan succeeix l'accident

	Nomenclatura	Descripció
Dades no mantingudes	*Mes*	Mes de l'any quan succeeix l'accident
	Descripcio_motiu_desplaçament _vianant	Motiu del desplaçament del vianant
	Descripcio_motiu_desplaçament _conductor	Motiu del desplaçament del conductor
	Descripcio_model	Model del vehicle implicat
	Descripcio_marca	Marca del vehicle implicat
	Coordenada_UTM_X	Coordenada UTM X del lloc de l'accident
	Coordenada_UTM_Y	Coordenada UTM Y del lloc de l'accident
	Longitud	Longitud cartogràfica del lloc de l'accident
	Latitud	Latitud cartogràfica del lloc de l'accident

Taula 4-3. Dades no seleccionades dels orígens de dades.

També es descarten les dades de *Descripcio_model* i *Descripcio_marca*. Aquesta decisió es pren principalment per la gran quantitat d'informació de que disposen, ja es té una descripció del tipus de vehicle i les condicions de l'accident. Junt amb això, la marca del vehicle no permet caracteritzar l'accident en el que està involucrat, ja que les marques tenen un rang molt ample de característiques dins els productes que ofereixen com es pot apreciar a [Taula 4-4]. En aquesta, es recull un conjunt de característiques que il·lustren la decisió d'eliminar dites variables.

Característica	Unitats	Valor més baix	Valor més alt
Potència del motor	[*CV*]	80 (Seat Ibiza)	310 (Seat León Cupra R)
Acceleració màxima (0 – 100 km/h)	[*s*]	15,3 (Seat Ibiza)	5,2 (Seat León Cupra R)
Llargària	[*m*]	3,55 (Seat Mii)	4,73 (Seat Tarraco)
Preu	*[€]*	13.400 (Seat Ibiza)	53.495 (Seat León Cupra R)
Pes	[*kg*]	933 (Seat Mii)	1868 (Seat Tarraco)
Ocupants	[*persones*]	4 (Seat Mii)	7 (Seat Tarraco)

Taula 4-4. Rang de característiques dels vehicles SEAT.

Amb la variable *Descripcio_model* passa que hi ha 2596 valors únics. Alguns d'aquests fan referència al mateix model, però caldria dedicar uns esforços massa elevats per fer la tasca manual de revisió de cadascun d'aquests, ja que són dades introduïdes manualment per l'agent corresponent. A més a més, la variable pren massa valors i amb aquests no se'n pot extreure informació.

En tot cas, es podria relacionar cada model i fabricant amb les característiques de potència, velocitat màxima, acceleració màxima i d'altres del vehicle, però no es disposa de la

informació suficient (en molts dels casos no s'indica el model exacte, per exemple Volkswagen passat 2.0 tdi 140cv advance bluemotion 2001).

També, els rangs de les característiques no són propis de cada marca, sinó que el ventall és prou ample a totes les marques perquè es comparteixin valors. És per això que, juntament amb el recull de característiques de [Taula 4-4], es determina que no es pot extreure cap conclusió de la variable *descripcio_marca* ni *descripcio_model* del vehicle i es descarten.

4.2. Adequació de les dades

Al llarg dels apartats compresos en aquest capítol es detalla la tasca de modificacions desenvolupada sobre les dades recollides per tal de poder avançar amb una sola taula d'accidents amb el conjunt de característiques de cadascun.

Dades d'accidents

En primer lloc, el que es fa és agrupar, per a cada font de dades, els dos arxius que corresponen a 2018 i 2019 en una mateixa taula. Aquesta agrupació ens permet disposar de cinc taules úniques on es guarda la informació referent als accidents. Per aconseguir aquesta agrupació es realitza una tasca manual de revisió del les fonts de dades, ja que al ser obtingudes durant un període de dos anys, alguns dels valors que prenen i la forma d'etiquetar-les canvia. Es modifica aquests aspectes sobre els arxius font per facilitat i rapidesa.

Més enllà d'aquestes modificacions preliminars, sobre les fonts de dades 1, 3 i 5 no s'hi realitzen transformacions i contenen, a més a més de les dades que relacionen tots els jocs de dades (*numero_expedient*) aquelles dades específiques seleccionades de cada una de les fonts de dades. A continuació es descriu les variables recollides de cada origen i es detalla les tasques d'adequació realitzades.

Origen 1

La taula de dades generada a través del primer origen de dades conté les dades comuns entre arxius (variables comuns), la variable única inclosa i l'índex utilitzat per relacionar els diferents arxius amb la resta de variables específiques.

La variable única inclosa, *descripcio_causa_mediata*, presenta les dades que indiquen la causa per la que els agents de la Guàrdia Urbana fan la mediació. Aquesta variable pren els valors següents:

No hi ha causa mediata, alcoholèmia, objectes o animals a la calçada, excés de velocitat o inadequada, drogues o medicaments, calçada en mal estat, factors meteorològics, estat de la senyalització.

Origen 2

En relació a la taula producte de la font de dades 2, disposa de l'índex i de les específiques següents:

- *desc_tipus_vehicle_implicat,* pren valors que descriuen el tipus de vehicle de que es tracta, presentant vint-i-dues formes diferents.
- *Descripcio_sexe* i *edat,* que descriuen el sexe i l'edat de la persona relacionada amb l'accident. Hi pot haver més d'una persona relacionada, sigui conductor, vianant o passatger. En relació a aquestes:

  1. S'assigna a l'accident, sempre que és possible, l'edat i el sexe del conductor.

  2. De no disposar de les dades sobre algun accident, es defineix la variable com a *Desconegut* per a aquell*.*

- *Descripcio_tipus_persona* presenta la relació de la víctima amb l'accident, segons si era conductor o No conductor.

Com es comenta amb les variables *Descripcio_sexe* i *edat,* sempre que és possible s'assigna un dels conductors implicats en l'accident. De no disposar-ne de les dades, es determina la categoria de la variable com a *No conductor.*

- *descripcio_situacio,* dades que descriuen d'onze maneres diferents l'element o espai de les vies circulatòries on s'ha produït l'accident. Per exemple, *fora del pas*, *a la vorera* o *en pas regulat per semàfor*, entre d'altres.
- *descripcio_victimitzacio*, variable sobre la que es realitza una transformació de les dades. Les categories que prenen les dades són:

Desconegut, es desconeix, ferit lleu: rebutja assistència sanitària, ferit lleu: Amb assistència sanitària en lloc d'accident, ferit lleu: hospitalització fins a 24h, ferit greu: hospitalització superior a 24h, mort.

La transformació feta agrupa els diferents valors que pren la variable i els agrupa en les següents categories:

1. Lleu, 2. Assistencia, 3. Hospitalitzat, 4. Greu, 5. Mort

La taula de dades descriu totes aquelles persones involucrades en un accident, de manera que la relació de les mostres d'aquesta font amb el nombre d'accidents estudiats no és d'un a un. En molts dels casos existeix més d'una persona implicada en l'accident. Aquesta variable identifica el nivell de gravetat de l'accident sobre la persona involucrada. Donat que en l'estudi es treballarà sense repetició d'accidents, cal reduir les dades de l'arxiu de dades 2 a una persona per accident.

Per fer-ho, es basa en el que es fa en estudis anteriors, com *Analysis of road traffic fatal accidents using data mining techniques* [14], i es defineix la gravetat de l'accident a partir de les conseqüències més greus d'entre aquelles víctimes involucrades en l'accident, ja sigui en condició de conductor, passatger o vianant.

Origen 3

La taula de dades de la font 3 conté, igual que les anteriors, les dades que s'utilitzaran per relacionar els diferents arxius (*numero_expedient*) i les dades específiques, en aquest cas només una:

- *tipus_accident*. Aquestes informen sobre com ha estat l'accident i per tant, com ha estat ferida la víctima. Alguns exemples en son *abast múltiple, caiguda (dues rodes)* o *sortida de via amb atropellament*.

Origen 4

La taula creada a partir de la font de dades 4 conté informació sobre els vehicles implicats i, igual que amb la taula que conté les dades de la font de dades 2, en alguns casos hi ha més d'un vehicle implicat en els accidents.

- *multiples_vehicles_implicats*: es procedeix afegint una nova variable on es defineix si l'accident té múltiples vehicles implicats (*multiples_vehicles_implicats*, prenent la categoria *true* quan l'accident implica a 2 o més vehicles, *false* si només n'implica un).

També, per a cada accident es descarten tots els vehicles implicats, menys un. Idealment, es buscaria mantenir com a vehicle implicat en l'accident aquell en el que circulava la víctima que ha patit les conseqüències més greus, però els jocs de dades de que es disposa no permeten establir aquesta relació entre víctima i vehicle implicat.

També seria interessant disposar d'aquell vehicle que hagi patit les conseqüències materials més severes, però tampoc es disposa d'aquesta informació. D'aquesta manera, davant la falta d'opcions millors, es manté un dels vehicles implicats en l'accident seleccionant-lo de manera aleatòria i descartant la resta.

També es mantenen les següents dades:

- *numero_expedient*, per relacionar-se amb la resta de dades.
- *descripcio_causa_vianant*, que pren les categories *no és causa del vianant, desobeir altres senyals, creuar per fora pas de vianants, altres, desobeir el senyal del semàfor, transitar a peu per la calçada*.

Les resta de dades mantingudes són purament descriptives del vehicle, el carnet del

conductor i l'antiguitat d'aquest:

Descripcio_model, descripcio_marca, descripcio_color, descripcio_carnet i antiguitat_carnet.

Origen 5

A continuació, s'adreça les dades obtingudes de la font 5 de dades. Aquestes, sobre la que no es fa cap transformació, manté les dades *numero_expedient* i aporta una llista enumerativa dels elements relacionats amb l'accident: vehicles i nombre de víctimes i gravetat:
Numero_morts, numero_lesionats_lleus, numero_lesionats_greus, numero_victimes, numero_vehicles_implicats.

Agrupació dels arxius de dades

En aquest punt, es pot agrupar les dades de totes cinc taules en una de sola, emprant *numero_expedient* que relaciona tots els jocs de dades. Es disposa d'una taula amb les dades dels accidents descrites durant el capítol.

S'afegeix a la taula de dades generada dues variables més, *torn* i *festiu* per aportar informació complementària als accidents.

torn substitueix la variable categòrica que és la hora del dia, que presenta vint-i-quatre categories al llarg del dia. Es planteja una variable amb sis categories per tal de classificar els accidents durant el dia, agrupant-los en franges en que una part important dels desplaçament tenen un denominador comú.

La classificació es fa de la següent manera, en cursiva les categories que pren la variable:

- 02h a 05h (*mitjanit*)
- 06h a 09h (*matí*)
- 10h a 13h (*migdia*)
- 14h a 17h (*tarda*)
- 18h a 21h (*vespre*)
- 22h a 01h (*nit*)

La classificació representa les franges horàries en que els desplaçaments tenen una finalitat generalment igual, és a dir, de 6h a 9h els desplaçaments són en una majoria, per anar al lloc de feina. En canvi, els desplaçaments de mitjanit són deguts a gent que es desplaça entre un lloc d'esbarjo i lleure i casa seva. Així, aquesta divisió de les hores del dia és la que millor s'adapta al cicle de vida de les persones que fan vida a la ciutat i agrupa els desplaçaments similars en una mateixa franja.

La variable *festiu* compleix una finalitat similar a la de la variable *torn*, que és compactar un seguit de dades en una variable categòrica de pocs valors. En aquest cas, substitueix la dada *descripcio_dia_setmana* i ho fa en la seva condició de *festiu* o *no festiu*.

Es classifica cada tram segons si correspon a un dia festiu o no, entenent per festiu els cap de setmana (dissabte i diumenge) i els dies festius de l'any. A més a més, donat que es coneix l'activitat de la ciutat, es pot afirmar que la nit abans d'un dia festiu té més activitat de desplaçaments i el motiu d'aquests encaixa molt més amb un dia festiu. D'igual manera, la nit d'un dia festiu previ a un dia laborable té moltes semblances amb la nit d'un dia laborable. Per aquesta raó, es defineix:

- *festiu*: el torn de nit d'un dia laborable previ a un dia festiu.
- *No festiu*: el torn de nit d'un dia festiu previ a un dia laborable.

Per acabar, es comenta les transformacions fetes sobre les dades meteorològiques obtingudes.

Dades sobre l'estat del trànsit

Tot seguit, es treballen les fonts de dades de trànsit. En primer lloc, els sensors de mesura retornen els valors següents, que representen l'estat de la circulació:

0 = sense dades, *1* = molt fluid, *2* = fluid, *3* = dens, *4* = molt dens, *5* = congestió, *6* = tallat

La primera transformació aplicada consisteix en crear un calendari a cadascuna de les taules mensuals de dades, per tal d'identificar l'any, mes, dia i hora de la mostra. Si l'estat de la mesura és *6* = tallat, es canvia el seu valor a zero. Més endavant se'n dona les raons. Per cada arxiu de dades mensual, es fa:

1. Transformació de l'estat de mesura *6* = tallat a zero, per evitar que al fer la mitjana tingui un efecte multiplicador de l'estat sense representar la circulació. Una via tallada no es pot circular.
2. Càlcul de la densitat mitjana en cada període d'una hora i per cada estació, sempre que l'estat sigui diferent de zero.
3. Agrupació de les densitats calculades de les diferents estacions disponibles, fent la mitjana d'aquestes.

En acabat, de cada arxiu mensual se'n obté una taula, amb les dades de calendari, nombre d'observacions computades i estat mitjà de la circulació mesurat a la ciutat. S'agrupen les dades calculades mes a mes en una sola taula. Per tal d'associar la taula d'accidents amb l'estat del trànsit a través de les hores del dia, s'utilitza una nova variable auxiliar i no

permanent a la taula d'accidents (*Calendari*) obtinguda a partir de la concatenació de les dades *any*, *mes_any, dia_mes, hora_dia* [Taula 4-2].

Dades meteorològiques

Es disposa d'un arxiu únic que conté totes les dades recollides de la Font de dades 7 [Taula 4-2]. Es descarta les variables *01.CODI_ESTACIO*, ja que només es recullen dades d'una estació. Per tant, només cal associar cada accident amb les característiques de la meteorologia que li corresponen, segons el seu tram temporal.

Per fer això, cal fer una petita transformació de la variable *02.DATA_LECTURA* per tal d'adequar-ne el format i a través de la variable auxiliar *Calendari*, que dona una visió de la hora i el dia que s'ha registrat l'accident i permetrà agrupar les dades meteorològiques amb cada accident. Es descarta la variable auxiliar utilitzada.

En aquest punt es disposa d'una taula única, generada a partir dels set diferents orígens de dades, que conté les dades següents:

	Nomenclatura	**Descripció**
Dades mantingudes	*Numero_expedient*	Codi de registre de l'accident
	Districte	Codi del districte
	Barri	Nom del barri
	Any	Abreviació del dia de la setmana
	Nom_mes	Nom del mes quan s'ha produït l'accident
	Descripcio_causa_mediata	Motiu pel que es fa la mediació
	Desc_Tipus_vehicle_implicat	Descriu la tipologia del vehicle
	Descripcio_sexe	Sexe de la víctima descrita
	Edat	Edat de la víctima descrita
	Descripcio_tipus_persona	Relació amb l'accident
	Descripcio_situacio	Lloc on es produeix l'accident
	Descripcio_victimitzacio	Gravetat de la lesió
	Tipus_accident	Forma com s'ha produït l'accident
	Descripcio_causa_vianant	Identifica si l'accident és causa del vianant
	Descripcio_color	Color del vehicle implicat
	Descripcio_carnet	Carnet de conduir del conductor del vehicle implicat
	Antiguitat_carnet	Antiguitat del carnet de conduir del conductor del vehicle implicat
	Multiples_vehicles_implicats	Indica si hi ha múltiples vehicles implicats en l'accident
	Numero_morts	Nombre de morts en l'accident
	Numero_lesionats_lleus	Nombre de lesionats lleus en l'accident
	Numero_lesionats_greus	Nombre de lesionats greus en l'accident

	Nomenclatura	Descripció
Dades mantingudes	*Numero_victimes*	Nombre de víctimes en l'accident
	Numero_vehicles_implicats	Nombre de vehicles implicats en l'accident
	Torn	Defineix el tram d'hores del dia en que s'ha produït l'accident
	Festiu	Indica si el torn es considera festiu o no
	Estat_mitja	Indica l'estat mitjà de la circulació a la via a l'hora en que es produeix l'accident
	03.TM	Temperatura mitjana [ºC]
	04.TX	Temperatura màxima [ºC]
	05.TN	Temperatura mínima [ºC]
	06.HRM	Humitat relativa mitjana [%]
	07.PPT24H	Precipitació acumulada [mm]
	08.HPA	Pressió atmosfèrica mitjana (valor no corregit a nivell del mar) [hPa]
	09.RS24H	Irradiació solar global [MJ/m^2]
	10.VVM10	Velocitat del vent a 10 m [m/s]
	11.DVM10	Direcció del vent a 10 m [º]
	12.VVX10	Ratxa màxima del vent 10 m [m/s]
	13.DVX10	Direcció de la ratxa màxima de vent a 10 m [º]

Taula 4-5. Dades seleccionades dels orígens de dades.

La taula anterior descriu aquelles variables que es mantenen després de la neteja preliminar de les dades i que s'utilitzen de cara al estudi dels accidents. D'igual manera, es descarta aquelles variables que clarament no són rellevants, ja sigui perquè son repetides o aporten informació que ja es captura en alguna altra o perquè clarament no són rellevants a causa de la seva naturalesa i la del present estudi. Són aquelles que fan referència, a través d'una variable contínua, a la situació geogràfica de l'accident al que acompanyen o la marca i model del vehicle implicat.

4.3. Anàlisi descriptiu

En la present secció es desenvolupa un anàlisi descriptiu de les dades mantingudes. Mitjançant aquest, s'entén millor els accidents i les seves característiques i es redueixen el grup de dades utilitzades. En primer lloc, es presenten aquelles variables obtingudes del pas anterior, diferenciant entre numèriques i categòriques. Sobre les dades numèriques es fa un anàlisi mitjançant mesures estadístiques bàsiques, com la mitjana o la mediana, agrupades segons la gravetat de l'accident. Les dades categòriques s'exploren en una taula de contingència d'on se'n deriva una primera interpretació de les característiques dels accidents i la seva influència en la gravetat d'aquests.

Les dades incloses a l'anàlisi descriptiu són les següents:

Variable	Tipus de variable
Districte	Categòrica
Barri	
Any	
Nom_mes	
Descripcio_causa_mediata	
Desc_Tipus_vehicle_implicat	
Descripcio_sexe	
Torn	
Festiu	
Descripcio_tipus_persona	
Descripcio_situacio	
Descripcio_victimitzacio	
Tipus_accident	
Descripcio_causa_vianant	
Descripcio_color	
Descripcio_carnet	
Multiples vehicles implicats	
Edat	Discreta
Antiguitat_carnet	
Numero_morts	
Numero_lesionats_lleus	
Numero_lesionats_greus	
Numero_victimes	
Numero_vehicles_implicats	
Estat_mitja	
03.TM	
04.TX	
05.TN	
06.HRM	
07.PPT24H	
08.HPA	
09.RS24H	
10.VVM10	
11.DVM10	
12.VVX10	
13.DVX10	

Taula 4-6. Variables incloses a l'anàlisi descriptiu.

Finalment, es realitza tests de correlació i dependència entre variables. Sobre les dades numèriques, es realitza el test de correlació de Pearson i Spearman. Sobre les variables

categòriques s'hi aplica el test de Cramer's V per identificar possibles dependències entre parelles de variables. A través d'aquests mètodes es justifica la decisió d'eliminar certes variables que mantenen una alta relació de semblança amb altres dades del joc de dades o que no presenten independència.

Anàlisi descriptiu. Dades numèriques

Al present apartat es descriuen aquelles variables numèriques que es seleccionen als apartats anteriors. Tot seguit es fa un anàlisi de cada una d'aquestes, plasmant la variable i la gravetat i nombre d'accidents, per intentar identificar de forma preliminar alguna relació entre aquestes. Es recorda que els valors que pren la gravetat de l'accident són:

1. Lleu, 2. Assistencia, 3. Hospitalitzat, 4. Greu, 5. Mort.

Es recolza les explicacions amb les taules [Annex I Taula I-1 fins Annex I Taula I-3], que recull l'anàlisi descriptiu realitzat sobre les dades numèriques. D'aquesta, se'n desprèn que en relació a les dades *edat* i *antiguitat_carnet* sembla que no tenen efecte sobre la gravetat de l'accident. La mitjana, que ens indica el valor esperable de la variable, es manté en un valor similar en tots els casos.

Si es representen aquestes dades gràficament [Figura 4-1] i [Figura 4-2], sembla que si que hi hagi un cert efecte, sobretot en l'edat. Existeix una oscil·lació de la mitjana que fa que els accidents greus es produeixin a mitjanes d'edat més altes, igual que els accidents més lleus. Respecte de l'antiguitat del carnet, sembla que presenti una oscil·lació, tot i que la variació de les dades és de tan sols un any i mig.

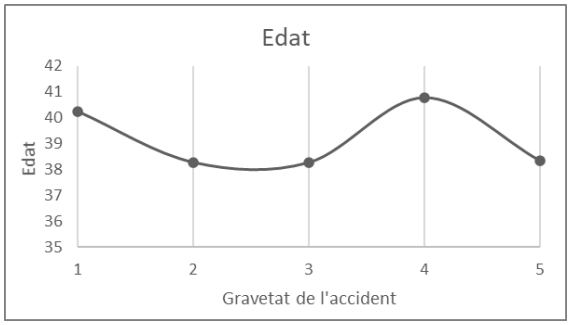


Figura 4-1. Edat – gravetat.

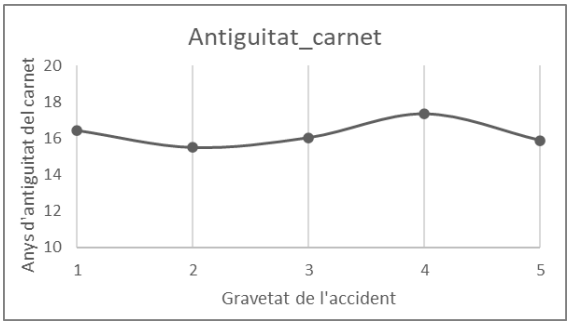


Figura 4-2. Antiguitat carnet – gravetat.

Pel que fa a la desviació estàndard, mesura de dispersió de les dades, ens mostra com de disperses es troben les dades respecte de la mitjana. A menor desviació, menor dispersió. De la taula se'n deriva que la dispersió de les diferents categories de gravetat és molt similar en totes dues variables. La mediana també és pràcticament la mateixa en les diferents categories. Finalment, es representa el nombre d'accidents totals segons l'edat i l'antiguitat del carnet, per veure la distribució dels accidents entre aquestes variables.

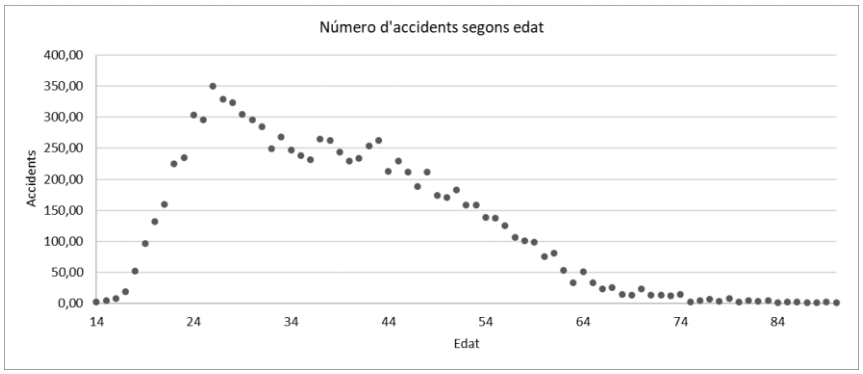


Figura 4-3. Edat – Volum d'accidents.

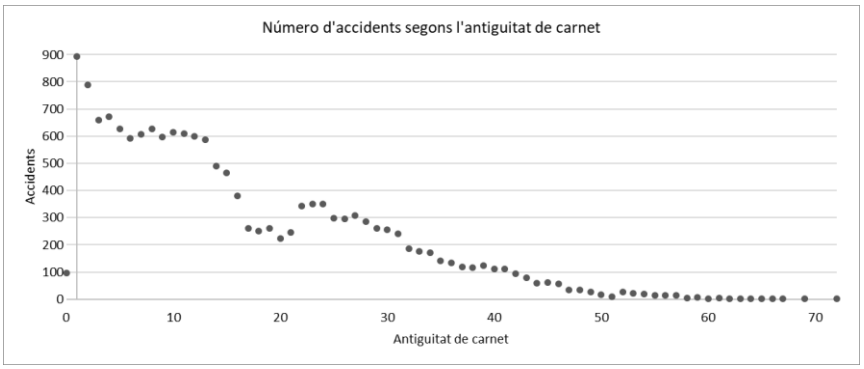


Figura 4-4. Antiguitat_carnet – Volum d'accidents.

De les anteriors figures se'n deriva que la distribució dels accidents al llarg de les variables d'edat i l'antiguitat del carnet no es manté sempre al mateix valor. En el cas de l'edat, hi ha un augment del volum d'accidents amb víctimes entre els 22 i els 45 anys, mentre que hi ha més accidents on el conductor fa entre 1 i 13 anys que va aprovar el carnet.

Sobre les dades de *numero_morts*, *numero_lesionats_lleus*, *numero_lesionats_greus* i *numero_victimes* cal comentar que mantenen una alta relació amb la gravetat de l'accident, donada la seva naturalesa. Són dades que es deriven de l'accident i per tant en son conseqüència. Tot i això, s'estudia les possibles relacions entre les víctimes i l'accident. Per fer-ho, es representa gràficament la seva relació [Figura 4-5] i [Figura 4-6].

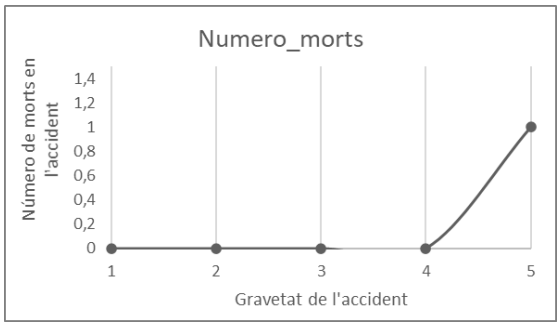


Figura 4-5. Nombre de morts – gravetat.

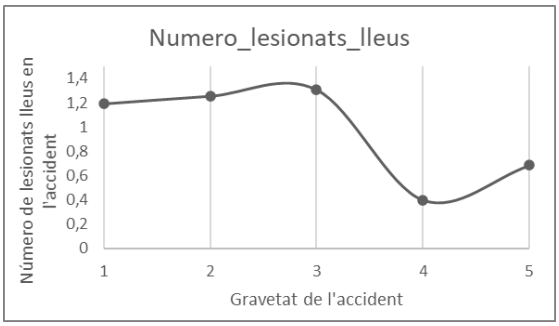


Figura 4-6. Nombre de lesionats lleus – gravetat.

Els accidents amb víctimes mortals s'identifiquen com accidents amb gravetat mort, de manera que clarament l'anàlisi ens mostra aquesta relació ja identificada. Els accidents amb nombre de morts guarda relació amb accidents de màxima gravetat.

Els accidents amb un *numero_lesionats_lleus* més elevat es relacionen amb una *gravetat* més baixa, mentre que a mesura que la *gravetat* de l'accident augmenta es redueixen les víctimes lleus.

Això pot ser explicat ja que, com que al vehicle hi circulen un nombre baix de persones (d'una a quatre o d'una a dues en vehicles de dues rodes, de forma general), el fet que una d'aquestes pateixi unes conseqüències severes fa que no les pateixi de lleus, de manera que quan l'accident suposa una víctima més greu, en general totes les víctimes (si hi ha altres passatgers) són de més gravetat. Hi pot haver víctimes lleus en totes les categories, de manera que permet veure una certa relació entre les variables.

Respecte la relació de les víctimes lleus amb el volum d'accidents, es pot identificar que a les dades recollides el més comú és que els accidents tinguin un o dos lesionats lleus. Concretament, s'identifica que hi ha un nombre molt elevat d'accidents on hi ha tan sols un lesionat lleu. Aquesta característica representa el 75% dels accidents dels que es disposa. Passa d'una manera similar amb les dades de *numero_lesionats_greus*. En aquestes dades, al 98% dels accidents no hi ha lesionats greus, mentre que al 2% dels accidents hi ha un lesionat greu.

A continuació, s'estudia la relació d'aquests accidents amb la gravetat dels accidents.

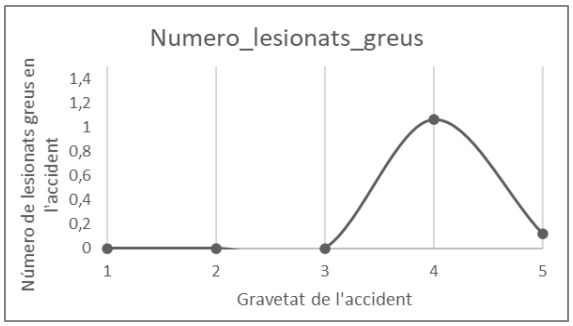


Figura 4-7. Nombre de lesionats greus – gravetat.

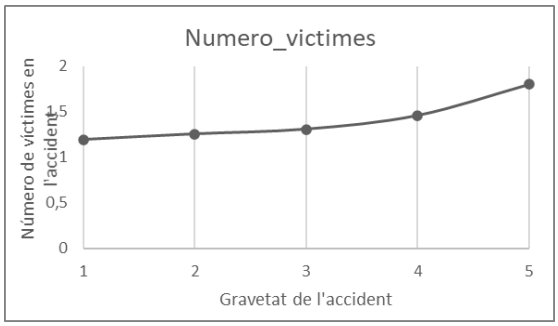


Figura 4-8. Nombre de víctimes – gravetat.

Clarament [Figura 4-7], si hi ha víctimes greus es tracta d'un accident de molta gravetat. Es pot notar que les víctimes amb lesions greus disminueix quan es tracta d'un accident mortal. Una hipòtesi és que als accidents mortals, si hi ha víctimes el més probable és que sigui mortal i no es quedi en una conseqüència greu. A través de les dades *numero_victimes* es pot veure que com més víctimes estan relacionades amb l'accident, més greu sol ser [Figura 4-8]. Aquesta variable passa d'una mitjana de 1,2 en accidents lleus a 1,8 en accidents

mortals. És explicable a través del fet que un accident lleu és menys probable que provoqui víctimes, degut a la seva baixa capacitat destructiva. En canvi, el sentit comú ens diu que un accident greu o mortal ha requerit d'una energia major per tal de provocar la mort o ferides d'alta gravetat, de manera que les possibilitats que hagi generat un impacte en terceres persones és més alt que no pas un accident lleu.

Estudiant *numero_vehicles_implicats*, sembla que els accidents mortals succeeixen amb un nombre més baix de vehicles implicats, tot i que el valor té una reducció poc rellevant [Figura 4-9].

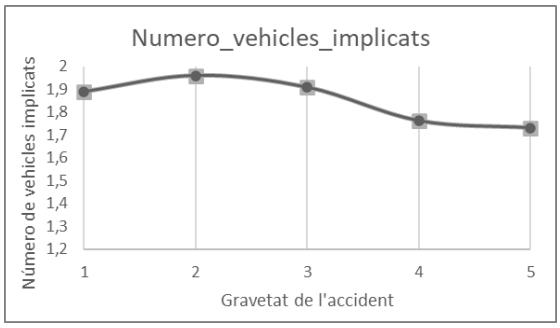


Figura 4-9. Nombre de vehicles implicats – gravetat.

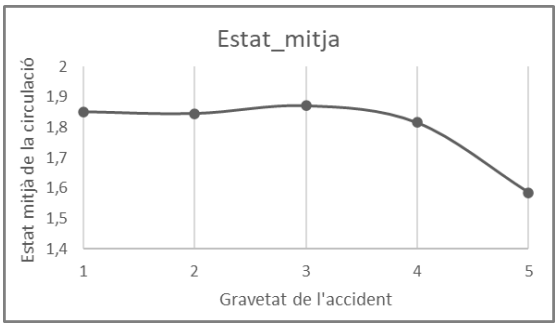


Figura 4-10. Estat mitjà de la circulació – gravetat.

També sembla que els accidents que tenen víctimes menys greus són aquells que impliquen menys vehicles. La dispersió d'aquestes dades es maximitza en els accidents de mitjana gravetat, aquells en que es rep assistència mèdica al lloc de l'accident, i els accidents de mínima i màxima gravetat són els que presenten la desviació més alta.

Estudiant les dades del nombre de vehicles implicats i el nombre d'accidents es pot identificar que hi ha un desequilibri en el nombre d'accidents segons el nombre de vehicles implicats. Hi ha una proporció molt elevada d'accidents que tenen dos vehicles implicats, concretament el 72% del total. De la resta, un 19% tenen un vehicle implicat i un 7% en tenen tres.

La variable que indica l'estat mitjà del trànsit reflecteix que un augment de la congestió fa que els accidents siguin de gravetat mitjana o baixa, com es pot veure a la [Figura 4-10]. Quan l'estat és menys fluid, sembla que els accidents requereixen assistència, ja sigui al lloc de l'accident o al hospital o són greus. Davant l'estat molt fluid, sembla que els accidents tendeixen a ser més greus.

Per acabar l'anàlisi de les variables numèriques, resta estudiar les dades meteorològiques. Les mitjanes de dels valors de les variables *03.TM* (temperatura mitjana), *04.TX* (temperatura màxima) i *05.TN* (temperatura mínima) pràcticament no presenten tendència quan es representen contra la gravetat dels accidents.

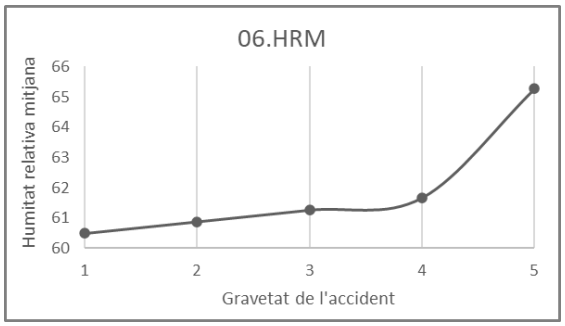


Figura 4-11. Humitat relativa mitjana – gravetat.

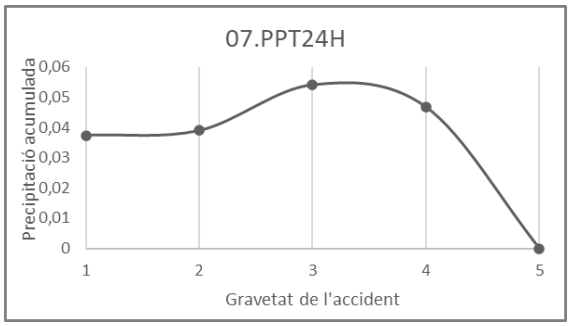


Figura 4-12. Precipitació acumulada – gravetat.

En canvi, la variable *06.HRM* (humitat relativa mitjana) és, de totes les meteorològiques, la que sembla que presenta una relació positiva amb la gravetat de l'accident. Sembla que els accidents són més greus quan la humitat relativa és més alta [Figura 4-11]. La relació amb la gravetat de la variable *07.PPT24H* (precipitació acumulada) [Figura 4-12] i amb la *08.HPA* (pressió atmosfèrica mitjana) sembla que suggereix un cert augment de la gravetat quan els valors d'aquestes augmenten, però els accidents mortals es caracteritzen per estar en un valor proper a zero en el cas de *07.PPT24H* i considerablement més baix en la variable pressió atmosfèrica mitjana.

Respecte de la resta de variables climàtiques, es poden agrupar en dos comportaments, ja que presenten unes tendències molt similars. Es presenta els gràfics més representatius.

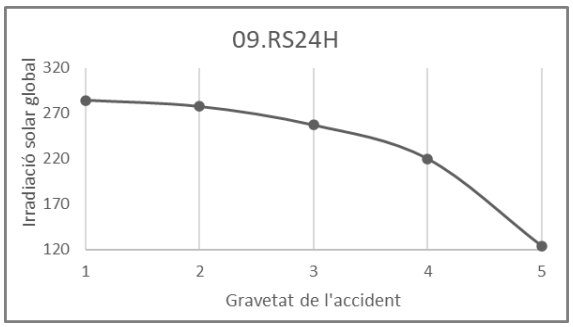


Figura 4-13. Humitat relativa mitjana – gravetat.

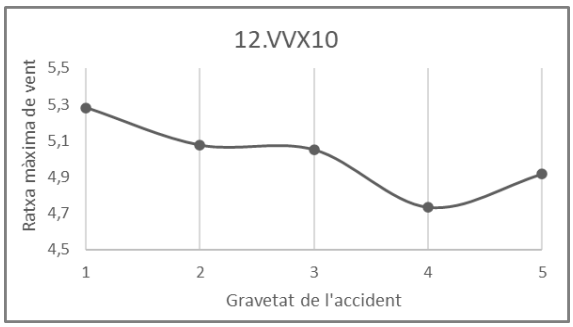


Figura 4-14. Precipitació acumulada – gravetat.

Les variables *11.DVM10* (direcció del vent) i *13.DVX10* (direcció de la ratxa màxima de vent) presenten una tendència paral·lela a la de la variable *09.RS24H* (irradiació solar global), en que la gravetat del accident augmenta amb la disminució del valor de la variable.

Per contra, la variable *10.VVM10* (velocitat del vent a 10 m) presenta una tendència similar a la de la variable *12.VVX10* (ratxa màxima del vent), on hi ha una reducció del valor de la variable segons augmenta la gravetat i canvia la tendència amb els accidents mortals.

Com a conclusió, sembla que les variables *edat* i *antiguitat_carnet* presenten un volum d'accidents molt variant segons el valor que pren la variable. Les dades de

numero_lesionats_lleus, *numero_victimes*, *numero_vehicles_implicats* i *estat_mitja* guarden una clara relació amb les conseqüències de l'accident. Les variables *numero_morts*, *numero_lesionats_greus* tenen una alta connexió, ja que comparteixen el factor que les defineix, la víctima més greu.

De les climatològiques, n'hi ha que tenen una tendència positiva, d'altres de negativa i finalment, les variables que sembla que no tenen efecte, on la gravetat de l'accident no sembla tenir una relació amb la temperatura mitjana, màxima i mínima. Tot i això, no es pot afirmar res amb total seguretat sobre la relació de les variables amb la gravetat de l'accident només revisant les taules [Annex I Taula I-1 a Annex I Taula I-3] i les diferents figures analitzades. A continuació, es desenvolupa l'anàlisi de les variables categòriques incloses a l'estudi preliminar [Taula 4-6].

Anàlisi descriptiu. Variables categòriques

Les taules disposades a [Annex I.i] mostren totes aquelles variables categòriques que conformen el grup de dades inicial i sobre el que s'està fent l'anàlisi descriptiu. Es comenta aquelles variables de les que se'n pot extreure una primera aproximació de com es reparteixen els accidents, segons la seva gravetat, a través de la variable i les diferents categories que aquesta pot prendre.

A primera vista es pot apreciar que hi ha un gran desequilibri en el nombre de dades que conformen els grups d'accidents segons la seva gravetat:

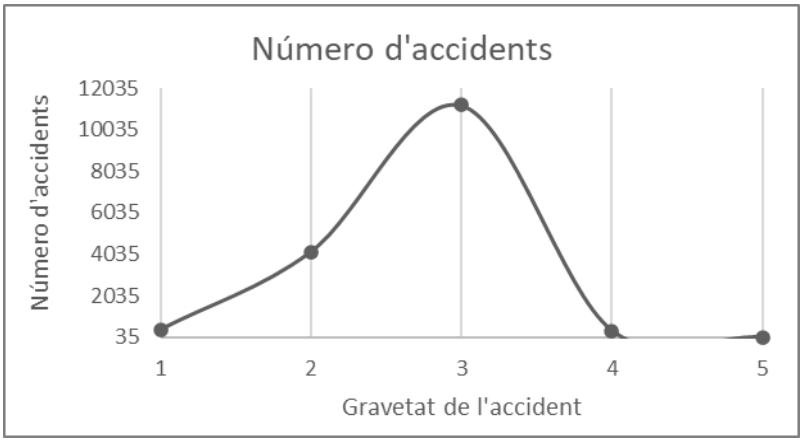


Figura 4-15. Nombre d'accidents – gravetat.

Gravetat	1	2	3	4	5	**Total**
Nombre d'accidents	399	4149	11210	358	41	16157
Proporció	2,5%	25,7%	69,4%	2,2%	0,3%	100%

Taula 4-7. Volum i proporció d'accidents segons la gravetat.

Les dades corresponents als accidents de gravetat *3. hospitalitzat* representen una mica més del 69% dels accidents. Juntament amb els accidents de tipus *2. assistencia* ja arriben a representar el 95% de les dades. Passa igual amb *descripcio_carnet*, degut a que la majoria de conductors tenen el carnet tipus *B* (64% del total) o amb *descripcio_causa_mediata*, on hi ha com a resposta *no hi ha causa mediata* al 95% dels accidents, *descripcio_causa_vianant*, on el 94% dels accidents es classifiquen com *no és causa del vianant*.

Sobre les dades de *districte*, es pot identificar un augment considerable dels accidents amb gravetat *3. hospitalitzat*.

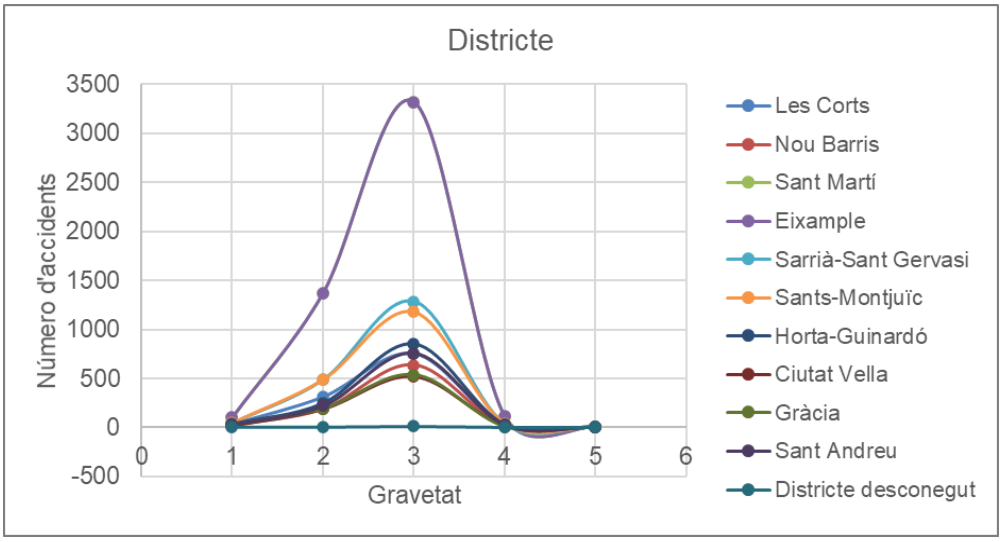


Figura 4-16. Nombre d'accidents – gravetat, segons districte.

Si es té en compte el nombre total d'accidents als districtes, a l'Eixample hi ha passat el 30% dels accidents. Amb això, sembla normal que hi hagi més accidents de tipus 3. Si es corregeix els valors, la proporció es manté, de manera que el valor d'accidents tipus 3 a l'Eixample no ens indica que hi hagi més accidents d'aquesta gravetat.

Gravetat	Gràcia	Sant Martí	Sants-Montjuïc	Sarrià-Sant Gervasi	Ciutat Vella	Eixample	Desc.	Les Corts	Sant Andreu	Nou Barris	Horta-Guinardó
Nombre d'accidents	1029	2496	2348	2426	985	6346	33	1478	1404	1194	1575
Proporció	4,8%	11,7%	11,0%	11,4%	4,6%	29,8%	0,2%	6,9%	6,6%	5,6%	7,4%

Taula 4-8. Nombre d'accidents i proporció, segons Districte.

Sobre la resta de districtes, tenen una mitja de 1124 accidents per districte al llarg dels dos anys estudiats i sembla que la distribució d'aquests al llarg dels districtes de Barcelona es manté igual.

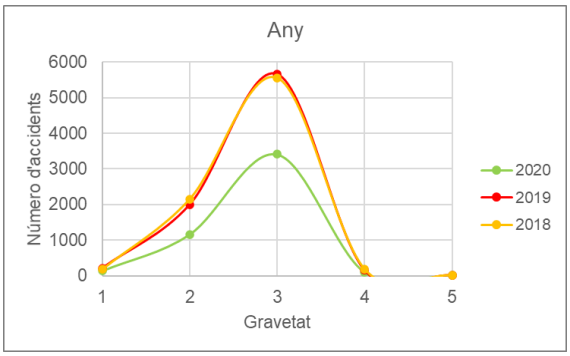


Figura 4-17. Nombre d'accidents en valor absolut – gravetat, segons any.

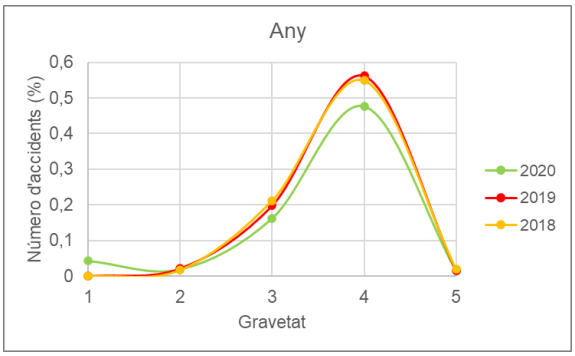


Figura 4-18. Nombre d'accidents, proporcionalment al total d'accidents – gravetat, segons any.

Pel que fa a l'any de l'accident, és interessant fer-ne l'anàlisi, ja que l'any 2020 és quan va aparèixer la pandèmia del Covid19, que va suposar una alteració de les vides d'aquells que viuen i es desplacen per la ciutat de Barcelona. És per l'anomalia que es va donar l'any 2020 que no s'inclou l'any a l'estudi.

Si s'estudien les dades de que es disposa a [Figura 4-17], es pot veure clarament la disminució dels accidents degut a les restriccions de circulació imposades als ciutadans. Sobre la figura [Figura 4-18] se'n pot treure que els accidents de gravetat tipus 3. Hospitalitzat són més baixes.

Gravetat	1	2	3	4	5	**Total**	**Proporció**
2020	143	1166	3413	108	14	4844	23,0%
2019	180	2149	5551	197	20	8097	38,6%
2018	219	2000	5659	161	21	8060	38,4%

Taula 4-9. Distribució dels accidents segons l'any.

A més a més, analitzant el detall de les dades [Taula 4-9] es pot veure que el nombre d'accidents dels anys 2018 i 2019 és gairebé el mateix, tenint el mateix nombre de dies hi ha una diferència de 37 accidents. Això significa que el 2019 hi va haver una reducció de, tan sols, el 0,46% en el nombre d'accidents respecte de 2018. Finalment, pel que fa a la proporció de la tipologia d'accidents, sembla que al 2020 es van reduir els accidents tipus *2. assistencia* i *3. hospitalitzat* respecte del 2018 i 2019, que es mantenen en la mateixa línia. L'any 2020, quan la congestió a la ciutat va ser més baixa, en molts casos inexistent, els accidents lleus i de gravetat mitja es redueixen.

Si a nivell anual no s'identifica una diferència entre els anys 2018 i 2019, a nivell mensual sembla que la proporció d'accidents, repartida a través dels mesos, també es manté igual. No destaca cap mes per la seva alta o baixa sinistralitat. A simple vista, sembla que segueix

una distribució força similar al llarg dels mesos, amb un volum d'accidents lleus més elevat, de tipus 2 i 3, que no pas categoritzats com greu o mort.

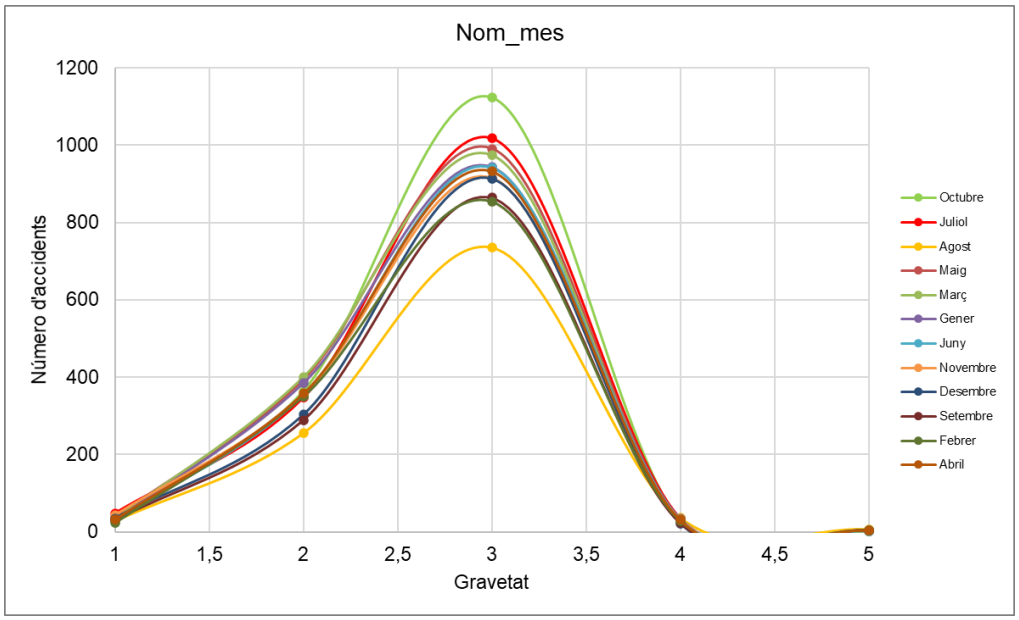


Figura 4-19. Nombre d'accidents – gravetat, segons nom_mes.

Si que es pot apreciar que, durant els mesos d'octubre, gener, juliol i febrer (especialment rellevant el febrer, que tot i tenir menys dies és dels mesos amb més sinistres) el volum d'accidents és més alt, després apareixen els mesos de desembre, novembre, setembre juny i març. Per últim, els mesos d'abril, maig i agost són els mesos de menys accidents.

En funció del sexe, se'n pot extreure una visió d'aquell grup que té més accidents, o que són més greus. Com que no es pot saber el total de conductors que han circulat, es fa difícil de dir quin dels dos sexes condueix millor. O, almenys, pateix menys accidents.

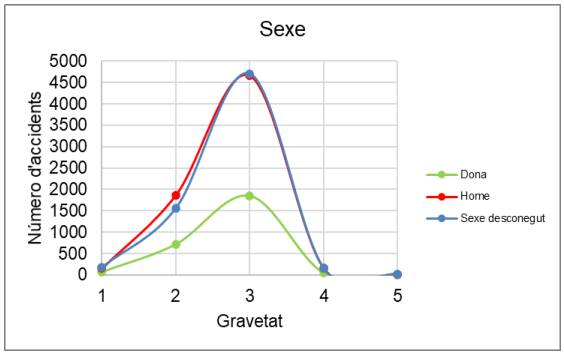


Figura 4-20. Nombre d'accidents en valor absolut – gravetat, segons sexe.

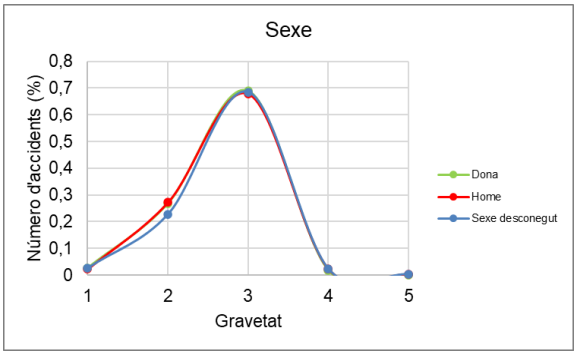


Figura 4-21. Nombre d'accidents, proporcionalment al total d'accidents – gravetat, segons sexe.

En aquest cas, sembla que hi ha un volum molt menor d'accidents entre el sexe femení

[Figura 4-20], això fent l'aventurada suposició que hi ha el mateix volum de conductors entre tots dos sexes. A simple vista, sembla que hi ha un canvi en el volum d'accidents, però no hi ha una clara distinció en quant a la gravetat d'aquests [Figura 4-21]. Si que es pot apreciar un volum més alt en els accidents de gravetat *2. assistencia* per part d'homes i un volum, proporcionalment parlant respecte del total d'accidents per sexes, per sota del de les dones quan a gravetat *3. hospitalitzat*.

En relació al *color* dels vehicles, la distribució dels accidents sembla que no presenta cap anomalia, la distribució de la gravetat d'aquests entre tots els colors registrats sembla que es manté. Quan es revisa la variable de *multiples vehicles implicats*, es pot observar que hi ha un volum major d'accidents on hi ha múltiples vehicles implicats, concretament gairebé quatre vegades més d'accidents múltiples.

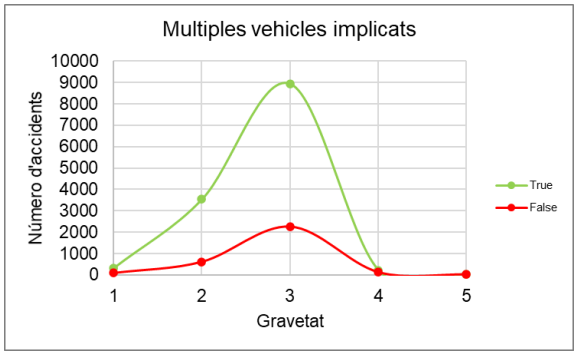


Figura 4-22. Nombre d'accidents en valor absolut – gravetat, segons multiples vehicles implicats.

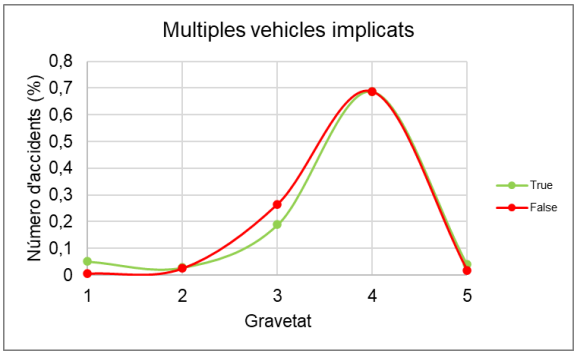


Figura 4-23. Nombre d'accidents, proporcionalment al total d'accidents – gravetat, segons multiples vehicles implicats.

A més a més, proporcionalment sembla que els accidents múltiples pateixen més accidents de categoria *2. assistencia* i menys de categoria 4*. greu*, mentre que la proporció es manté igual en la resta de categories d'accident. Estudiant el *torn*, es pot veure que en alguns torns hi ha més accidents que en d'altres [Figura 4-24], com és a la tarda, vespre i migdia per contraposició del matí, nit i mitjanit, quan en volum total hi ha menys accidents.

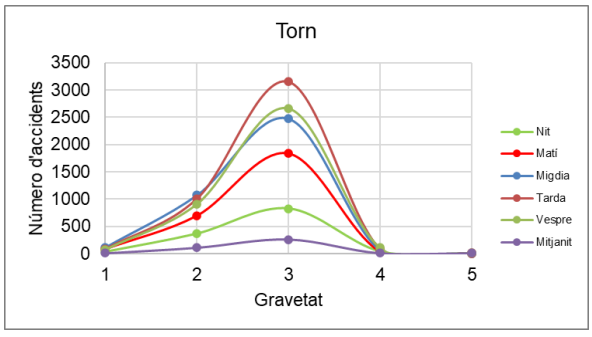


Figura 4-24. Nombre d'accidents en valor absolut – gravetat, segons torn.

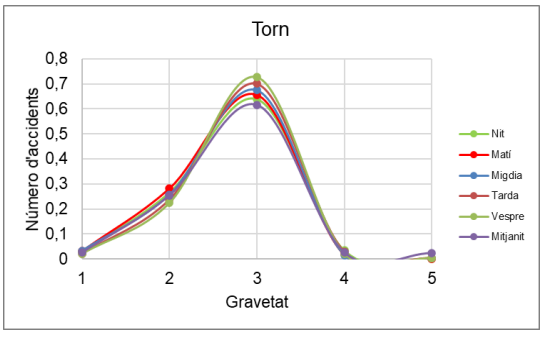


Figura 4-25. Nombre d'accidents, proporcionalment al total d'accidents – gravetat, segons torn.

En canvi, revisant la proporció d'aquests [Figura 4-25] sembla que durant la *tarda* i el *vespre* els accidents es caracteritzen per ser de gravetat *3. hospitalitzat* i durant la *mitjanit* i la *nit* els accidents semblen més radicals.

O no s'informa cap víctima, on clarament es desmarca de les altres franges, o aquesta víctima (o víctimes) és mortal, on també destaquen la *nit* i la *mitjanit*. Finalment, analitzant les franges del dia caracteritzades com *festiu* i no *festiu*, es pot veure el següent:

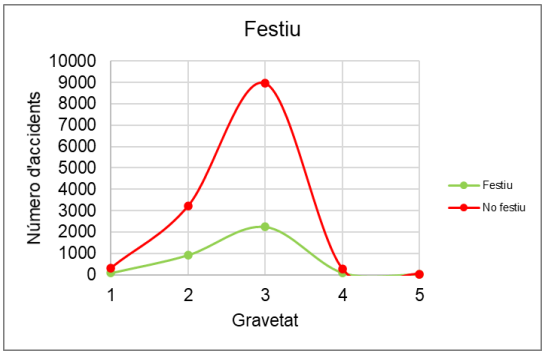


Figura 4-26. Nombre d'accidents en valor absolut – gravetat, segons festiu.

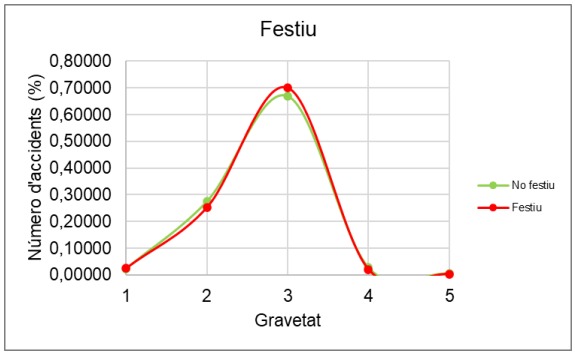


Figura 4-27. Nombre d'accidents, proporcionalment al total d'accidents – gravetat, segons festiu.

Clarament, es dies *no festius* tenen més accidents. Això, previsiblement, és degut al major volum de dies *festius* contra *no festius*. En canvi, si s'avalua la proporció dels tipus d'accidents, sembla que els dies *festius* els accidents solen ser en certa mesura, menys greus que no pas els dies *no festius*, quan augmenten clarament els accidents tipus *3. hospitalitzat*. També es pot apreciar que els accidents on hi ha víctimes *greus* i *mortals* (categories *4* i *5*) són més freqüents en dies festius.

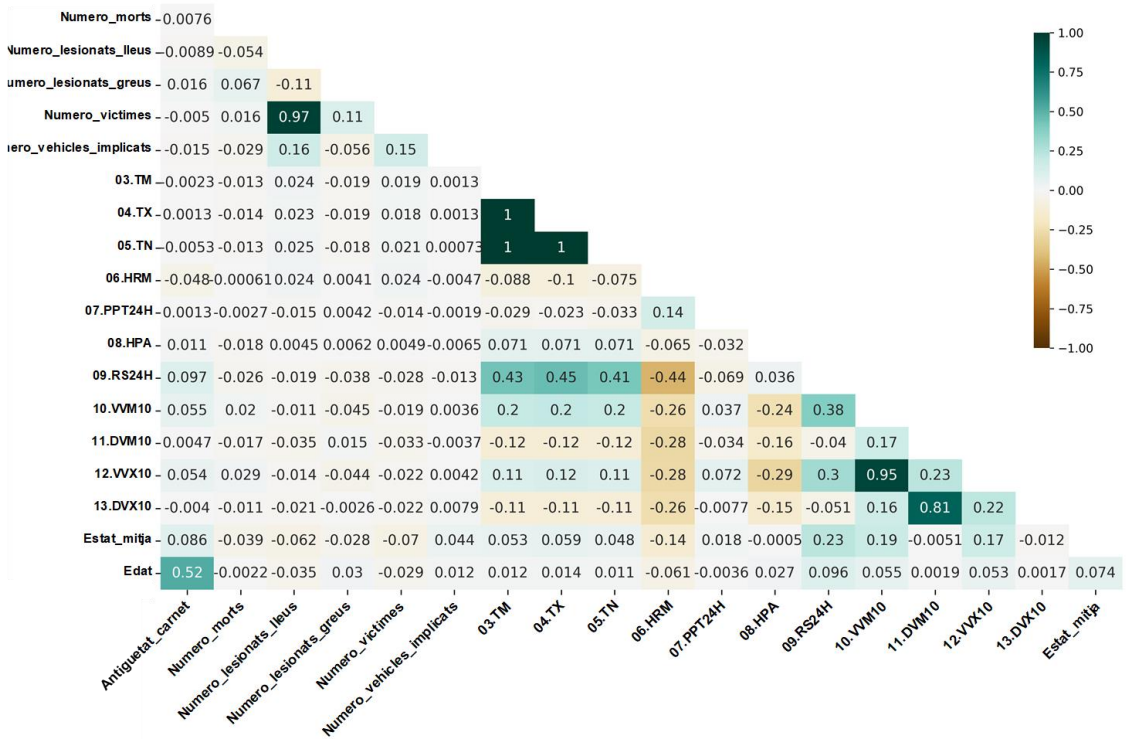
4.4. Correlació i dependència entre variables explicatives

Com es comenta a la introducció [4.3 Anàlisi descriptiu], aquest apartat de la memòria pren com a objectiu identificar i retirar del model aquelles variables que aporten informació altament relacionada, de manera que només es mantingui una d'aquestes.

Les correlacions i dependències entre variables s'identifiquen a través dels mètodes d'anàlisis de variables aplicats sobre les dades numèriques i categòriques, per separat. En primer lloc, es desenvolupa l'anàlisi de les variables numèriques. Sobre aquestes s'empren dos mètodes: els anàlisis de correlació de variables de Pearson i de Spearman.

A continuació, es mostren les taules de correlació obtingudes amb els dos mètodes exposats. En primer lloc, la taula de relació entre variables, amb els coeficients de correlació Pearson. Juntament amb la informació que aporten els diagrames de correlació es calcula els P-Values de cada parella de dades, disposats a [Taula 4-11] i [Taula 4-13]. Aquest valor

mostra la probabilitat d'arribar al mateix resultat si la relació entre variables fos realment inexistent.



Taula 4-10. Taula de correlació entre variables. Coeficients Pearson.

A simple vista, es pot identificar que hi ha una alta relació entre les següents parelles de variables:

Edat – antiguitat_carnet

numero_victimes – numero_lesionats_lleus

*04. TX (*temperatura màxima*) – 03. TM (*temperatura mitjana*)*

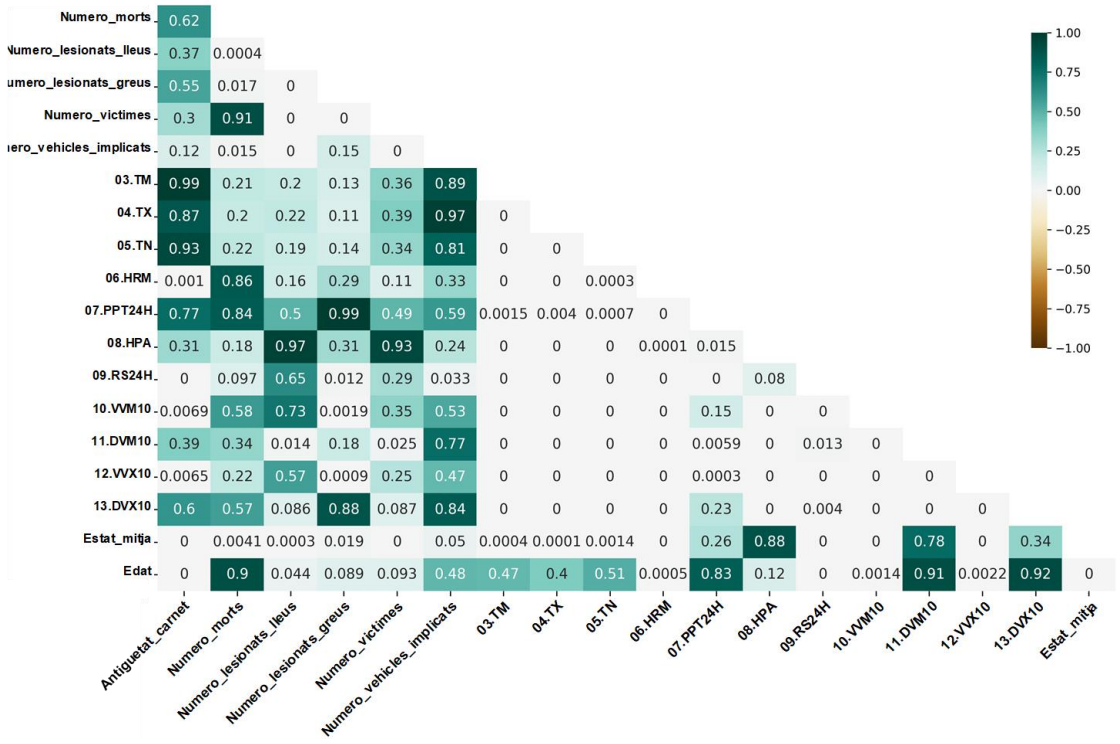
*05. TN (*temperatura mínima*) – 03. TM (*temperatura mitjana*)*

*05. TN (*temperatura mínima*) –04. TX (*temperatura màxima*)*

*12. VVX10 (*ratxa màxima de vent*) – 10. VVM10 (*Velocitat del vent a 10 m*)*

13. DVX10 (Direcció de la ratxa màxima de vent) – *11. DVM10* (Direcció del vent)

Totes elles amb coeficients de correlació de Pearson superiors a 0.81 i uns P-Value<0.0001, valors que indiquen la relació entre parelles de variables i la seva rellevància estadística.

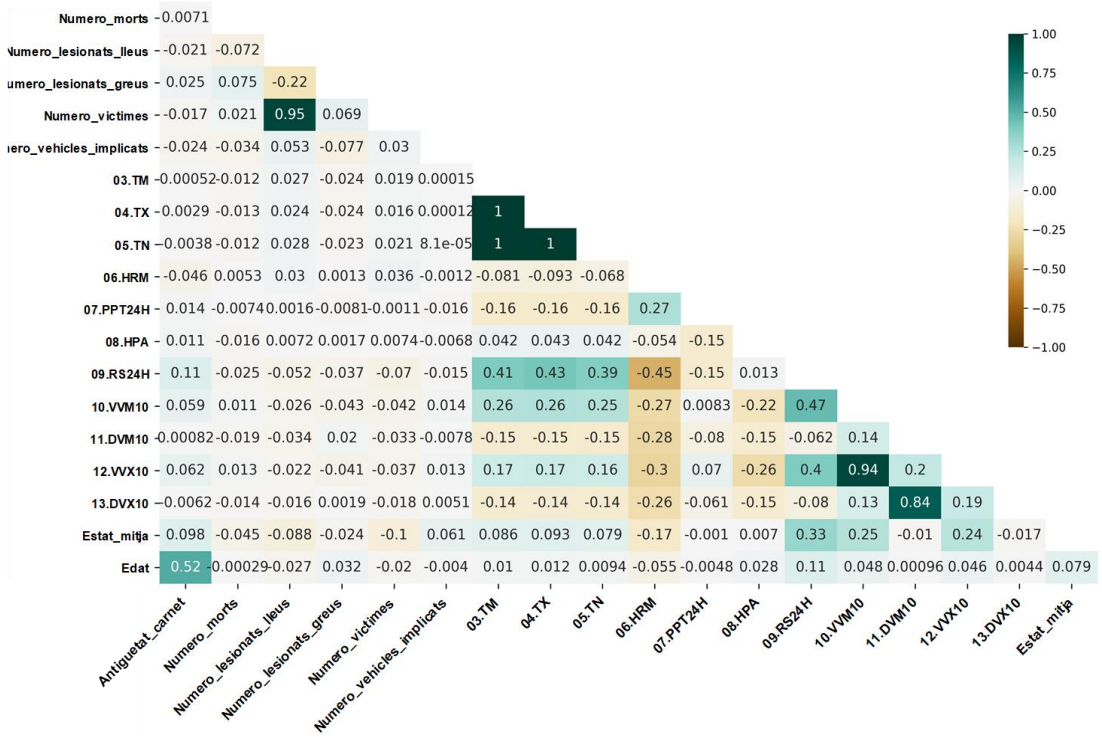


Taula 4-11. Taula de P-Values de la correlació Pearson.

Sobre la resta de relacions entre variables, recolzant-se en [Taula 4-11], es pot arribar a la conclusió que hi ha una relació mitja entre la variable *irradiació solar global* amb les diferents variables de *temperatura* i *humitat relativa,* amb la que presenta una relació inversa. També entre la *velocitat del vent mitjana* i la *irradiació solar global* i entre la *ratxa de vent màxima* i la *irradiació solar global .* Totes elles presenten P-values molt propers a zero, que en fan que les conclusions siguin estadísticament rellevants.

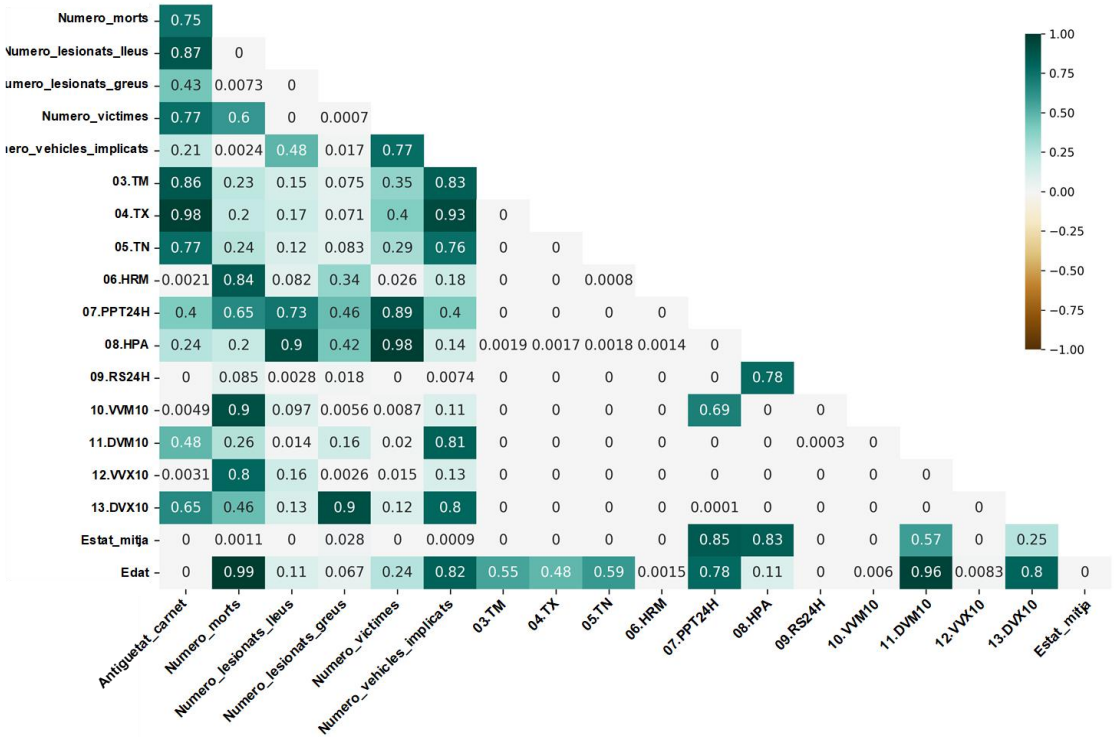
De la resta de parelles de variables que són estadísticament rellevants es pot afirmar que són independents entre elles. Finalment hi ha les parelles on el P-Value corresponent indica que no es pot extreure una conclusió amb robustesa de la taula de coeficients de Pearson.

En segon lloc, la taula de correlació amb els coeficients de correlació Spearman i la taula de P-Values corresponent:



Taula 4-12. Taula de correlació entre variables. Coeficients Spearman.

De la taula de coeficients de Spearman [Taula 4-12] se'n conclou que hi ha una alta relació entre les mateixes parelles de variables identificades mitjançant els coeficients de Pearson.



Taula 4-13. Taula de P-Values de la correlació Spearman.

Aquestes parelles de variables presenten uns coeficients de correlació de Spearman superiors al 0.84 en tots els casos i uns P-Value < 0.0001, que determinen que les relacions extretes són estadísticament rellevants. La resta de relacions de variables presenten un paral·lelisme absolut a les conclusions obtingudes amb el test de Pearson, tant a nivell de correlació com de significació estadística. De les parelles que s'identifica que aporten informació molt relacionada es mantenen les variables següents:

Edat
numero_lesionats_lleus
03. TM (temperatura mitjana)
10. VVM10 (velocitat del vent a 10 m)
11.DVM10 (direcció del vent a 10 m)

Sobre les variables amb correlació mitja o baixa no se'n descarta cap. La resta de parelles no presenten una correlació significativa. Addicionalment, es descarta la variable *11.DVM10* (direcció del vent a 10 m) ja que no sembla lògic que els resultats tinguin cap influència sobre els accidents a la ciutat de Barcelona.

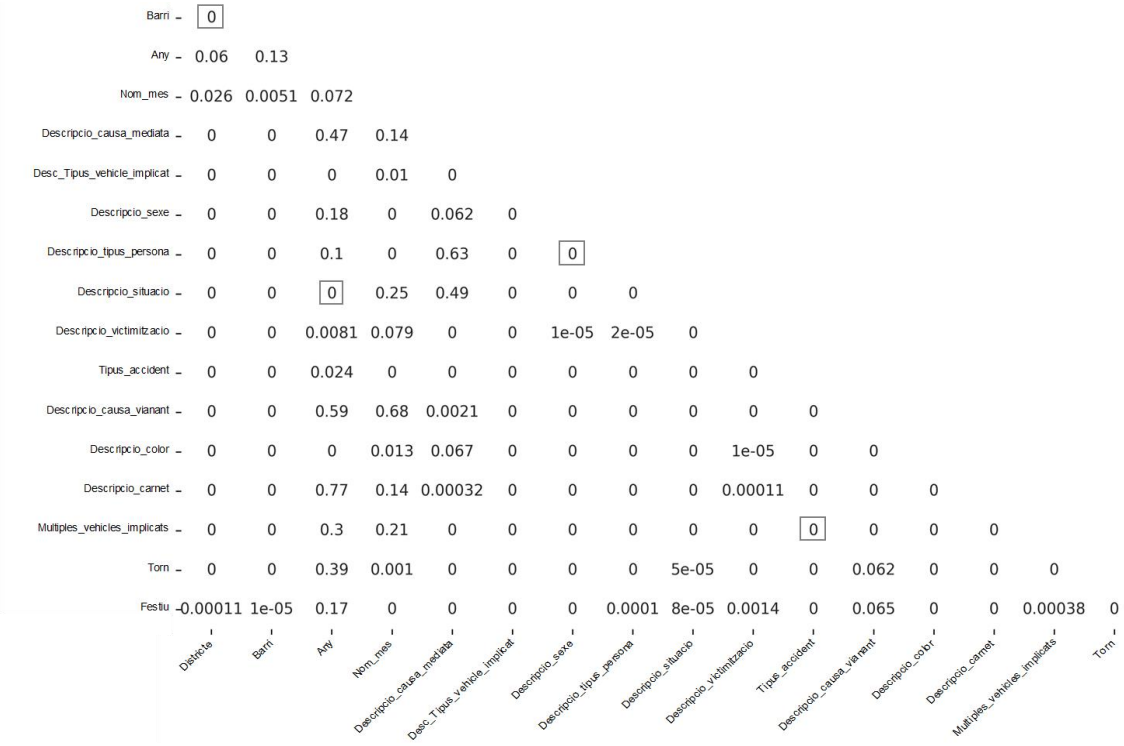
A continuació, es desenvolupa l'anàlisi de dependència de les variables categòriques. En aquest cas, s'utilitza el test de Cramer's V, amb la complementació dels P-Value corresponents.

Dels resultats obtinguts a la taula de coeficients [Taula 4-14], juntament amb els P-values que mostren si les conclusions a les que s'arriba són estadísticament rellevants, s'identifica que hi ha una relació de dependència entre les següents parelles de variables:

barri – districte
descripcio_situacio – any
descripcio_sexe – descripcio_tipus_persona
multiples_vehicles_implicats – tipus_accident

Totes les parelles tenen un valor de P-value inferior al 0.05 marcat com a llindar, de manera que es pot afirmar que les parelles anteriors mantenen una certa dependència.

Taula 4-14. Taula del test Cramer's V.



Taula 4-15. Taula de P-Values del test de Cramer's V.

A la següent taula es disposen els valors del test de Cramer's V, fent els resultats més visuals.

Parella de variables	Coeficient de dependència	Grau de dependència
barri – districte	0.87	Alt
descripcio_situacio – any	0.67	
descripcio_sexe – descripcio_tipus_persona	0.87	
multiples_vehicles_implicats – tipus_accident	0.74	

Taula 4-16. Resultats de l'anàlisi de dependència de variables categòriques.

Els resultats mostren el grau de dependència entre les parelles de variables indicades [Taula 4-16]. D'aquesta manera, les variables descartades i mantingudes respecte del joc de dades sobre el que es fa l'anàlisi descriptiu són:

Variables mantingudes	Variables descartades
Districte	*Numero_victimes*
Nom_mes	*04.TX*
Descripcio_causa_mediata	*05.TN*
Desc_Tipus_vehicle_implicat	*11.DVM10*
Descripcio_sexe	*12.VVX10*
Descripcio_situacio	*13.DVX10*
Descripcio_victimitzacio	*Antiguitat_carnet*
Tipus_accident	*Numero_victimes*
Descripcio_causa_vianant	*Barri*
Descripcio_color	*Any*
Descripcio_carnet	*Descripcio_tipus_persona*
Numero_morts	*Multiples vehicles implicats*
Numero_lesionats_lleus	*Descripcio_model*
Numero_lesionats_greus	*Descripcio_marca*
Numero_vehicles_implicats	
Torn	
Festiu	
03.TM	
06.HRM	
07.PPT24H	

Variables mantingudes	Variables descartades
08.HPA	
09.RS24H	
10.VVM10	
Estat_mitja	
Edat	

Taula 4-17. Variables mantingudes i descartades al joc de dades, post anàlisi de dependència.

A continuació, s'opta per fer aplicar el mètode de reducció de variables Principal Components (PCA) amb la finalitat de reduir les característiques que tenen els accidents i disposar d'un conjunt de dades més reduït amb el que poder treballar.

4.5. Anàlisi previ al clustering

Al present apartat de la memòria es realitza una primera aproximació a l'agrupació mitjançant tècniques de clustering. S'empren diferents mètodes per tal d'estudiar les dades mantingudes, la seva agrupació en Principal Components i com caracteritzen els accidents, definint el nombre de clústers que minimitza els errors de classificació del conjunt de les dades. De representar amb els Principal Components de manera significativa el joc de dades, suposaria una reducció dimensional important. Respecte del nombre òptim de clústers, és un primer anàlisi que permet entendre la proximitat entre les dades i un punt de partida per a l'explotació dels algorismes utilitzats posteriorment.

Factor Analysis of Mixed Data (FAMD)

El primer dels anàlisis es tracta d'un Factor Analysis of Mixed Data. Analitzant els resultats, sembla que hi ha una separació força clara seguint la gravetat entre accidents, sobretot entre els de més alta (*mort*) i les restants. Ara, no fa una bona separació entre la resta d'accidents. Cal tenir en compte el percentatge de la variabilitat que s'explica emprant aquests dos Principal Components. En aquest cas, combinant-los tots dos, s'explica gairebé un 11% de la variabilitat del joc de dades, valor insuficient per seguir treballant amb aquestes dades a través dels components principals aconseguits. Estudiant el nombre de Principal Components necessaris per explicar una elevada part de la variabilitat del joc de dades, es conclou que aquest és massa elevat per considerar-lo. Es requereix de treballar amb 20 Principal Components per aconseguir explicar el 85.5% de la variabilitat.

Cal tenir en compte que hi ha variables categòriques que fan que el total de variables, quan aquestes es converteixen a numèriques, sigui de 96 variables. Es segueix treballant amb el conjunt de dades indicat a la [Taula 4-17], degut a que no s'aconseguirà la representació en

dues dimensions, una de les principals avantatges de la reducció dimensional mitjançant PCA.

Test de Silhouette

Aquest és un mètode de validació de l'agrupació de dades en clústers, ja introduït al capítol [3]. Aplicant el mètode sobre el joc de dades, es distingeix que les millors agrupacions es troben entre els 5 i 6 clústers, amb una nota de Silhouette de 0.38 i 0.36 respectivament. A continuació es disposen els resultats de les agrupacions de cada un dels accidents, segons 5 i 6 clústers.

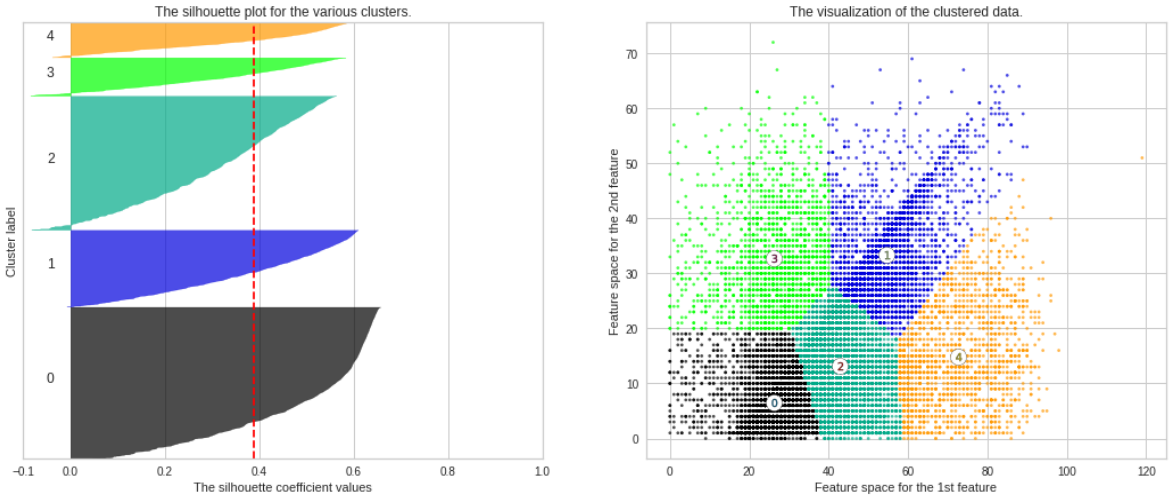


Figura 4-28. Anàlisi de Silhouette per agrupació mitjançant K-Prototypes, 5 clústers.

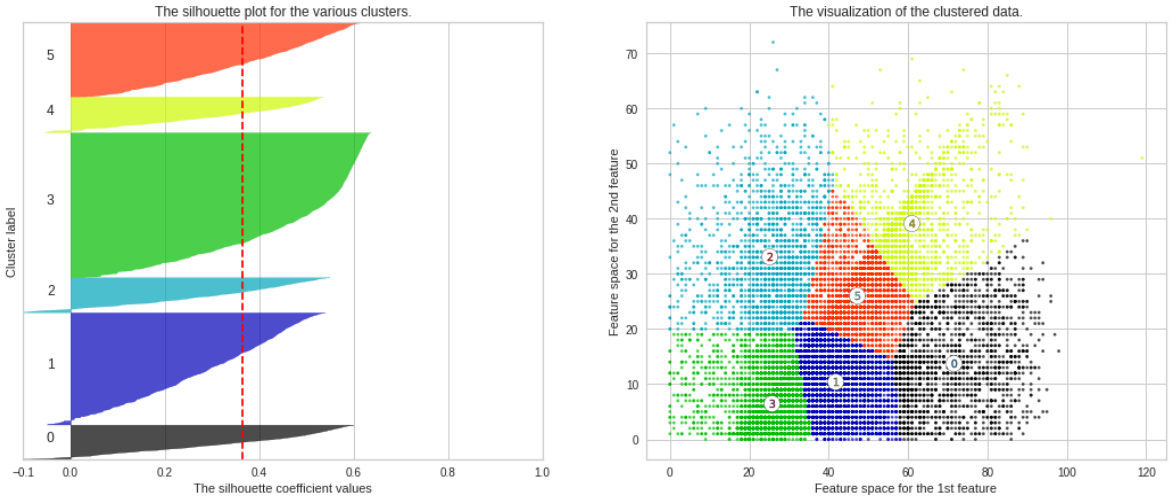


Figura 4-29. Anàlisi de Silhouette per agrupació mitjançant K-Prototypes, 6 clústers.

Com és apreciable a ambdues figures, apareixen valors negatius a les puntuacions de Silhouette, el que indica que les agrupacions tenen accidents mal classificats.

Test del colze

Per tal de confirmar els resultats, es realitza un Elbow Test sobre el joc de dades, per trobar el punt el en que el sumatori quadrat de les distàncies entre els punts i el seu centroide ja no millora significativament quan augmenten el nombre d'agrupacions creades.

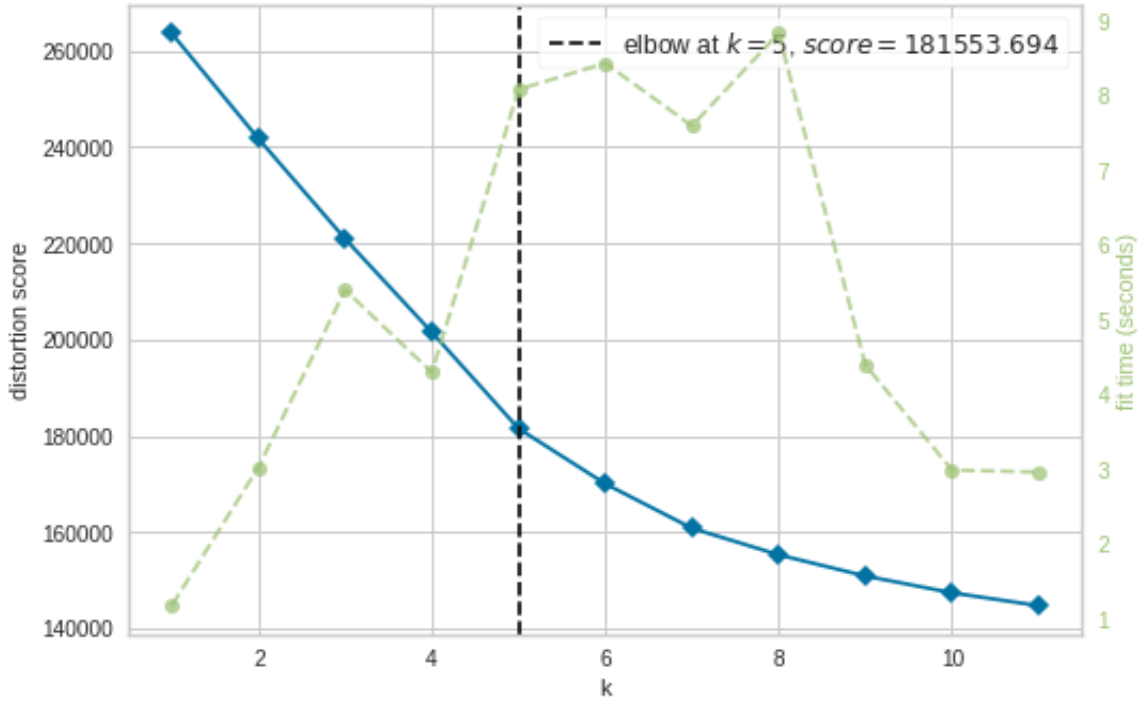


Figura 4-30. Elbow test sobre les dades.

El test confirma que el nombre de clústers mitjançant als que s'arriba a les agrupacions amb menys errors de classificació de les dades es troben entre els 5 i els 6 clústers, on la puntuació (distorsion score) ja no millora significativament quan s'augmenta el nombre de clústers.

Tot i això, considerant que el present treball no es busca les millors de les agrupacions, els valors de 5 i 6 clústers serveixen de referència i punt de partida per les primeres aproximacions als clústers, però no es prenen com a objectiu o condició de disseny.

4.6. Transformacions sobre el joc de dades

Per tal d'adequar les dades als algorismes seleccionats es realitzen una sèrie de transformacions sobre les variables, com es comenta a la definició de la metodologia de treball [3.1]. A continuació, es detalla el producte del procediment, per a cada algorisme.

K-means

Per a l'aplicació de l'algorisme K-means sobre el joc de dades, cal realitzar la transformació de les dades categòriques en variables binàries, juntament amb una estandardització del conjunt de variables restants. Això fa que apareguin un total de 2452 variables binàries, necessàries per representar el total de categories preses per part de les variables.

K-modes

Tot seguit es presenta les transformacions de dades requerides per poder aplicar les proves amb l'algorisme K-modes, distingint entre variables numèriques i categòriques.

Variables numèriques

Fent l'agrupació de variables numèriques a categòriques es perd informació aportada per aquestes, però és una conseqüència assumida en el disseny de l'experiment. D'aquesta manera, es fa les agrupacions de les variables recollides a [Annex II.i.1]. És important que presentin un volum d'accidents representatiu, sempre que sigui possible. Es busca que la categoria creada presenti el 15% o més del total d'accidents estudiats.

Les variables queden compactades de la següent manera:

Edat	**Joves**: 14-23 anys **Adolescents**: 24-31 anys **Adults**: 32-43 anys **Grans**: 44-53 anys **Ancians**: +54 anys
Numero_lesionats_lleus	**Cap**: 0 lesionats **Un**: 1 lesionat **Dos o més**: +2 lesionats
Numero_lesionats_greus	**Cap**: 0 lesionats **Un o més**: +1 lesionat
Numero_vehicles_implicats	**Un**: 1 vehicle implicat **Dos**: 2 vehicles implicats **Tres o més**: +3 vehicles implicats

03.TM	**GT1**: de 0ºC a 12.99ºC **GT2**: de 13ºC a 15.59ºC **GT3**: de 15.60ºC a 25.49ºC **GT4**: més de 25.50ºC
06.HRM	**GH1:** de 0% a 45% **GH2:** de 46% a 54% **GH3:** de 55% a 72% **GH4:** més de 73%
07.PPT24H	**GP1:** de 0mm a 0.29mm **GP2:** més de 0.29mm
08.HPA	**GPA1:** de 0 hPa a 1007 hPa **GPA2:** de 1008 hPa a 1016 hPa **GPA3:** més de 1017 hPa
10.VVM10	**GV1:** de 0 m/s a 1.1 m/s **GV2:** de 1.2 m/s a 2.9 m/s **GV3:** més de 3 m/s

Taula 4-18. Agrupació de les variables numèriques.

Variables categòriques

Algunes variables categòriques prenen una gran quantitat de valors que fan que la variable original agafi molta importància al model pel simple fet de tenir moltes categories. Per adreçar aquest problema es realitza una compactació d'aquestes variables [Annex II.i.2]. Concretament, a través de l'anàlisi desenvolupat a [4.3] s'identifica les variables que s'agrupa els seus valors tal com s'indica a les figures compreses entre [Annex II Figura II-10] i [Annex II Figura II-14].

Les variables queden agrupades tal com es disposa a continuació:

Descripcio_situacio	Presentat Situació desconegut Altres situacions
Tipus_accident	Tipus accident 1 Tipus accident 2 Tipus accident 3 Tipus accident 4
Descripcio_color	Tipus de color 1 Tipus de color 2 Tipus de color 3
Descripcio_carnet	Tipus de carnet 1 Tipus de carnet 2 Tipus de carnet 3

Desc_Tipus_vehicle_implicat	Motocicleta Turisme Ciclomotor Altres vehicles

Taula 4-19. Agrupació de les variables categòriques.

A les figures compreses entre [Annex II Figura II-10] i [Annex II Figura II-14] s'hi pot trobar més detall sobre les categories que integren cadascuna de les anteriors.

K-prototypes

La transformació requerida per aplicar l'algorisme K-prototypes és una estandardització del conjunt de variables numèriques. Realitzada aquesta, ja que l'algorisme realitza agrupacions amb totes dues tipologies de dades, es pot aplicar el procés de prova dissenyat sobre el joc d'accidents.

5. Obtenció dels clústers

Aquest correspon a l'execució de les proves aplicant els algorismes seleccionats, iterant sobre les agrupacions obtingudes i realitzant les modificacions de variables corresponents segons la seva rellevància entre els resultats. També es presenten aquests juntament amb una descripció de les situacions identificades.

5.1. Execució de proves

Al punt actual, es disposa de tres jocs de dades sense correlacions ni dependències que presenten informació sobre 16157 accidents de trànsit a la ciutat de Barcelona. Aquestes variables han estat transformades i adaptades segons el que es detalla als anteriors aparats. Tal com es defineix al capítol metodològic, el mètode de treball segueix per fer les execucions de proves aplicant els algorismes de clustering seleccionats.

Joc de dades complet

Partint del joc de dades amb els 16157 accidents i les variables que els caracteritzen [Taula 4-14], sobre les que es fa les diferents transformacions descrites, s'aplica tres vegades el mètode de proves definit [Figura 3-1], canviant en cadascuna l'algorisme de clustering.

La primera de les execucions es fa amb l'algorisme K-means i el joc de dades amb totes les variables transformades a numèriques i normalitzades. Analitzant les diferents fases de la prova i els resultats, s'identifica que aquests no són bons.

En primer lloc, no és directe fer l'anàlisi d'aquests resultats, ja que no és trivial interpretar que indica i la significació de la distància que separa dues variables categòriques.

En segon lloc, la transformació de variables categòriques a numèriques crea una gran quantitat de variables auxiliars, utilitzades per representar totes les possibles categories que es prenen i que dilueixen les variables numèriques. L'algorisme interpreta molt més clarament les variables binàries, de manera que té tendència a agrupar segons aquestes. Això fa que els resultats presentin uns clústers agrupats segons les variables categòriques i que no pren en consideració les numèriques.

És per aquestes raons que es pren la decisió de no seguir amb les proves mitjançant l'algorisme K-means. La segona execució de prova es realitza amb l'algorisme de clustering K-prototypes, que treballa amb tots dos tipus de variables.

En aquestes proves, es detecta que les variables numèriques prenen protagonisme a l'hora d'agrupar els accidents en clústers. Això s'explica ja que l'algorisme està prenent les variables numèriques com a categòriques, transformant-les en un gran nombre de variables

binàries. Això fa que, igual que amb l'anterior prova, unes variables prenguin més importància al model per tenir una gran quantitat de variables binàries que les representen. Aquestes proves es rebutgen per no ser representatives dels resultats, ja que el mètode de clustering utilitzat no és adient per a les dades de que es disposa.

Finalment, la tercera de les proves s'executa a partir de l'algorisme K-modes, amb les transformacions de dades indicades a 4.6].

Veient que les proves presenten unes agrupacions on es considera tots dos tipus de variables i a partir dels resultats obtinguts com a agrupació definitiva, s'identifica una sèrie de modificacions que poden fer que el resultat millori considerablement, des del punt de vista d'aconseguir situacions que siguin representatives.

Hi ha uns accidents on no es té identificat el conductor i que per tant presenten les variables *Edat* i *Descripcio_sexe* com a "*desconegut*". Això fa que en molts dels clústers s'agrupi els accidents en funció de si es coneix o no les dades del conductor i no de les dades pròpiament. Buscant aconseguir uns resultats més rellevants i millorar la operativa, es descarta els accidents que presenten la categoria "*desconegut*" a les variables *Descripcio_sexe* i *Edat*, característiques que identifiquen al conductor d'un dels vehicles implicats en l'accident. Aquests són un total de 6613 accidents.

D'aquesta manera, es torna a fer l'aplicació del mètode de clustering considerant les categories de les variables de les que se'n disposa informació i per tant, que serveixen per caracteritzar l'accident al que fan referència.

El joc de dades amb les modificacions comentades conté 9544 mostres. A base de fer iteracions sobre l'arxiu, es conclou que hi ha una sèrie de variables que cal descartar.

Les característiques dels accidents que es mantenen i que es rebutgen per no identificar els clústers en l'agrupació definitiva són:

Variables mantingudes	**Variables descartades**
Tipus vehicle implicat	*Nom_mes*
Descripcio_sexe	*Descripcio_causa_mediata*
Torn	*Descripcio_situacio*
Edat	*Descripcio_victimitzacio*
Numero_lesionats_lleus	*Tipus_accident*
03.TM	*Descripcio_causa_vianant*
06.HRM	*Descripcio_color*
10.VVM10	*Descripcio_carnet*
	Numero_morts
	Numero_lesionats_greus
	Numero_vehicles_implicats

Variables mantingudes	Variables descartades
	Festiu
	07.PPT24H
	08.HPA
	09.RS24H
	Estat_mitja

Taula 5-1. Variables mantingudes i que no identifiquen els clústers. Joc de dades complet.

Subconjunt d'accidents greus

D'igual manera que al punt anterior, s'acaba concloent que l'únic algorisme del que se'n podrà treure uns resultats rellevants i operativament correctes és K-modes.

Procedint com amb el joc de dades complet, eliminant els accidents dels que no es coneixen les característiques del conductor, es treballa amb un total de 227 accidents. A base de fer iteracions sobre l'arxiu, es conclou que per arribar a l'agrupació definitiva algunes de les variables que es disposa al joc de dades són rebutjades.

Les característiques dels accidents que es mantenen i que es rebutgen per no identificar els clústers són:

Variables mantingudes	Variables descartades
Desc_Tipus_vehicle_implicat	*Districte*
Tipus_accident	*Nom_mes*
Torn	*Descripcio_causa_mediata*
Festiu	*Descripcio_sexe*
Numero_lesionats_lleus	*Descripcio_situacio*
Numero_vehicles_implicats	*Descripcio_victimitzacio*
Estat_mitja	*Descripcio_causa_vianant*
03.TM	*Descripcio_color*
06.HRM	*Descripcio_carnet*
	Numero_morts
	Numero_lesionats_greus
	07.PPT24H
	08.HPA
	09.RS24H
	10.VVM10
	Edat

Taula 5-2. Variables mantingudes i que no identifiquen els clústers. Subconjunt d'accidents greus.

Realitzades les proves i generats els resultats, cal fer-ne la presentació i descripció de les situacions que es detecta. A continuació, es descriuen els resultats obtinguts de les proves.

5.2. Resultats

Realitzades les proves sobre els tres jocs de dades complets, es veu que tan sols d'un d'aquestes se'n deriven proves vàlides. També es presenta els resultats obtinguts del subconjunt de dades format pels accidents més greus.

Joc de dades complet

Amb el joc de dades i les variables mantingudes a partir de la prova mitjançant K-modes s'arriba als resultats disposats a [Taula 5-4]. La caracterització d'un clúster per part d'una de les variables s'estableix si el valor d'aquesta variable és comú en una majoria dels elements del grup, es ressalta en color verd els valors majoritaris a [Taula 5-4].

Com es pot apreciar a la taula, més del 65% dels accidents es classifiquen en 3 grups, clústers 1 – 3. El clúster 1 és clarament el més rellevant, ja que agrupa un 47.7% dels accidents. Per sota d'aquest, hi ha els clústers 2 i 3 que agrupen un 10% i gairebé un 8% de les mostres, respectivament. No hi ha una clara classificació segons les causes de la mediació, de manera que es conclou que no es detecten situacions que assenyalin més cap a una causa de l'accident.

La distribució de la gravetat de l'accident entre els clústers es disposa a [Taula 5-1]. Es pot constatar que els diferents clústers difereixen en relació a la gravetat dels accidents, però en un grau molt baix si es compara amb les dades de [Taula 4-7], taula de proporció d'accidents segons la gravetat. Se'n pot derivar que la gravetat de l'accident no té una gran dependència de les condicions que el defineixen o que no es captura les variables que descriuen les condicions que fan més o menys greu un accident.

	Gravetat de l'accident				
	1. Lleu	**2. Assistencia**	**3. Hospitalitzat**	**4. Greu**	**5. Mort**
Clúster 1	2,1%	27,0%	68,6%	2,2%	0,2%
Clúster 2	2,2%	28,5%	67,0%	2,2%	0,0%
Clúster 3	2,4%	26,7%	67,6%	2,6%	0,7%
Clúster 4	4,2%	29,2%	65,6%	0,6%	0,3%
Clúster 5	2,9%	27,4%	67,3%	2,0%	0,3%
Clúster 6	1,2%	21,4%	73,3%	3,7%	0,4%
Clúster 7	2,9%	27,6%	66,7%	2,5%	0,4%
Clúster 8	2,3%	29,3%	66,8%	1,4%	0,2%
Clúster 9	2,2%	26,5%	69,5%	1,8%	0,0%

Taula 5-3. Distribució de la gravetat entre els clústers. Joc de dades complet.

La majoria dels clústers inclouen accidents de *motocicletes*, que és el tipus de vehicle que més accidents pateix. Tot i així, és remarcable que el clúster 4 agrupa el 6.5% dels accidents i que tenen categoria *Turisme*. També hi ha una majoria d'accidents on el conductor és un *home*, menys als clústers 2, 7 i 9 on és una *dona*. Aquests clústers contenen el 20.5% dels accidents i són força distints en la majoria dels aspectes.

Mentre en tots tres grups (clústers 2, 7 i 9) el vehicle que predomina als accidents és una *motocicleta*, els torns varien entre el *matí*, el *migdia* i el *vespre*. Les edats de les conductores es troben entre els 24 i els 31 anys als clústers 2 i 9, mentre que tenen entre 32 i 53 anys al grup 7 i les franges del dia són el *vespre*, el *matí* i el *migdia*, respectivament.

Els clústers 2 i 7 només tenen un *lesionat lleu*, mentre que al 9 n'hi ha dos o més. Pel que fa a les condicions climàtiques, tots tres grups apunten a combinacions de característiques diferents: *temperatura* baixa amb una alta *humitat*, *temperatura* molt baixa amb una *humitat* elevada i *velocitat del vent* molt baixa i una *temperatura* mitja, *humitat* baixa i *velocitat del vent* suau.

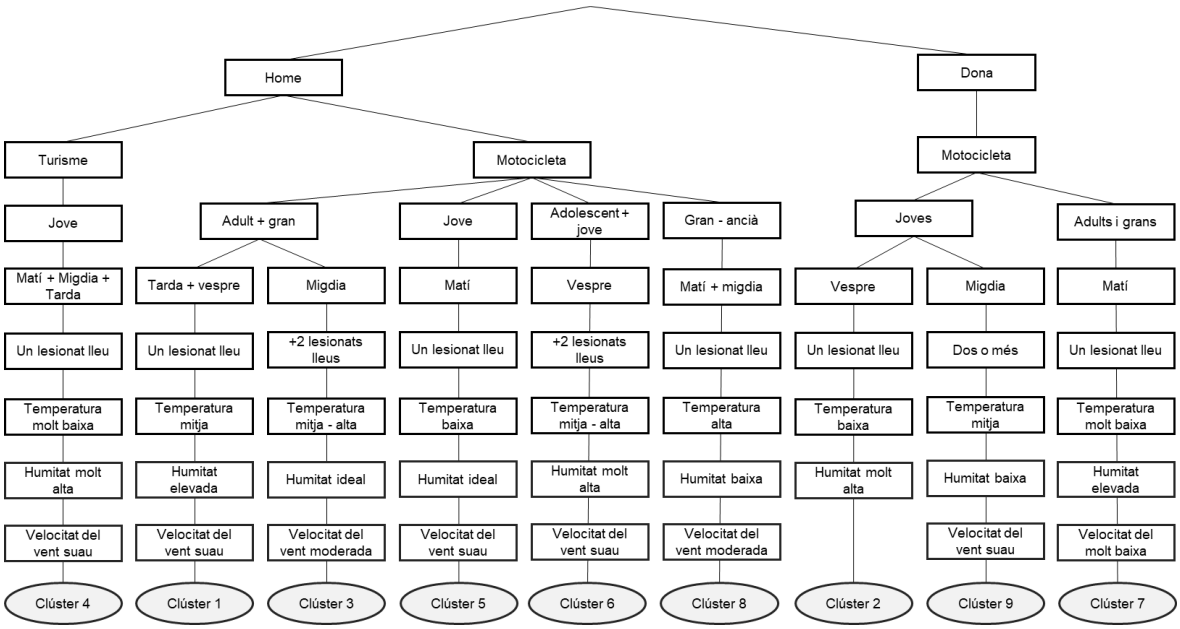


Figura 5-1. Característiques dels clústers. Joc de dades complet.

Als altres 6 clústers, integrats per accidents on els conductors són majoritàriament *homes*, el vehicle implicat en tots ells és una *motocicleta*, menys un grup on és un *turisme*. Aquest clúster inclou el 6.5% del total d'accidents.

Es donen majoritàriament entre el *matí* i la *tarda* de dies amb una *temperatura* molt baixa, una *humitat relativa* molt elevada i una *velocitat del vent* suau. El conductor és *jove*, entre 24 i 31 anys i només impliquen un *lesionat lleu*.

Sobre les altres cinc agrupacions, que impliquen un conductor *home* circulant amb una *motocicleta*, el clúster 1 agrupa una gran part dels accidents i és per això que la situació que presenta és molt significativa.

Es pot observar que fa referència a la *tarda* i el *vespre*, quan més circulació i desplaçaments hi ha. És explicable degut a que la gent torna a casa després de tot el dia d'activitat. Així ho confirma l'edat, que és la franja on es troba gran part de la població activa. La *temperatura mitja*, en combinació amb una *humitat* molt elevada fa que la sensació tèrmica sigui més alta que la temperatura real, de manera que poden aparèixer signes de cansament i possible fatiga deguts a una exposició prolongada, que segons la DGT estan relacionats directe o indirectament amb entre un 20% i un 30% de tots els accidents de trànsit [36].

Sobre les 4 agrupacions restants, els accidents es donen entre el *matí*, el *migdia* i el *vespre*. La resta de variables també varien: l'*edat* es mou entre *adolescents* fins a persones *grans* (14 – 53 anys), en dos dels clústers hi ha un *lesionat lleu* i als altres dos n'hi ha dos o més. És rellevant comentar que al clúster 6 es presenta una situació que és una *temperatura mitja* i alta i una *humitat relativa* molt alta. Com es comenta anteriorment, això fa augmentar la sensació tèrmica, que pot desencadenar en fatiga i cansament.

A continuació, es presenta la taula de resultats per a la prova de clustering definitiva sobre el joc de dades complet. Es ressalta en color verd aquells valors que són majoria al clúster, amb un 50% o més de presència entre els accidents agrupats. També es presenta el nombre d'accidents al grup i la proporció que representen sobre el total dels accidents estudiats. S'ordenen els clústers d'esquerra a dreta en ordre decreixent en nombre de mostres al clúster.

Variable	Categoria	Home entre 32 i 53 anys, accident en motocicleta durant la tarda i vespre, un lesionat lleu. Temperatura mitja, humitat elevada i velocitat del vent suau	Dona entre 24 i 31 anys, accident en motocicleta durant el vespre, un lesionat lleu. Temperatura baixa i humitat molt alta.	Home entre 32 i 53 anys, accident en motocicleta durant el migdia, dos o més lesionats lleus. Temperatura mitja-alta, humitat ideal i velocitat del vent moderada	Home entre 24 i 31 anys, accident en turisme durant el matí, migdia i tarda, un lesionat lleu. Temperatura molt baixa, humitat molt alta i velocitat del vent suau	Home entre 24 i 31 anys, accident en motocicleta durant el matí, un lesionat lleu. Temperatura baixa, humitat ideal i velocitat del vent suau	Home entre 14 i 31 anys, accident en motocicleta durant el vespre, dos o més lesionats lleus. Temperatura mitja - alta, humitat molt alta i velocitat del vent suau	Dona entre 32 i 53 anys, accident en motocicleta durant el matí i el migdia, un lesionat lleu. Temperatura molt baixa, humitat elevada i velocitat del vent molt baixa	Home entre 44 i 90 anys, accident en motocicleta durant el matí i el migdia, un lesionat lleu. Temperatura alta, humitat baixa i velocitat del vent moderada	Dona entre 24 i 31 anys, accident en motocicleta durant el migdia, dos o més lesionats lleus. Temperatura mitja, humitat baixa i velocitat del vent suau.
		Clúster 1	**Clúster 2**	**Clúster 3**	**Clúster 4**	**Clúster 5**	**Clúster 6**	**Clúster 7**	**Clúster 8**	**Clúster 9**
Número de d'accidents al cluster		4557	946	735	617	594	566	561	515	453
Proporció del total d'accidents		47,7%	9,9%	7,7%	6,5%	6,2%	5,9%	5,9%	5,4%	4,7%
Tipus vehicle implicat	**Ciclomotor**	9,4%	16,4%	7,6%	16,2%	7,2%	11,0%	12,7%	5,2%	9,7%
	Motocicleta	73,6%	67,9%	64,2%	24,0%	83,5%	64,7%	69,5%	77,9%	58,3%
	Turisme	10,0%	10,1%	17,8%	50,0%	4,0%	14,7%	13,4%	11,1%	23,4%
	Altres vehicles	7,1%	5,6%	10,3%	9,9%	5,2%	9,7%	4,5%	5,8%	8,6%
Descripcio_sexe	**Dona**	18,7%	70,8%	8,6%	14,3%	13,8%	12,0%	71,5%	15,5%	82,8%
	Home	81,3%	29,2%	91,4%	85,7%	86,2%	88,0%	28,5%	84,5%	17,2%
Torn	**Matí**	11,9%	7,3%	9,0%	13,3%	52,0%	7,4%	59,2%	37,5%	12,6%
	Migdia	16,3%	13,7%	57,3%	33,9%	15,0%	2,8%	5,2%	35,7%	55,8%
	Tarda	42,9%	17,4%	16,1%	10,0%	12,6%	7,4%	5,2%	19,2%	15,7%
	Vespre	19,2%	51,9%	9,1%	19,0%	12,0%	69,8%	16,9%	6,0%	6,8%
	Nit	7,0%	8,1%	6,9%	16,5%	7,6%	11,0%	9,4%	1,4%	7,1%
	Mitjanit	2,7%	1,5%	1,6%	7,3%	0,8%	1,6%	4,1%	0,2%	2,0%
Edat	**Adolescents**	8,2%	9,3%	9,0%	8,5%	7,7%	32,3%	10,5%	6,6%	6,6%
	Joves	16,68%	53,49%	11,8%	50,0%	56,9%	18,7%	14,3%	13,59%	51,66%
	Adults	47,07%	14,06%	16,5%	13,6%	15,2%	20,5%	20,7%	17,86%	19,43%
	Grans	17,09%	13,64%	47,6%	13,8%	11,6%	15,9%	47,2%	13,59%	13,47%
	Ancians	10,97%	9,51%	15,1%	14,1%	8,6%	12,5%	7,3%	48,35%	8,83%
Numero_lesionats_lleus	**Cap**	1,5%	1,6%	2,2%	0,3%	1,5%	3,5%	2,3%	1,2%	1,1%
	Un	82,9%	88,3%	31,6%	78,9%	89,4%	14,1%	87,2%	89,1%	37,7%
	Dos o més	15,5%	10,1%	66,3%	20,7%	9,1%	82,3%	10,5%	9,7%	61,1%
03.TM	**Molt baixa**	11,0%	14,4%	32,9%	62,2%	12,1%	13,6%	62,2%	11,1%	7,7%
	Baixa	13,2%	50,3%	12,8%	11,7%	56,1%	23,1%	10,0%	8,3%	7,7%
	Mitja	62,8%	24,4%	31,0%	15,2%	16,8%	30,2%	18,5%	24,5%	77,5%
	Alta	13,0%	10,9%	23,3%	10,9%	15,0%	33,0%	9,3%	56,1%	7,1%
06.HRM	**Baixa**	9,6%	10,4%	18,1%	12,5%	8,6%	7,2%	4,3%	64,9%	57,2%
	Ideal	12,1%	12,6%	50,9%	7,5%	68,0%	9,2%	14,1%	11,1%	10,8%
	Elevada	63,2%	21,8%	21,4%	18,3%	15,7%	24,4%	72,2%	19,4%	24,9%
	Molt alta	15,1%	55,3%	9,7%	61,8%	7,7%	59,2%	9,4%	4,7%	7,1%
10.VVM10	**Molt baixa**	9,9%	21,6%	12,7%	17,3%	8,8%	15,5%	63,3%	7,6%	6,8%
	Suau	67,7%	29,6%	18,5%	75,5%	86,4%	81,4%	28,9%	21,2%	72,8%
	Moderada	22,4%	48,8%	68,8%	7,1%	4,9%	3,0%	7,8%	71,3%	20,3%

Taula 5-4. Resultats clustering sobre el joc de dades complet.

ETSEIB

Resultats parcials. Joc de dades complet

A més a més del resultat que presenta la millor de les agrupacions de la prova definitiva, s'indica alguns resultats parcials que s'obtenen i que són prou rellevants per ser comentats. Aquests poden incloure algunes de les variables descartades per no ser característiques als resultats definitius, però sí ser-ho en alguna de les proves intermèdies realitzades.

Als resultats presentats a [Taula 5-5] s'hi distingeix tres agrupacions que presenten la característica de *tipus d'accident* 1, 2 i 3, respectivament [Annex II Figura II 11]. Es pot veure que en tots tres casos el vehicle implicat és una *motocicleta*. Al primer clúster s'agrupa una mica més del 4% dels accidents, el conductor és sempre una *dona* que condueix majoritàriament en torn de *nit* de dia *festiu*, on hi ha *dos vehicles implicats* i ha patit *accident de tipus 1*, col·lisió lateral, fronto-lateral o abast.

El segon clúster agrupa un 4% dels accidents, que succeeixen en *motocicleta* conduïda per un *home*, durant el *vespre* i de dia *no festiu*. Amb *dos vehicles implicats*, pateix un accident de caiguda o atropellament. L'últim clúster agrupa *accidents de tipus 3*, que són caigudes a l'interior del vehicle, xoc contra element estàtic i abast múltiple. Es tracta d'accidents de *motocicletes* conduïdes per un *home* durant el *matí* de dia *no festiu* i amb *tres o més vehicles implicats*. Representen el 3.5% del total d'accidents estudiats.

Variable	Categoria	Dona, conduïnt una motocicleta en torn de tarda de dia festiu i hi ha dos vehicles implicats. L'accident és de tipus 1	Home, conduïnt una motocicleta en torn de vespre de dia no festiu i hi ha dos vehicles implicats. L'accident és de tipus 2	Home, conduïnt una motocicleta en torn de matí de dia no festiu i hi ha tres o més vehicles implicats. L'accident és de tipus 3
Número d'accidents al cluster		394	365	323
Proporció del total d'accidents		4,1%	3,8%	3,4%
Tipus vehicle implicat	**Ciclomotor**	12,5%	10,1%	7,6%
	Motocicleta	59,6%	72,4%	75,4%
	Turisme	20,6%	11,2%	11,7%
	Altres vehicles	7,2%	6,3%	5,4%
Descripcio_sexe	**Dona**	100,0%	25,2%	31,5%
	Home	0,0%	74,8%	68,5%
Tipus_accident	**Tipus accident 1**	84,8%	27,1%	8,4%
	Tipus accident 2	2,5%	62,7%	1,2%
	Tipus accident 3	9,9%	9,6%	89,8%
	Tipus accident 4	2,8%	0,5%	0,6%
Torn	**Nit**	0,4%	2,8%	0,0%
	Migdia	5,9%	2,2%	49,2%
	Matí	3,1%	8,6%	50,7%
	Vespre	2,4%	79,4%	0,0%
	Tarda	88,2%	6,9%	0,1%
	Mitjanit	0,0%	0,1%	0,0%
Festiu	**No festiu**	30,5%	94,5%	81,4%
	Festiu	69,5%	5,5%	18,6%
Numero_vehicles_implicats	**Un**	4,6%	35,3%	8,0%
	Dos	86,80%	59,73%	6,50%
	Tres o més	8,63%	4,93%	85,45%

Taula 5-5. Resultats parcials, tipus accident. Joc de dades complet.

Els resultats parcials que es presenten tot seguit tenen relació amb l'aparició del *ciclomotor* com a vehicle implicat. Els clústers representen conjuntament el 6.1% dels accidents. Es distingeix entre conductor *home jove* en torn de *vespre* i *dona gran* i *anciana* durant el *matí* i el *migdia*. El *nombre de lesionats lleus* és un en tots dos clústers, la *temperatura* és baixa i la *humitat relativa* presenta una gran similitud. La *velocitat del vent* és molt baixa al clúster 1 i suau al clúster 2.

Variable	Categoria	Home jove, conduïnt un ciclomotor al vespre i amb un lesionat lleu. Temperatura mitja, humitat relativa elevada i molt alta, velocitat del vent molt baixa	Dona gran i anciana, conduïnt un ciclomotor al matí i migdia i amb un lesionat lleu. Temperatura mitja, humitat relativa molt alta, velocitat del vent suau
Número de d'accidents al cluster		376	211
Proporció del total d'accidents		3,9%	2,2%
Tipus vehicle implicat	**Ciclomotor**	63,5%	58,3%
	Motocicleta	21,2%	16,6%
	Turisme	6,9%	16,6%
	Altres vehicles	8,4%	8,5%
Descripcio_sexe	**Dona**	17,7%	91,9%
	Home	82,3%	8,1%
Torn	**Matí**	9,4%	45,5%
	Migdia	9,9%	11,4%
	Tarda	8,4%	15,2%
	Vespre	59,1%	19,9%
	Nit	10,3%	6,6%
	Mitjanit	3,0%	1,4%
Edat	**Adolescents**	14,3%	12,8%
	Joves	62,6%	10,0%
	Adults	12,3%	27,0%
	Grans	8,4%	17,5%
	Ancians	2,5%	32,7%
Numero_lesionats_lleus	**Cap**	1,5%	1,9%
	Un	87,68%	88,63%
	Dos o més	10,84%	9,48%
03.TM	**Molt baixa**	6,9%	6,2%
	Baixa	7,4%	6,6%
	Mitja	82,8%	79,6%
	Alta	3,0%	7,6%
06.HRM	**Baixa**	33,0%	6,6%
	Ideal	13,3%	19,0%
	Elevada	27,1%	21,3%
	Molt alta	26,6%	53,1%
10.VVM10	**Molt baixa**	67,5%	4,7%
	Suau	10,8%	85,8%
	Moderada	21,7%	9,5%

Taula 5-6. Resultats parcials, tipus de vehicle. Joc de dades complet.

Per acabar, es comenta uns resultats agrupen segons la variable *districte*. En aquest cas, s'identifica i agrupa un volum important d'accidents que comparteixen un mateix *districte*. Aquest és el *districte* de l'Eixample.

Els accidents en tots dos casos tenen com a vehicle implicat una *motocicleta*, mentre que en uns el conductor és un *home* i als altres és una *dona*. També es distingeix que al clúster 1 s'agrupa *accidents tipus 1* en torn de *nit festiu* per part d'*homes adults* i *grans* i els accidents al clúster 2 es caracteritzen per patir *accident tipus 3* en torn de *tarda no festiu* per part

d'*homes joves*. Al clúster 1 els accidents involucren *dos vehicles implicats* amb *dos o més lesionats lleus*, mentre que al clúster 2 tenen *tres o més vehicles implicats* i *un lesionat lleu*.

Variable	Categoria	Dona adulta i gran, conduïnt una motocicleta en horari de nit de dia festiu al districte de l'Eixample. Dos o més lesionats lleus i dos vehicles implicats. L'accident és de tipus 1	Home jove, conduïnt una motocicleta a la tarda de dia no festiu pel districte de l'Eixample. Un lesionat lleu i tres o més vehicles implicats. L'accident és de tipus 3
Número de d'accidents al cluster		473	853
Proporció del total d'accidents		5,0%	8,9%
Districte	**Gràcia**	5,3%	2,6%
	Sant Martí	7,8%	4,1%
	Sants-Montjuïc	4,4%	7,0%
	Sarrià-Sant Gervasi	7,6%	7,6%
	Ciutat Vella	2,5%	3,0%
	Eixample	53,1%	57,6%
	Districte desconegut	0,0%	0,0%
	Les Corts	5,5%	5,9%
	Sant Andreu	4,2%	4,1%
	Nou Barris	5,3%	2,8%
	Horta-Guinardó	4,2%	5,3%
Tipus vehicle implicat	**Ciclomotor**	12,5%	7,6%
	Motocicleta	59,6%	75,4%
	Turisme	20,6%	11,7%
	Altres vehicles	7,2%	5,4%
Descripcio_sexe	**Dona**	100,0%	31,5%
	Home	0,0%	68,5%
Tipus_accident	**Tipus accident 1**	84,8%	8,4%
	Tipus accident 2	2,5%	1,2%
	Tipus accident 3	9,9%	89,8%
	Tipus accident 4	2,8%	0,6%
Torn	**Matí**	0,4%	0,0%
	Migdia	5,9%	49,2%
	Tarda	3,1%	50,7%
	Vespre	2,4%	0,0%
	Nit	88,2%	0,1%
	Mitjanit	0,0%	0,0%
Festiu	**No festiu**	30,5%	81,4%
	Festiu	69,5%	18,6%
Edat	**Adolescents**	10,8%	7,2%
	Joves	13,5%	67,1%
	Adults	19,5%	0,0%
	Grans	42,3%	17,2%
	Ancians	14,0%	8,6%
Numero_lesionats_lleus	**Cap**	4,7%	1,1%
	Un	16,5%	93,8%
	Dos o més	78,9%	5,2%
Numero_vehicles_implicats	**Un**	4,6%	8,0%
	Dos	86,80%	6,50%
	Tres o més	8,63%	85,45%

Taula 5-7. Resultats parcials, districte. Joc de dades complet.

El que es desprèn dels resultats obtinguts de les proves fetes sobre el joc de dades complet és que s'identifica tres grups molt rellevants en matèria d'accidents ocorreguts a Barcelona, ja que contenen més del 65% de les mostres, mitjançant els que es detecta aquelles situacions repetitives que donen peu a aquests sinistres. De la mateixa manera, s'exposen situacions menys representatives però igualment vàlides i amb valor a l'hora d'identificar escenaris on es desenvolupen accidents de forma habitual a la ciutat.

Subconjunt d'accidents greus

Els anteriors resultats presenten unes agrupacions centrades en la gravetat dels accidents *3. Hospitalitzat* i no descriuen les situacions relacionades amb els accidents més greus, que són els que suposen unes conseqüències més severes i que tenen major impacte. Donat que es disposa de la variable *Descripcio_victimitzacio*, mitjançant la qual es pot separar els accidents, com es descriu en capítols anteriors s'aplica el model d'execució de proves sobre el joc de dades que conté els accidents classificats com *4. Greu* i *5. Mort.* Igual que es fa a les proves amb tots els accidents, es retira del model aquells accidents que presenten la categoria "*desconegut*" a les variables *Descripcio_sexe* i *Edat.*

D'aquesta manera, es busca definir les situacions que propicien els accidents d'alta gravetat. Es treballa amb un total de 227 accidents. Seguint el mateix esquema que en les proves anteriors, detallat a la [Figura 3-1], s'arriba a una solució que agrupa els accidents en 7 clústers. Aquests es presenten a la [Taula 5-11] amb les situacions identificades resumides a la [Taula 5-9] i [Taula 5-10].

En primer lloc, avaluant la gravetat dels grups d'accidents, com es disposa a la [Taula 5-8], es pot constatar que els diferents clústers difereixen en relació a la gravetat, però en un grau molt baix. Les proporcions entre *4. Greu* i *5. Mort* de les dades originals són 90% i 10% respectivament, de manera que els resultats obtinguts presenten una relació entre categories similar. Analitzant-ne la distribució, no es pot afirmar en cap cas que s'agrupi els accidents mortals en cap dels clústers, ja que no és majoria en cap d'aquests, però sí que es pot comentar que les situacions presentades als clústers 3 i 7 tenen un volum d'accidents mortals més elevat.

	Gravetat de l'accident	
	4. Greu	**5. Mort**
Clúster 1	92%	8%
Clúster 2	87%	13%
Clúster 3	85%	15%
Clúster 4	89%	11%
Clúster 5	100%	0%
Clúster 6	89%	11%
Clúster 7	79%	21%

Taula 5-8. Distribució de la gravetat entre els clústers.

Per comentar els resultats de les agrupacions pren especial rellevància el volum d'accidents agrupats en cadascuna de les situacions. En els quatre primers clústers s'agrupa més del 77.5% dels accidents, dels que s'identifiquen les situacions que els propicien.

Clúster 1	Clúster 2	Clúster 3
Accidents de tipus 1 durant la tarda i el vespre de dia no festiu on el conductor té entre 24 i 43 anys. No hi ha cap lesionat lleu, hi ha dos vehicles implicats i estat del trànsit fluid. Temperatura mitja i humitat elevada.	Accidents entre febrer i març i durant la tarda i el vespre de dia no festiu on l'accident és de tipus 3. L'edat del conductor es troba entre 44 i 53 anys, no hi ha lesionats lleus, hi ha dos vehicles implicats i el trànsit és fluid. La temperatura és molt baixa.	Accidents entre març i juny durant la tarda i el vespre de dia no festiu on l'accident és de tipus 2. El conductor té entre 24 i 43 anys, hi ha un lesionat lleu i un vehicle implicat. El trànsit és fluid, la temperatura és mitja i humitat elevada.

Taula 5-9. Situacions dels clústers 1-3 amb accidents greus.

Els accidents inclosos als tres primers clústers coincideixen en la franja del dia, que es troba entre la *tarda* i el *vespre*. També en que és una franja classificada com a *no festiu* i que l'estat del trànsit és *fluid*. En canvi, divergeixen en l'època de l'any: el primer clúster no presenta una agrupació rellevant en aquesta dimensió.

El 2 ajunta una majoria d'accidents entre *febrer* i *març* i el 3 entre *març* i *juny*. El tipus d'accident també és diferent a les tres situacions, variant entre accidents de tipus *1, 3* i *2* respectivament. El primer clúster, que agrupa un 40% de les dades i per això és especialment rellevant, presenta accidents on el conductor té una edat entre *24 i 43 anys*, no hi ha *cap lesionat lleu* i hi ha *dos vehicles implicats*. La *temperatura* és mitja i la *humitat relativa* a l'ambient és *elevada*.

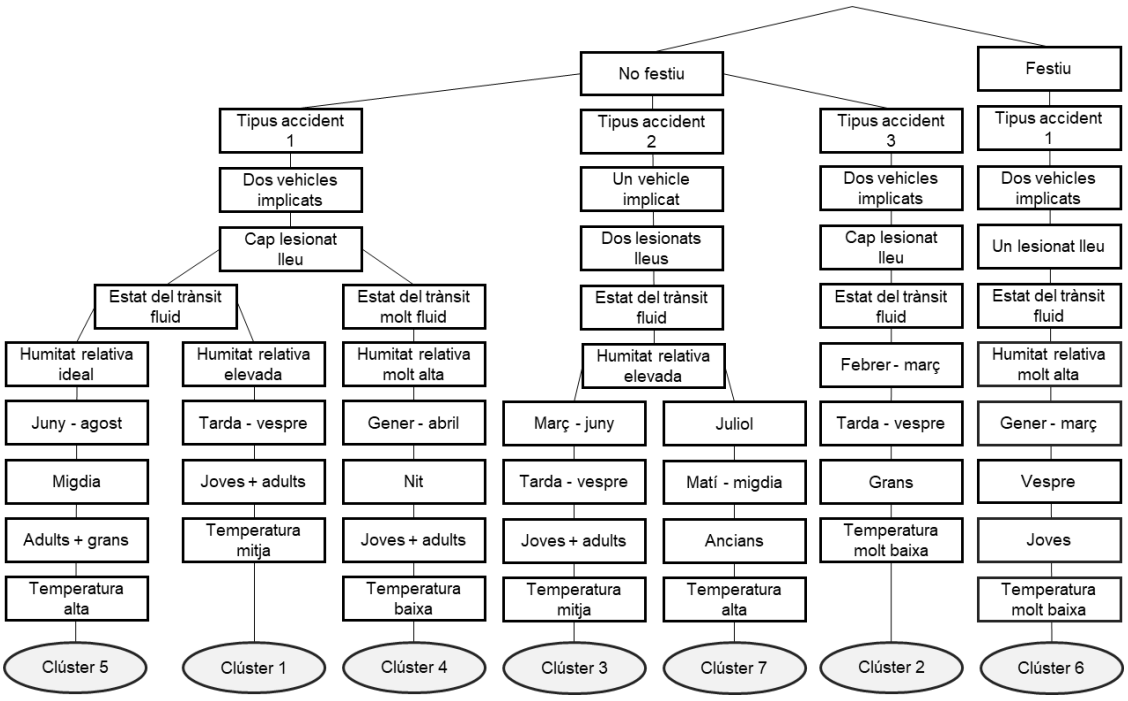


Figura 5-2. Característiques dels clústers. Subconjunt d'accidents greus.

En canvi, el clúster 2 agrupa un 13,2% dels accidents on el conductor té entre *44 i 53 anys*, tampoc hi ha *cap lesionat lleu* i hi ha *dos vehicles implicats*. La temperatura és *molt baixa*. El clúster 3, amb gairebé un 12% dels accidents si que presenta *un lesionat lleu* i *un vehicle implicat* als sinistres agrupats. La temperatura és *mitja* i la humitat relativa és *elevada*.

Clúster 4	Clúster 5	Clúster 6	Clúster 7
Accidents entre gener i abril durant la nit no festiva on l'accident és de tipus 1. El conductor té entre 24 i 43 anys, no hi ha cap lesionat lleu i hi ha dos vehicles implicats. El trànsit és molt fluid, la temperatura és baixa i la humitat relativa és molt alta.	Accident entre juny i agost durant el migdia de dia no festiu on l'accident és de tipus 1. El conductor té entre 32 i 53 anys, no hi ha cap lesionat lleu i hi ha dos vehicles implicats. El trànsit és fluid, la temperatura és alta i la humitat és ideal.	Accidents entre gener i març durant el vespre de dia festiu on l'accident és de tipus 1. El conductor té entre 24 i 31 anys, hi ha un lesionat lleu i dos vehicles implicats. L'estat del trànsit és fluid, la temperatura és molt baixa i la humitat relativa és molt alta.	Accidents durant els matins i migdia de juliol en torn no festiu on l'accident és de tipus 2. El conductor té més de 54 anys, no hi ha lesionats lleus i hi ha un vehicle implicat. L'estat del trànsit és fluid, la temperatura és alta i la humitat relativa és elevada.

Taula 5-10. Situacions dels clústers 4-7 amb accidents greus.

El quart clúster, que també agrupa gairebé un 12% dels accidents, presenta unes mostres que es donen principalment de *gener* a *abril*, siguin *accidents de tipus 1* durant la *nit no festiva*. Els conductors tenen entre *24 i 43 anys*, igual que els clústers 1 i 3, *cap lesionat lleu*, *dos vehicles implicats* i un estat del trànsit *molt fluid*. Finalment, presenten una temperatura *baixa* i una humitat relativa *molt alta*.

El clúster 5 agrupa un 8.5% dels accidents, accidents que succeeixen entre *juny* i *agost*, època estival. Igual que els del clúster 7, amb una proporció del 6.2%, que agrupa principalment accidents durant el *juliol*. Al 5, són *accidents de tipus 1*, en torn de *migdia no festiu* i on els conductors del vehicle tenen entre *32 i 53 anys*. No hi ha *lesionats lleus* i sí que es presenten *dos vehicles implicats*. L'estat del trànsit és *fluid*, la temperatura és *alta* i la humitat és *ideal*. Al 7, els accidents també són en una franja matinal, entre *matí* i *migdia*. Es donen en torn *no festiu* i, a diferència del clúster 5, són de tipus 2. Els conductors són *ancians*, de més de 54 anys, no hi ha *lesionats lleus* i sí *un vehicle implicat*. L'estat del trànsit és *fluid*, la temperatura és *alta* i la humitat relativa és *elevada*.

Finalment, el clúster 6, amb un 7.9% dels accidents, agrupa accidents de tipus 1 durant els mesos de gener a març, en torn de vespre festiu. Els conductors són principalment joves, entre 24 i 31 anys. L'estat del trànsit és fluid, la temperatura és alta i la humitat relativa és elevada. Això, juntament amb l'edat del conductor, pot desencadenar en un estat de cansament i fatiga que propicia l'accident, a més a més de ser en torn de vespre i de franja classificada com a festiva. Hi ha un lesionat lleu i dos vehicles implicats.

Variable	Categoria	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Clúster 6	Clúster 7
Número de d'accidents al cluster		92	30	27	27	19	18	14
Proporció del total d'accidents		40,5%	13,2%	11,9%	11,9%	8,4%	7,9%	6,2%
Mes	Gener	13,0%	3,3%	7,4%	7,4%	15,8%	27,8%	0,0%
	Febrer	1,1%	40,0%	11,1%	11,1%	0,0%	16,7%	14,3%
	Març	5,4%	10,0%	7,4%	0,0%	0,0%	11,1%	7,1%
	Abril	9,8%	3,3%	25,9%	37,0%	0,0%	0,0%	0,0%
	Maig	18,5%	0,0%	11,1%	7,4%	0,0%	0,0%	0,0%
	Juny	8,7%	3,3%	11,1%	0,0%	31,6%	0,0%	0,0%
	Juliol	9,8%	3,3%	3,7%	7,4%	10,5%	11,1%	50,0%
	Agost	10,9%	0,0%	7,4%	0,0%	26,3%	0,0%	21,4%
	Setembre	5,4%	6,7%	3,7%	11,1%	5,3%	11,1%	0,0%
	Octubre	8,7%	10,0%	7,4%	3,7%	5,3%	0,0%	7,1%
	Novembre	4,3%	10,0%	3,7%	3,7%	0,0%	11,1%	0,0%
	Desembre	4,3%	10,0%	0,0%	11,1%	5,3%	11,1%	0,0%
Tipus_accident	Tipus accident 1	75,0%	23,3%	3,7%	66,7%	94,7%	66,7%	0,0%
	Tipus accident 2	1,1%	16,7%	74,1%	7,4%	0,0%	5,6%	64,3%
	Tipus accident 3	10,9%	50,0%	14,8%	14,8%	5,3%	22,2%	14,3%
	Tipus accident 4	13,0%	10,0%	7,4%	11,1%	0,0%	5,6%	21,4%
Torn	Matí	13,0%	23,3%	3,7%	7,4%	10,5%	0,0%	42,9%
	Migdia	9,8%	6,7%	14,8%	3,7%	52,6%	11,1%	21,4%
	Tarda	22,8%	40,0%	44,4%	7,4%	21,1%	0,0%	7,1%
	Vespre	46,7%	16,7%	18,5%	3,7%	10,5%	66,7%	21,4%
	Nit	4,3%	6,7%	11,1%	63,0%	0,0%	5,6%	0,0%
	Mitjanit	3,3%	6,7%	7,4%	14,8%	5,3%	16,7%	7,1%
Festiu	No festiu	81,5%	83,3%	88,9%	66,7%	89,5%	16,7%	85,7%
	Festiu	18,5%	16,7%	11,1%	33,3%	10,5%	83,3%	14,3%
Edat	Adolescents	3,3%	10,0%	14,8%	14,8%	5,3%	0,0%	7,1%
	Joves	15,2%	10,0%	22,2%	48,1%	15,8%	55,6%	21,4%
	Adults	47,8%	10,0%	37,0%	14,8%	21,1%	11,1%	7,1%
	Grans	15,2%	53,3%	14,8%	18,5%	47,4%	11,1%	7,1%
	Ancians	18,5%	16,7%	11,1%	3,7%	10,5%	22,2%	57,1%
Numero_lesionats_lleus	Cap	80,4%	86,7%	14,8%	81,5%	84,2%	22,2%	64,3%
	Un	15,2%	0,0%	77,8%	18,5%	10,5%	50,0%	28,6%
	Dos o més	4,3%	13,3%	7,4%	0,0%	5,3%	27,8%	7,1%
Numero_vehicles_implicats	Un	5,4%	10,0%	85,2%	18,5%	0,0%	0,0%	64,3%
	Dos	79,3%	86,7%	14,8%	74,1%	94,7%	88,9%	28,6%
	Tres o més	15,2%	3,3%	0,0%	7,4%	5,3%	11,1%	7,1%
Estat_mitja	Sense dades	1,1%	0,0%	3,7%	0,0%	0,0%	0,0%	0,0%
	Molt fluid	12,0%	20,0%	22,2%	88,9%	5,3%	16,7%	28,6%
	Fluid	85,9%	76,7%	74,1%	11,1%	94,7%	83,3%	71,4%
	Dens	1,1%	3,3%	0,0%	0,0%	0,0%	0,0%	0,0%
03.TM	Molt baixa	13,0%	63,3%	11,1%	18,5%	10,5%	50,0%	21,4%
	Baixa	7,6%	13,3%	22,2%	51,9%	5,3%	27,8%	7,1%
	Mitja	62,0%	16,7%	59,3%	25,9%	26,3%	11,1%	0,0%
	Alta	17,4%	6,7%	7,4%	3,7%	57,9%	11,1%	71,4%
06.HRM	Baixa	8,7%	36,7%	7,4%	3,7%	15,8%	5,6%	0,0%
	Ideal	10,9%	16,7%	18,5%	3,7%	68,4%	16,7%	21,4%
	Elevada	56,5%	23,3%	51,9%	40,7%	15,8%	27,8%	71,4%
	Molt alta	23,9%	23,3%	22,2%	51,9%	0,0%	50,0%	7,1%

Taula 5-11. Resultats clustering sobre el joc de dades de més gravetat.

6. Planificació temporal

Al present capítol es realitza una planificació de totes aquelles activitats que conformen el projecte, sense entrar al detall de l'operativa o de les tasques puntuals a realitzar. D'aquesta manera, s'aconsegueix una disposició de les tasques principals en forma de calendari, presentant la possibilitat d'identificar prelacions entre activitats i la distribució d'aquestes al llarg del projecte. Addicionalment, és útil per identificar els recursos requerits en cada fase de treball, alhora que conèixer les hores dedicades i elaborar un pressupost del projecte.

S'estableix unes fites importants per al treball, com són l'inici i finalització del projecte. L'inici es marca al dia 10 de gener de 2022, alhora que la finalització es programa per al dia 24 d'abril de 2022.

També es defineixen fites intermèdies, corresponents a la finalització dels principals blocs de treball definits i que serveixen de punts de control en relació al desenvolupament del projecte.

Sobre les hores dedicades, el total de crèdits matriculats al present treball final d'estudis és de 24 unitats, ja que es cursa un doble màster. Tenint en compte que la equivalència són 30 hores per cada crèdit [37], agrupant tots els tipus de dedicació descrits, es disposa de 720 hores de treball.

Aquestes hores distribuïdes entre les fites d'inici i finalització del projecte, suposen dedicar de mitjana diària unes 6.9 hores. Aquest és un ritme lleugerament superior a una jornada laboral completa, ja que també es treballa en cap de setmana i dies festius.

El total d'hores de treball es divideixen entre esforços dedicats al disseny i investigació, de major valor afegit, tasques d'execució i verificació, que són més repetitives i manuals i tasques d'elaboració dels documents, que consisteixen en la preparació de tota la documentació necessària, alhora que redactar els documents i disposar-los amb el format adequat. A la planificació es detalla l'assignació d'aquests a cada tasca.

A partir de la planificació realitzada, s'identifica que la utilització de recursos es distribueix de la següent manera:

Tasques	Disseny i investigació	Execució i verificació	Elaboració dels documents	**Total dedicació**
Dedicació [%]	9%	73%	18%	100%
Dedicació [*h*]	66	524	130	720

Taula 6-1. Distribució de la dedicació.

Els blocs de treball s'indiquen de color blau i integren un conjunt d'activitats amb relació entre elles, tant de precedència i dependència com a nivell temàtic. Els blocs amb més

dedicació en la dimensió d'hores són la programació del codi, l'elaboració documental i les activitats prèvies, en aquest ordre. Aquestes conformen l'esquelet del treball, ja que en gran part de la realització del projecte és requisit treballar contra el programa de càlcul, fent-hi modificacions o obtenint-ne resultats.

Passa igual amb l'elaboració documental: es planteja un mètode centrat en la realització de tasques i que encara que en alguns casos calgui tirar enrere i realitzar modificacions, en molts altres moments és possible realitzar la tasca i plasmar el procediment seguit i els resultats obtinguts a la documentació.

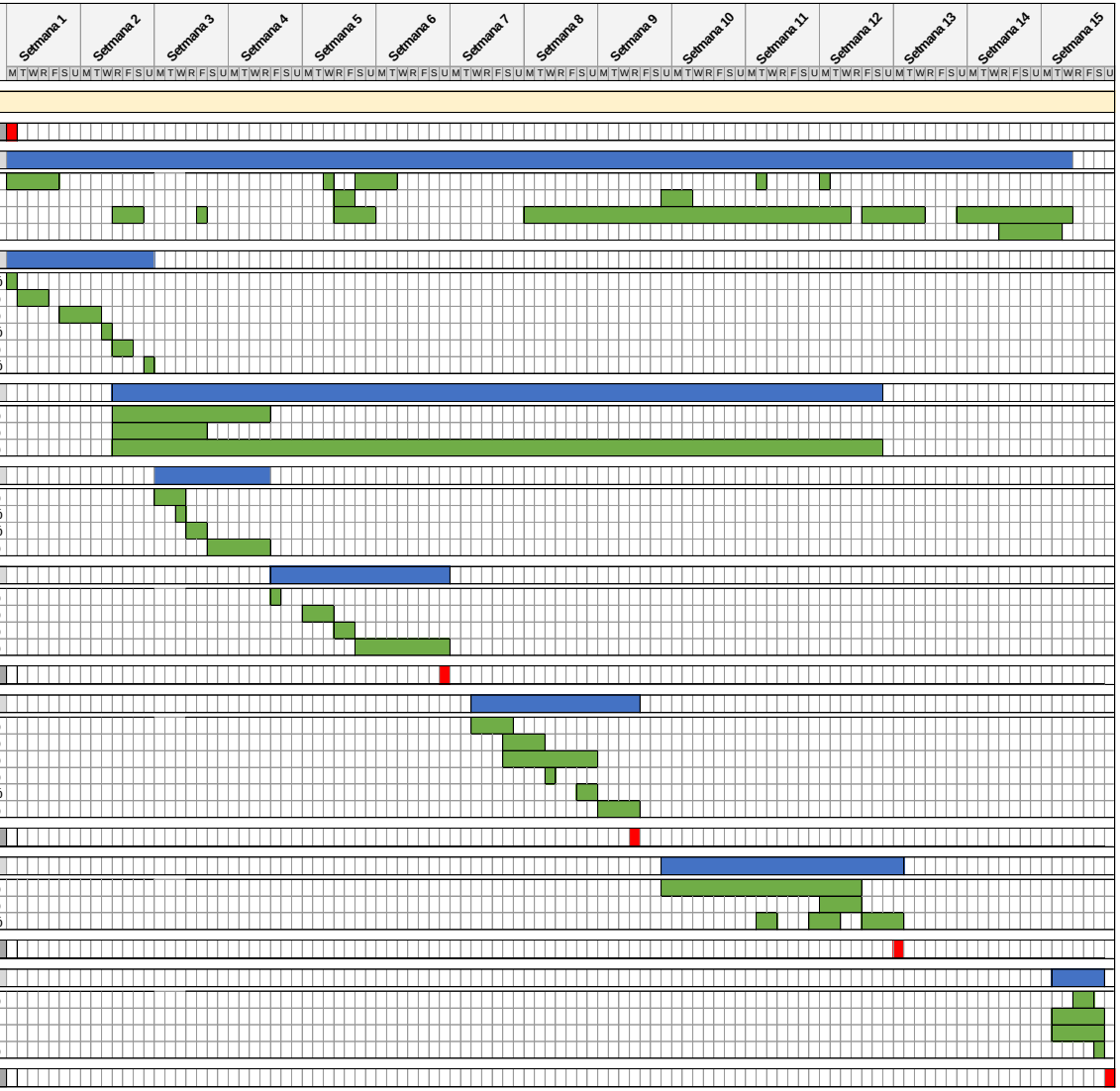
També es pot veure que les tasques, indicades en color verd, que suposen més esforços són l'elaboració del codi, la redacció de la memòria, l'execució de les proves i l'anàlisi descriptiu de les variables.

Degut a que l'autor del projecte no ha treballat mai amb tècniques de Machine Learning com el clustering ni ha tingut cap formació prèvia en llenguatge de programació Python, cal dedicar una part del temps a formació i aprenentatge d'ús de les eines i conceptes.

En tot moment es manté un marge de seguretat en relació a la dedicació estimada a cadascuna de les tasques, per permetre flexibilitat i marge de maniobra a l'hora d'haver d'allargar alguna de les tasques programades.

A la següent pàgina es mostra el diagrama de Gantt generat amb les tasques realitzades i els recursos assignats, juntament amb les fites marcades per al projecte.

Nom de la tasca	Dia inici	Dia final	Dedicació [h]	Recurs assignat
Treball Final de Màster	**10/01/2022**	**24/04/2022**	**720**	-
Inici del projecte	10/01/2022			-
Elaboració documental	10/01/2022	20/04/2022	105	-
Preparació de la documentació	10/01/2022	28/03/2022	8	Documentació
Elaboració de les plantilles d'anàlisi	10/02/2022	15/03/2022	10	Documentació
Redacció de la memòria	20/01/2022	20/04/2022	67	Documentació
Redacció dels annexes	14/04/2022	19/04/2022	20	Documentació
Activitats prèvies	10/01/2022	23/01/2022	80	-
Definició del projecte	10/01/2022	10/01/2022	8	Disseny i investigació
Formació temàtica	11/01/2022	13/01/2022	25	Execució i verificació
Estudi d'antecedents	15/01/2022	18/01/2022	16	Execució i verificació
Definició d'objectius i abast	19/01/2022	19/01/2022	8	Disseny i investigació
Necessitats d'informació	20/01/2022	21/01/2022	15	Execució i verificació
Plantejament de possibles mètodes	23/01/2022	23/01/2022	8	Disseny i investigació
Programació del codi	20/01/2022	02/04/2022	251	-
Formació Python	20/01/2022	03/02/2022	30	Execució i verificació
Formació Google Colab	20/01/2022	28/01/2022	21	Execució i verificació
Elaboració del codi	20/01/2022	02/04/2022	200	Execució i verificació
Diseny i definició del mètode	24/01/2022	03/02/2022	41	-
Revisió de la literatura	24/01/2022	26/01/2022	12	Execució i verificació
Anàlisi de recursos	26/01/2022	26/01/2022	6	Disseny i investigació
Plantejament de la metodologia	27/01/2022	28/01/2022	15	Disseny i investigació
Validació del mètode	29/01/2022	03/02/2022	8	Execució i verificació
Primer bloc de treball	04/02/2022	20/02/2022	57	-
Recull de dades	04/02/2022	04/02/2022	3	Execució i verificació
Adequació i transformació	07/02/2022	09/02/2022	12	Execució i verificació
Anàlisi de les dades	10/02/2022	11/02/2022	7	Execució i verificació
Anàlisi descriptiu	12/02/2022	20/02/2022	35	Execució i verificació
Finalització primer bloc de treball	20/02/2022			-
Segon bloc de treball	23/02/2022	10/03/2022	76	-
Anàlisi de correlació	23/02/2022	26/02/2022	12	Execució i verificació
Anàlisi de dependència	26/02/2022	01/03/2022	17	Execució i verificació
Anàlisi qualitatiu	26/02/2022	06/03/2022	25	Execució i verificació
Anàlisi d'algorismes de clustering	02/03/2022	02/03/2022	6	Execució i verificació
Definició de proves	05/03/2022	06/03/2022	6	Disseny i investigació
Transformacions de dades	07/03/2022	10/03/2022	10	Execució i verificació
Finalització segon bloc de treball	10/03/2022			-
Tercer bloc de treball	13/03/2022	04/04/2022	70	-
Execució de proves tots els accidents	13/03/2022	31/03/2022	40	Execució i verificació
Execució de proves accidents greus	28/03/2022	31/03/2022	15	Execució i verificació
Exposició i descripció dels resultats	22/03/2022	04/04/2022	15	Disseny i investigació
Finalització tercer bloc de treball	04/04/2022			-
Tancament	21/04/2022	23/04/2022	40	-
Estudi de la memòria	21/04/2022	22/04/2022	10	Execució i verificació
Revisió de la documentació	19/04/2022	23/04/2022	20	Documentació
Retocs i petites modificacions	19/04/2022	23/04/2022	5	Documentació
Tancament definitiu de la memòria	23/04/2022	23/04/2022	5	Execució i verificació
Finalització del projecte	24/04/2022			-



Taula 6-2. Planificació temporal del projecte.

7. Estudi econòmic

Al present apartat es detalla el cost de realització del present estudi, considerant tot allò relacionat amb el material utilitzat, el software que ha estat necessari i el temps invertit en desenvolupar tot el treball.

Aquest temps de treball, com es detalla a la planificació, es distingeix segons la tasca realitzada en hores de disseny i investigació, hores d'execució i hores d'elaboració dels documents que cal entregar.

Tal com es comenta al capítol anterior, es disposa d'un total de 720 hores a la realització del projecte. La distribució identificada a través de la planificació, segons el tipus de tasques realitzades, es presenta a [Taula 6-1].

Sobre els materials utilitzats que tenen una vida útil de més duració que el propi treball, s'imputa una part proporcional a aquesta.

- Lenovo ThinkPad T480 + paquet MS Office: S'estima que té una vida útil de 3 anys. Considerant un ús mig diari de 8h (una jornada laboral completa) i 253 dies laborables per any, la vida útil d'un ordinador portàtil és de 6072 hores. S'estableix la imputació com 720 hores / 6072 hores.
- Material d'oficina: Es considera com un 30% del cost total de material.

Cost del projecte				
Eina de treball	**Unitats** [*u*]	**Imputació** [%]	**Cost per unitat** [*€/u*]	**Cost total** [*€*]
Lenovo ThinkPad T480	1	12	1.141,91	137,02
Paquet Microsoft Office	1	12	50,00	6,00
Material d'oficina	1	100	13,70	41,10
			Total material [*€*]:	184,12

Tasques	**Temps** [h]	**Cost per unitat** [*€/h*]	**Cost total** [*€*]
Disseny i investigació	66	25,00	1.650,00
Execució i verificació	524	15,00	7.860,00
Elaboració dels documents	130	12,00	1.560,00
		Total tasques [*€*]:	11.070,00
		Total projecte [*€*]:	11.254,12

Taula 7-1. Detall del cost del projecte.

8. Impacte econòmic, social i ambiental

En aquesta secció del projecte s'adreça els impactes derivats de la realització d'aquest, ens les tres dimensions: econòmica, social i ambiental.

8.1. Impactes negatius

En relació als aspectes negatius, el projecte no presenta impacte negatiu en l'àmbit social, donat que es tracta amb dades públiques, sense relació amb persones físiques o amb tractament de dades sensibles. La publicació de la present memòria no suposa un risc per a cap persona, entitat o empresa. Pel que fa a l'impacte econòmic, es quantifica un cost de realització del projecte de 11.254,12 €, principalment derivats de la mà d'obra d'enginyeria. No hi ha altres impactes econòmics negatius en relació al projecte.

Finalment, considerant l'impacte mediambiental, el projecte implica un consum de recursos, com ara l'electricitat o altres matèries relacionades amb els materials i estris consumits. No es genera residus extraordinaris derivats del projecte. Es considera negligible les emissions derivades del consum d'energia o els petits residus domèstics generats.

8.2. Impactes positius

Els impactes positius en la dimensió social són derivats del propi objectiu del treball. Es busca identificar situacions habituals en les que es desenvolupen accidents a la ciutat de Barcelona. La consecució d'aquests objectius, a través dels resultats obtinguts, obre la porta a adreçar aquestes situacions, mitjançant les polítiques que millor hi encaixin amb l'objectiu de reduir els sinistres a les vies de circulació.

En segon lloc, identificar aquestes casuístiques impacta, en cas de ser emprades per reduir els accidents de trànsit, en una reducció dels sinistres i la problemàtica a les vies de circulació, que presenta un efecte directe sobre les dimensions econòmica i mediambiental. Per una banda, es redueixen les emissions per part dels vehicles que es troben aturats a causa dels accidents i totes les tasques requerides per atendre als actors afectats: retencions, mobilització de vehicles de rescat o socors, agents de mediació i altres. També, té un efecte sobre la quantitat de residus que apareixen com a conseqüència d'un accident, com poden ser parts dels vehicles implicats o elements de la calçada malmesos.

En relació amb la dimensió econòmica, la possible disminució d'accidents desencadena en una clara reducció de tots els costos derivats dels sinistres, com reparacions, despeses mèdiques, de reparació de danys o substitució d'elements. El projecte presenta un balanç molt positiu en relació als impactes sobre les tres dimensions analitzades.

9. Conclusions

Aquest projecte s'emmarca en la recerca per la millora i reducció de l'accidentalitat de trànsit. Concretament, és el primer estudi que aplica tècniques de clustering per a la classificació i detecció de situacions comunes entre les dades d'accidents de trànsit de la ciutat de Barcelona. Amb l'objectiu de detectar aquestes situacions i poder adreçar-les per part dels òrgans competents, es defineix un mètode per agrupar els accidents en clústers significatius segons les seves pròpies característiques i les de l'entorn, presentant els resultats obtinguts.

A partir dels coneixements adquirits durant els màsters cursats i una part important de revisió d'antecedents i literatura, metodologia i aprenentatge autònom sobre la matèria s'ha definit un procediment d'estudi centrat en tres algorismes de clustering per aplicar sobre les dades i obtenir els resultats. També s'ha fet el tractament, anàlisi de variables i transformacions per aconseguir el joc de dades utilitzat.

Finalment, realitzat el procés de tractament de dades corresponent i definida la metodologia i requeriments per aplicar-la, s'executa el procediment de prova dissenyat sobre dos conjunts de dades: tots els accidents per una banda i per triplicat, aplicant un algorisme diferent en cadascun dels cicles; el subconjunt d'accidents de màxima gravetat per l'altra. D'aquesta manera, a més a més d'identificar les situacions relacionades amb el global dels accidents, es fa un esforç per identificar-ne sobre els sinistres que presenten un impacte més elevat social, econòmica i mediambientalment parlant.

A través de la primera d'aquestes proves s'identifica que només un dels algorismes plantejats inicialment retorna uns resultats útils i representatius i es detalla la problemàtica en relació als altres dos algorismes. Dels resultats en deriven tres situacions que conjuntament representen més del 65% dels accidents i que es caracteritzen segons les variables de tipus de vehicle implicat en l'accident, sexe i edat del conductor, torn del dia, nombre de lesionats lleus i característiques climàtiques com la temperatura mitjana, la humitat relativa i la velocitat del vent. La primera d'aquestes situacions identificades agrupa gairebé el 47.7% dels accidents ocorreguts a la ciutat de Barcelona entre els anys 2018 i 2019. Respecte del subconjunt d'accidents greus, mitjançant altres variables característiques, també s'arriba a identificar situacions representatives, amb una d'aquestes identificant un conjunt de variables comuns en el 40.5% dels accidents més greus. Es presenta una descripció de la resta de situacions identificades, juntament amb agrupacions generades en iteracions parcials realitzades que són interessants per al lector.

En relació als objectius establerts, s'ha complert els objectius en relació al plantejament del

cas i la seva resolució, disseny i execució de proves, alhora que tot el procés d'adequació de les dades i generació de resultats significatius i amb un volum d'accidents i característiques destacables. En canvi, no s'aconsegueix identificar situacions on es donen accidents greus i mortals, de forma distintiva. El model, les variables de que es disposa i el procés de prova definit no generen agrupacions que separin els accidents greus i els mortals entre ells.

La qualitat de les situacions identificades i del procés de clustering és satisfactòria, però cal tenir en compte les limitacions que s'ha identificat en desenvolupar el present estudi. En primer lloc, es tracta d'un mètode d'anàlisi d'accidents de trànsit del que hi ha molt poca literatura de la que se'n pugui treure orientacions i millors pràctiques a l'hora de plantejar el procés d'operació. No existeix cap estudi anterior on s'apliqui algorismes de clustering en accidents urbans de la tipologia dels d'aquest estudi.

Sobre la informació de que es disposa, no es tenen característiques de detalls que sembla que podrien ajudar a identificar millor les situacions, com les relacionades amb el vehicle, amb la situació de l'accident i les condicions del moment del sinistre o sobre els conductors implicats en aquests i el seu estat físic en el moment de l'accident. També és important comentar que l'autor no havia rebut formació en relació a aquestes tècniques ni al llenguatge de programació Python prèviament a la realització del treball.

Possibles tasques futures

No és fàcil adreçar la falta de variables altament descriptives en relació als actors implicats en els accidents, ja que s'ha emprat les dades recollides in-situ per part de la guàrdia urbana, juntament amb dades de trànsit i meteorològiques. Per tal de superar aquestes limitacions, es proposa fer una recerca de noves variables a utilitzar o promoure la recollida de variables complementàries als òrgans encarregats. També es podria realitzar una ampliació de les característiques dels vehicles implicats, identificant en el model d'aquests algunes noves variables relacionades amb els vehicles.

Per tal de millorar el procediment d'operació, es proposa ampliar les tècniques i algorismes de Machine Learning utilitzats amb aquelles opcions que es consideri explorar. També es planteja la possibilitat d'ampliar l'àmbit de recerca, recollint dades relacionades amb altres ciutats similars en característiques a Barcelona, per tal de tenir una major volumetria de mostres.

10. Agraïments

En primer lloc i principalment, m'agradaria donar les gràcies al Doctor Jordi Olivella Nadal, tutor del treball, per la confiança dipositada en mi per a la realització d'aquest estudi. A més a més, aprecio molt el seguiment que ha fet en tot moment i el suport prestat en totes les fases del treball, orientant sobre la direcció a prendre i fent aportacions de molt valor durant el desenvolupament d'aquest.

En segon lloc, voldria agrair als diferents companys i amics que m'han acompanyat en aquesta aventura a l'escola, compartint vivències i moments molt especials i que han estat un recolzament fonamental durant aquests anys.

Finalment, no pot faltar el reconeixement a la família, que ha donat suport en aquells moments més complicats, mostrant interès i il·lusió en allò que faig, oferint consells i estant al meu costat en tot moment.

11. Bibliografia

[1] WORLD HEALTH ORGANIZATION. *World report on road traffic injury prevention*. (09 de Febrer de 2004). Recollit el 14 de Març de 2022, de t.ly/2X7U.

[2] WORLD HEALTH ORGANISATION. *Road Traffic Safety*. (21 de Juny de 2021). Recollit el 14 de Març de 2022, de t.ly/YEAS.

[3] UNITED NATIONS. *Transforming our World: The 2030 Agenda for Sustainable Development.* (2015). Recollit de t.ly/zovu.

[4] ARTIME RÍOS, E., Suárez Sánchez, A., Sánchez Lasheras, F., & Seguí Crespo, M. Genetic algorithm based on support vector machines for computer vision syndrome classification in health personnel. *Neural Computing and Applications* (2018). *32*, 1239 - 1248.

[5] SANTOS, D., PAULO QUARESMA, J., & BEIRES NOGUEIRA, V. Machine Learning Approaches to Traffic Accident Analysis and hotspot prediction. *Computers* (2021).

[6] REVERTER LÓPEZ, E. *Predicting The Severity Of Road Traffic Accidents In The City Of Barcelona.* Barcelona: Escola Tècnica Superior d'Enginyeria Industrial de Barcelona. (2021).

[7] CHEN, C., ZHANG, G., QIAN, Z., A. TAREFDER, R., & TIAN, Z. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis and Prevention* (2016)*, 90*, 128 - 139.

[8] GANDHI, R. *Towards data science*. Consultat el 20 / desembre / 2021, a Support Vector Machine — Introduction to Machine Learning Algorithms (07 / juliol / 2018): t.ly/HkIT.

[9] CHANG, L.-Y., & WANG, H.-W. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention (2006), 38*, 1019-1027.

[10] KAPLAN, S., & PRATO, C. Cyclist–Motorist Crash Patterns in Denmark: A Latent Class Clustering Approach. *Traffic Injury Prevention* (2013), 725 - 733.

[11] DEPAIRE, B., WETS, G., & VANHOOF, K. Traffic Accident Segmentation by Means of Latent Class Clustering. *Accident Analysis & Prevention* (2008)*, 40*, 1257 - 1266.

[12] KIM, K., & YUKIO YAMASHITA, E.. Using a K-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. *Journal of Advanced Transportation* (2007), 69 - 89.

[13] SHWETA, J. Y., BATRA, K., & GOEL, A. K. A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. *Journal of Physics* (2021).

[14] L. LI, S. SHRESTHA, & G. HU. Analysis of road traffic fatal accidents using data mining techniques. *15th International Conference on Software Engineering Research, Management and Applications (SERA)* (2017) (págs. 363-370). IEEE.

[15] ALBALATE, D., & FAGEDA, X. On the relationship between congestion and road safety in cities. *Transport Policy, 105*. Recollit de On the relationship between congestion and road safety in cities (Maig / 2021).

[16] HARWOOD, D. W., BAUER, K. M., & POTTS, I. B. Development of Relationships between Safety and Congestion for Urban Freeways. *Transportation Research Record: Journal of the Transportation Research Board* (1 de Gener de 2013)*, 2398*.

[17] ABDEL-ATY, M., & LEE, J. A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accident Analysis & Prevention* (2016)*, 88*.

[18] STATE HIGHWAY ADMINISTRATION. *The relationship between congestion levels and accidents.* Research Report (2003), University of Maryland, Maryland.

[19] KEAY, K, & SIMMONDS, I. Road accidents and rainfall in a large Australian city. *Accid. Anal. Prev, 38*, 445 – 454 (2006). Recollit de t.ly/99ZA.

[20] BLACK, A.W., VILLARINI, G., & MOTE, T.L. Effects of Rainfall on Vehicle Crashes in Six U.S. States. *Weather Clim. Soc* (2017)*, 9*. Recollit de t.ly/POPn.

[21] EISENBERG, D. The mixed effects of precipitation on traffic crashes. *Accid. Anal. Prev.* (2004)*, 36*, 637 - 647. Recollit de t.ly/PZlG.

[22] FRIDSTRØM, L., IFVER, J., INGEBRIGTSEN, S., KULMALA, R., & THOMSEN, L.K. Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accidents. *Accid. Anal. Prev.* (1995)*, 27*, 1 - 20. Recollit de t.ly/ffkK.

[23] GAO, J., CHEN, X., WOODWARD, A., LIU, X., WU, H., LU, Y., LIU, Q. The association between meteorological factors and road traffic injuries: a case analysis from Shantou city, China. *Scientific Reports (Sci Rep)* (2016)*, 6*. Recollit de https://rdcu.be/cHiL6.

[24] BISMART. (2021). Consultat el 19 / desembre / 2021, a The differences between supervised and unsupervised Machine Learning: t.ly/cDoI.

[25] BROWNLEE, J. *Machine Learning Mastery*. Consultat el 19 / desembre / 2021, a Difference Between Classification and Regression in Machine Learning (11 / desembre / 2017): t.ly/Hk69.

[26] MISHRA, S. *Towards Data Science*. Consultat el 19 / desembre / 2021, a Unsupervised Learning and Data Clustering (2021): t.ly/bid4.

[27] BROWNLEE, J. *Machine Learning Mastery*. Consultat el 19 / desembre / 2021, a

Supervised and Unsupervised Machine Learning Algorithms (20 / agost / 2020): t.ly/28h0.

[28] ZHEXUE HUANG, J. Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* (1998)*, 2*, 283–304.

[29] AJUNTAMENT DE BARCELONA. *Open Data BCN*. (Ajuntament de Barcelona) Consultat el 20 / Novembre / 2021, a t.ly/AZuT.

[30] COHEN, J. *Statistical Power Analysis for the Behavioral Sciences* (2ª edició ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers (1988).

[31] MAHMOOD, S. (12 / Juliol / 2021). *Factor Analysis of Mixed Data*. Recollit de Towards Data Science: t.ly/mdEB.

[32] A. GUPTA (09 de Febrer de 2021). *Elbow Method for optimal value of k in KMeans*. Obtenido de GeeksforGeeks: t.ly/fga1.

[33] BHARDWAJ, A. (26 / Maig / 2020). *Silhouette Coefficient*. Recollit de Towards Data Science: t.ly/OTUJ.

[34] VIDAL CLARAMUNT, O. *TFM_Code*. Recollit de GitHub: t.ly/f0xM.

[35] DADES METEOROLÒGIQUES DE LA XEMA. (Dades Obertes Catalunya) Consultat el 22 / Febrer / 2022, a t.ly/W2i8.

[36] TRÁFICO, D. G. *Otros factores de riesgo: La fatiga* (2022).

[37] UNIVERSITAT POLITÈCNICA DE CATALUNYA. *Crèdits per activitats d'extensió universitària*. Recollit de Servei de Gestió Acadèmica: t.ly/fyGp.