**Calanuya road safety analyzes and Report Malihe Alavi**

**30 jan 2024** [**m.alavi.649@gmail.com**](mailto:m.alavi.649@gmail.com) **+989370336116**

**Introduction:**

Road safety is a critical issue that affects millions of people worldwide. The World Health Organization (WHO) reports that road traffic accidents are the leading cause of death among young people aged 5-29 years 1. Machine learning (ML) has emerged as a promising technology to address road safety problems. ML algorithms can analyze large amounts of data and identify patterns that can help predict and prevent accidents. However, the use of ML in safety-critical systems poses significant engineering challenges.

Too many studies have worked on road safety problems for dataset of different countries. In most of these researches, authors have tried to train a classification model to predict if an accident results in death or sever injuries or not [7][8][9][10]. These studies have used common classification models like XGboost, logistic regression, KNN, random forest, SVM and neural network for classification task. On the other hand, several studies have worked on road-safety problem and tried to learn the pattern and predict future events by time series models. Most of these studies have used ARIMA model for this type of challenges.[1][2][3]

In this study, first we investigate the general trends and patterns that effect on the accidents, then we develop 3 ML models for predicting the accident based on environmental factors like weather, light, road type, speed limit and … and also based on the time series pattern between days and accidents types. Despite of the previous study that developed a classification model, the target value of Catalunya dataset has different range of integer values which leads us to train regression models with random forest regressor and SVM regressor, also we have developed a time series model for predicting the upcoming year.

In order to train regression models, we had to first preprocess the data because it had missing values that should be replaces with mean or median or should be removed from dataset, also there were too many categorical features which should be encoded to numeric features.

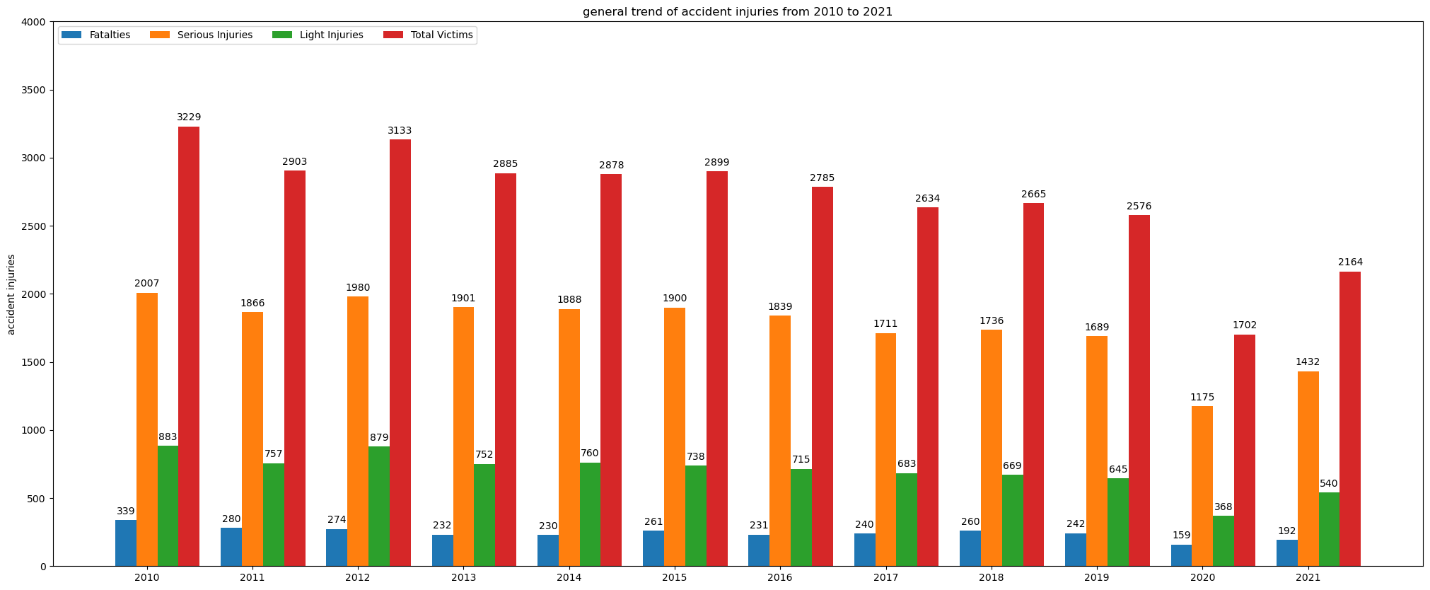
**Key finding:**

* Random forest regressor outperforms SVR in terms of time complexity, accuracy and simplicity for Calanuya dataset
* The most important factors on accidents were: hour of the day, wind condition, weather condition, light condition and the type of vehicle that was involved in the accident.
* Based on yearly trend, one can concludes that the amounts of total victims and light injuries in accidents has slightly decreased over 10 years
* Most of the accidents have occurred in Barcelona which is about 4 times more than the other province.
* Most of the accidents occurred from 10 a.m. to 8 p.m. and during this period, the evenings about (4p.m. to 8p.m.) are the pic of crashes
* Type of road and speed limit info also have great impact on the accidents
* the accident rates are more in months May, June, July and October respect to serious Injuries.
* Fatalities rates are relatively the same in all months
* light injuries also increases in July.

**Data Visualizations and Analyzes:**

1. **General Trends**

**What are the overall trends in traffic accidents,** **fatalities, and serious injuries in Catalonia from 2010-2021?**

****

As can be seen from the chart, the overall total victims have been slightly decreased during 2010 to 2021, the number of total victims in 2010 was 3229 and it has reduced to 2164 in 2021. However, there is no significant trend in fatalities rate but a slight reduction also can be seen in serious injuries and light injuries rate. As they reduced from 2007 and 883 in 2010 to 1432 and 540 in 2021 respectively.

**2. Accident Characteristics**

**What common characteristics (time of day, type of road, etc.) are observed in the most severe accidents?**

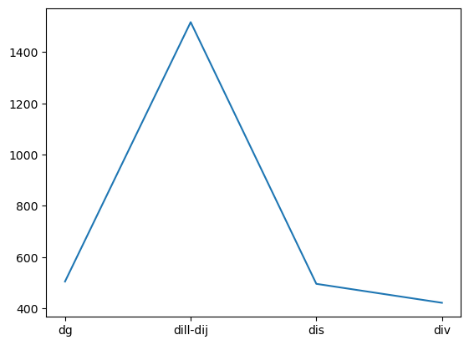
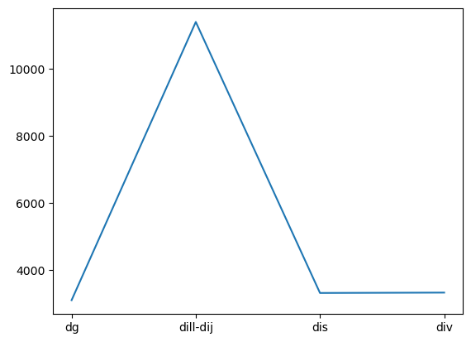
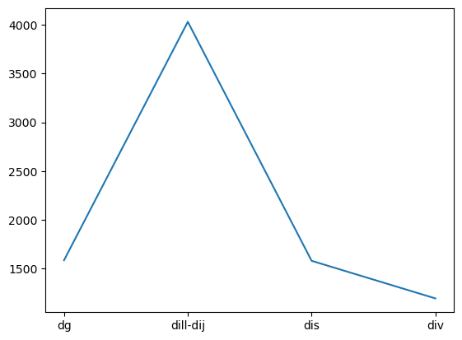
Number of fatalities, serious Injuries, light Injuries and total victims respect to **Road** (top 20)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Road | Tot. vic. | | SE | 14351 | | AP-7 | 951 | | N-II | 947 | | N-340 | 662 | | A-2 | 650 | | C-31 | 642 | | N-260 | 546 | | CR | 530 | | C-14 | 527 | | C-58 | 444 | | C-55 | 440 | | C-12 | 429 | | C-17 | 366 | | C-32 | 304 | | C-16 | 298 | | N-240 | 277 | | C-25 | 271 | | C-35 | 246 | | B-10 | 219 | | C-13 | 209 | | |  |  | | --- | --- | | road | fatalities | | SE | 814 | | AP-7 | 140 | | N-340 | 136 | | N-II | 124 | | A-2 | 80 | | C-31 | 78 | | C-14 | 76 | | CR | 74 | | C-12 | 62 | | C-55 | 58 | | N-260 | 57 | | C-25 | 51 | | C-17 | 49 | | C-16 | 39 | | C-58 | 39 | | N-240 | 39 | | C-32 | 37 | | N-420 | 32 | | C-63 | 30 | | C-15 | 30 | | |  |  | | --- | --- | | road | Ser. Inj | | SE | 10786 | | N-II | 501 | | AP-7 | 440 | | C-31 | 356 | | N-340 | 335 | | CR | 326 | | N-260 | 324 | | A-2 | 319 | | C-14 | 254 | | C-58 | 240 | | C-55 | 231 | | C-17 | 199 | | C-12 | 198 | | C-32 | 184 | | C-16 | 175 | | C-35 | 129 | | C-25 | 128 | | B-10 | 128 | | N-240 | 122 | | C-13 | 121 | | |  |  | | --- | --- | | Road | Li. Inj | | SE | 2751 | | AP-7 | 371 | | N-II | 322 | | A-2 | 251 | | C-31 | 208 | | C-14 | 197 | | N-340 | 191 | | C-12 | 169 | | C-58 | 165 | | N-260 | 165 | | C-55 | 151 | | CR | 130 | | C-17 | 118 | | N-240 | 116 | | C-25 | 92 | | C-35 | 89 | | C-16 | 84 | | C-32 | 83 | | C-65 | 70 | | B-10 | 69 | |

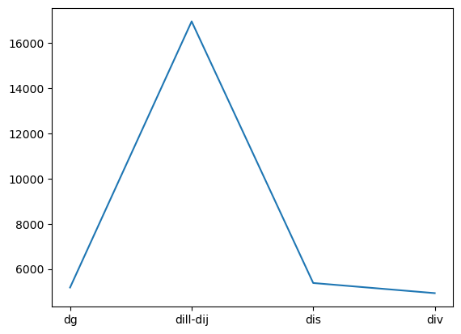
Number of fatalities, serious Injuries, light Injuries and total victims respect to **Day Type**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Day Type | Fatality | | dg | 505 | | dill-dij | 1517 | | dis | 496 | | div | 422 | | |  |  | | --- | --- | | Day Type | Ser. Injuries | | Dg | 3093 | | dill-dij | 11404 | | dis | 3308 | | div | 3319 | | |  |  | | --- | --- | | Day Type | Total Vic. | | Dg | 5183 | | dill-dij | 16953 | | dis | 5383 | | div | 4934 | | |  |  | | --- | --- | | Day Type | Lig. Injuries | | dg | 1585 | | dill-dij | 4032 | | dis | 1579 | | div | 1193 | |

fatalities serious injuries Light injuries

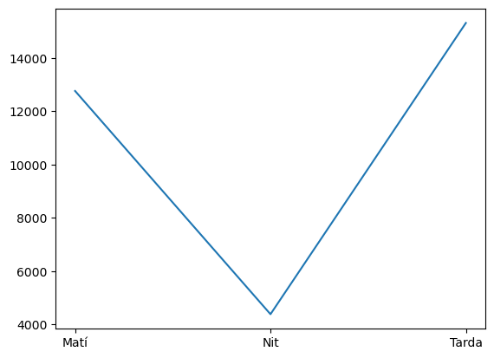
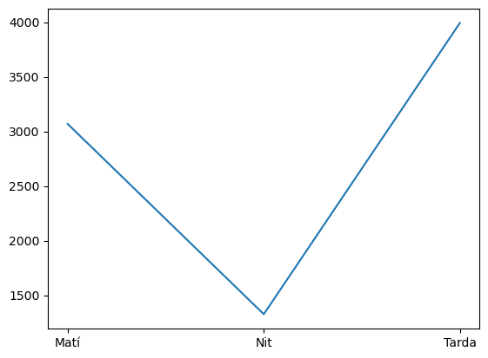
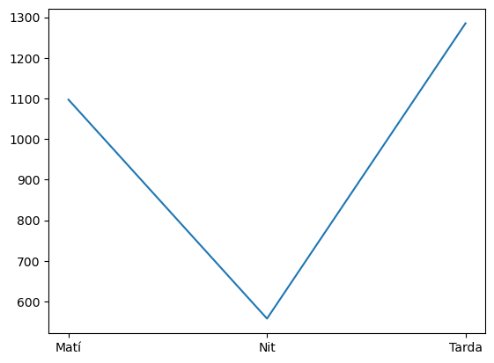


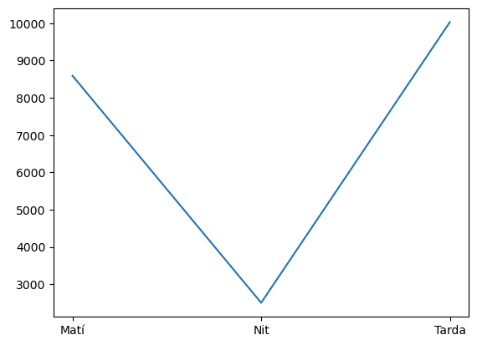
total victims



Number of fatalities, serious Injuries, light Injuries and total victims respect to **Time of Day Grouping**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | TOD Grouping | Lig. Inj. | | Matí | 3069 | | Nit | 1327 | | Tarda | 3993 | | |  |  | | --- | --- | | TOD Grouping | Tot. vic. | | Matí | 12758 | | Nit | 4387 | | Tarda | 15308 | | |  |  | | --- | --- | | TOD Grouping | fatality | | Matí | 1097 | | Nit | 558 | | Tarda | 1285 | | |  |  | | --- | --- | | TOD Grouping | Ser. Inj. | | Matí | 8592 | | Nit | 2502 | | Tarda | 10030 | |

 Light injuries total victims fatalities

Serious Injuries

**3. Geographical Insights**

**Which municipalities or counties in Catalonia have the highest incidence of traffic accidents? How does this correlate with population density or road network characteristics?**

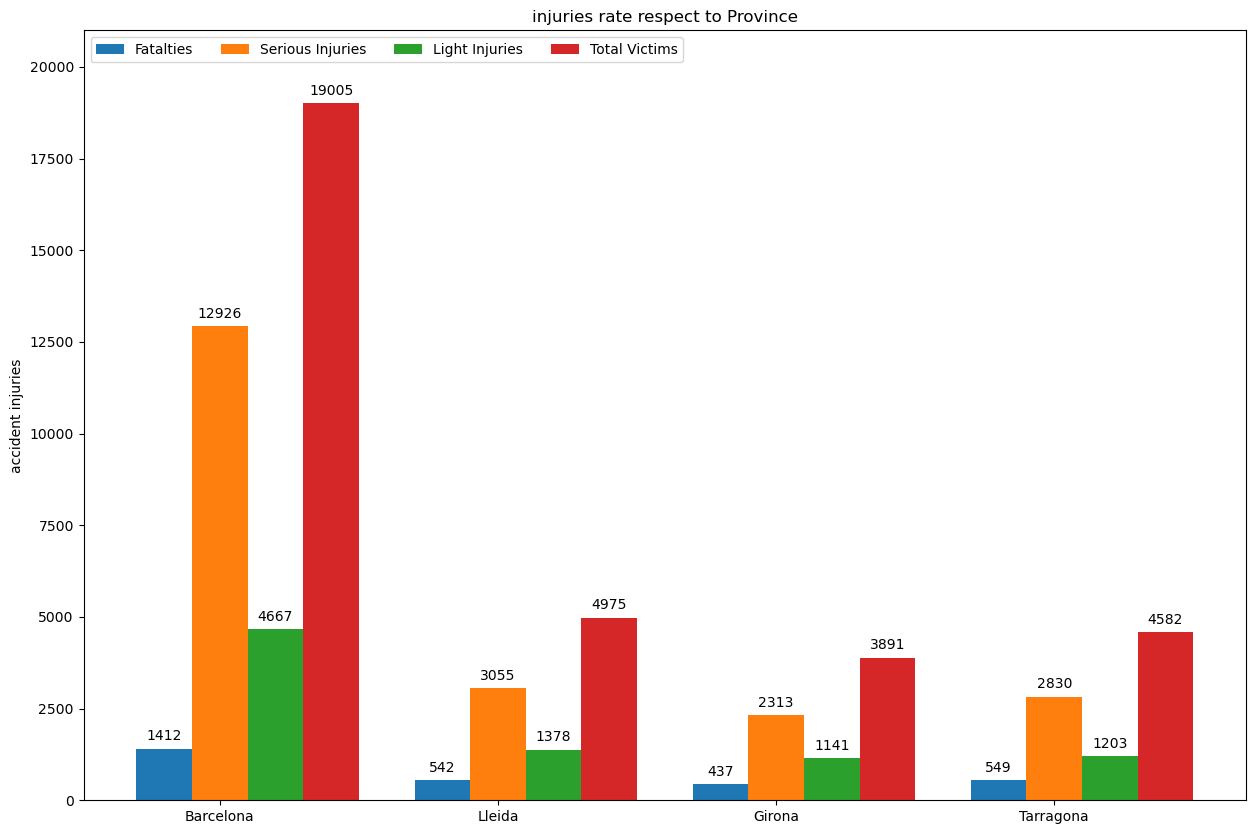
Number of fatalities, serious Injuries, light Injuries and total victims respect to **Municipality Name** (top 20)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Municipality Name** | **Tot. vic.** | | BARCELONA | 4029 | | TERRASSA | 746 | | LLEIDA | 732 | | SABADELL | 708 | | BADALONA | 544 | | TARRAGONA | 527 | | HOSPITALET DE LLOBREGAT | 504 | | GIRONA | 444 | | REUS | 395 | | MATARO | 390 | | MANRESA | 372 | | SANT BOI DE LLOBREGAT | 318 | | VILANOVA I LA GELTRU | 293 | | VENDRELL, EL | 292 | | SANT CUGAT DEL VALLES | 263 | | VIC | 243 | | CERDANYOLA DEL VALLES | 240 | | GRANOLLERS | 233 | | LLORET DE MAR | 233 | | PRAT DE LLOBREGAT, E | 212 | | |  |  | | --- | --- | | **Municipality Name** | fatality | | BARCELONA | 302 | | LLEIDA | 52 | | BADALONA | 46 | | TARRAGONA | 44 | | SABADELL | 30 | | TERRASSA | 29 | | REUS | 28 | | GIRONA | 28 | | HOSPITALET DE LLOBREGAT | 26 | | MANRESA | 26 | | VENDRELL, EL | 25 | | ROCA DEL VALLES, LA | 25 | | AMPOSTA | 24 | | VILANOVA I LA GELTRU | 21 | | MATARO | 21 | | SANT CUGAT DEL VALLES | 20 | | MONT-ROIG DEL CAMP | 19 | | GRANOLLERS | 19 | | VILAFRANCA DEL PENEDES | 18 | | BALAGUER | 17 | | |  |  | | --- | --- | | **Municipality Name** | Ser. inj. | | BARCELONA | 2656 | | TERRASSA | 574 | | SABADELL | 530 | | LLEIDA | 507 | | HOSPITALET DE LLOBREGAT | 384 | | TARRAGONA | 378 | | BADALONA | 360 | | GIRONA | 320 | | MATARO | 303 | | REUS | 293 | | MANRESA | 264 | | SANT BOI DE LLOBREGAT | 243 | | VILANOVA I LA GELTRU | 213 | | SANT CUGAT DEL VALLES | 187 | | VIC | 187 | | LLORET DE MAR | 184 | | CERDANYOLA DEL VALLES | 179 | | GRANOLLERS | 176 | | VENDRELL, EL | 172 | | CORNELLA DE LLOBREGAT | 162 | |
| |  |  | | --- | --- | | **Municipality Name** | Li. Inj. | | BARCELONA | 1701 | | LLEIDA | 173 | | SABADELL | 148 | | TERRASSA | 143 | | BADALONA | 138 | | TARRAGONA | 105 | | GIRONA | 96 | | VENDRELL, EL | 95 | | HOSPITALET DE LLOBREGAT | 94 | | MANRESA | 82 | | REUS | 74 | | MATARO | 66 | | SANT QUIRZE DEL VALLES | 63 | | ROCA DEL VALLES, LA | 62 | | AMPOSTA | 62 | | SANT BOI DE LLOBREGAT | 59 | | VILANOVA I LA GELTRU | 59 | | SANT CUGAT DEL VALLES | 56 | | CERDANYOLA DEL VALLES | 51 | | PRAT DE LLOBREGAT, E | 50 | |  |  |

Number of fatalities, serious Injuries, light Injuries and total victims respect to **County Name** (top 20)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | County Name | Total vic. | | Barcelones | 5371 | | Valles Occidental | 3279 | | Baix Llobregat | 2439 | | Valles Oriental | 1799 | | Maresme | 1395 | | Segria | 1300 | | Selva | 1214 | | Alt Emporda | 1208 | | Bages | 1159 | | Girones | 1052 | | |  |  | | --- | --- | | County Name | Ser. Inj. | | Barcelones | 3626 | | Valles Occidental | 2335 | | Baix Llobregat | 1737 | | Valles Oriental | 1155 | | Maresme | 1007 | | Segria | 814 | | Bages | 745 | | Selva | 737 | | Alt Emporda | 724 | | Tarragones | 693 | | |  |  | | --- | --- | | County Name | Light Inj. | | Barcelones | 1354 | | Valles Occidental | 757 | | Baix Llobregat | 545 | | Valles Oriental | 488 | | Segria | 363 | | Alt Emporda | 347 | | Selva | 344 | | Girones | 306 | | Bages | 303 | | Maresme | 296 | |
|  |  |  |
| |  |  | | --- | --- | | County Name | fatalities | | Barcelones | 391 | | Valles Occidental | 187 | | Baix Llobregat | 157 | | Valles Oriental | 156 | | Alt Emporda | 137 | | Selva | 133 | | Segria | 123 | | Bages | 111 | | Girones | 110 | | Baix Ebre | 109 | |  |  |

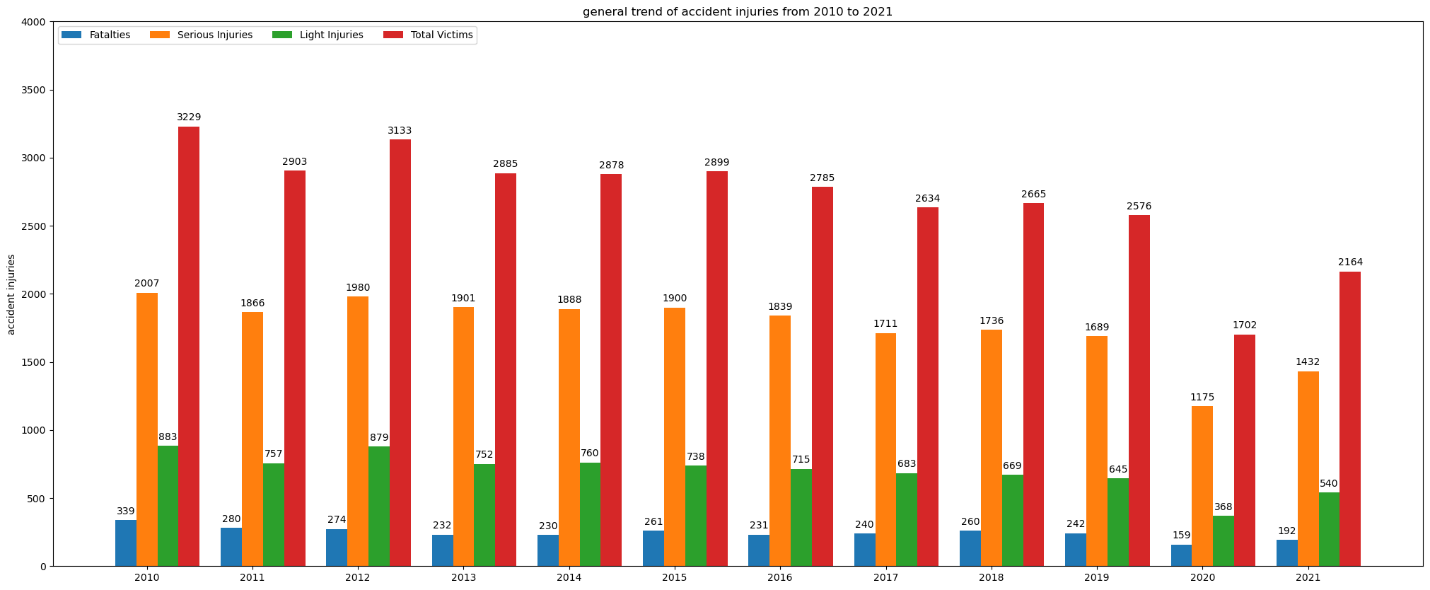
Number of fatalities, serious Injuries, light Injuries and total victims respect to **Province Name**



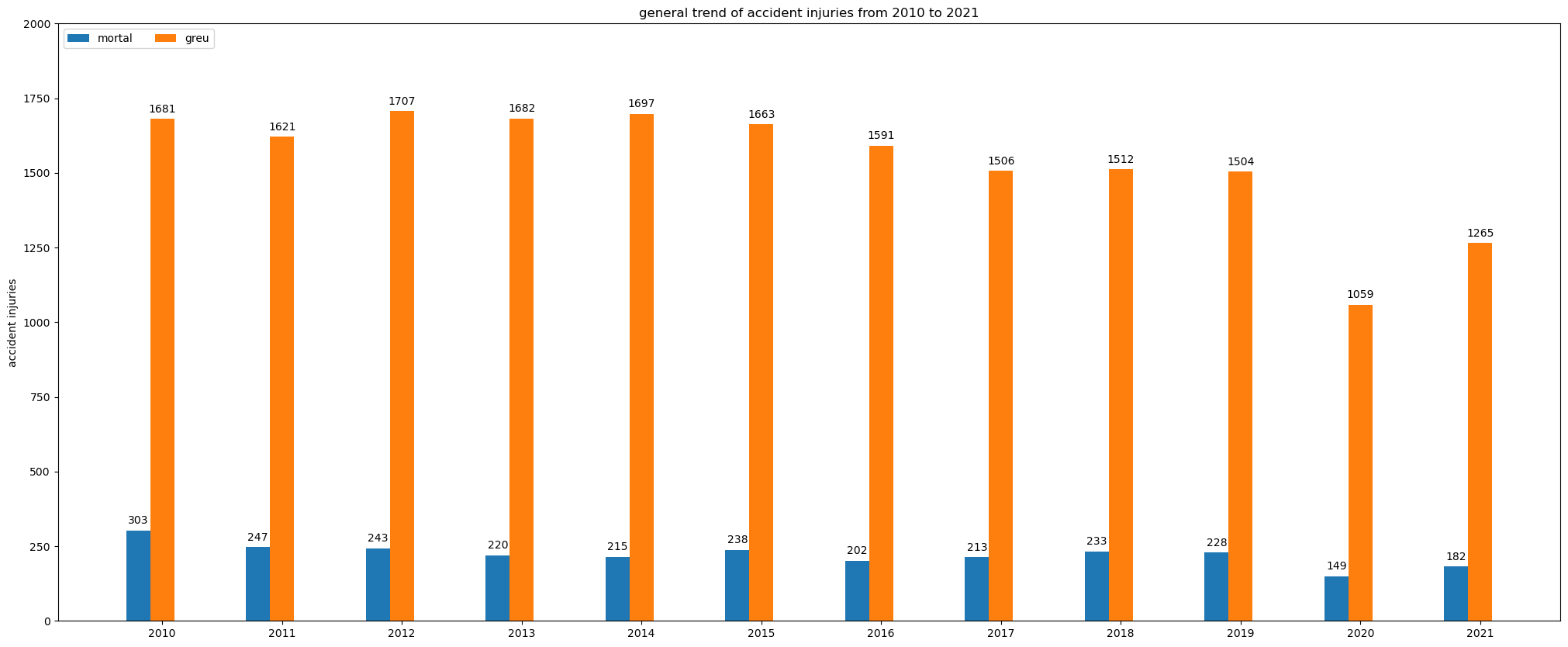
4. **Yearly Trends**

How have traffic accident patterns (frequency, severity) changed yearly from 2010 to 2021?

Number of fatalities, serious Injuries, light Injuries and total victims respect to **Year**

****

Number of mortal and greu (severity of accident ) respect to **Year**



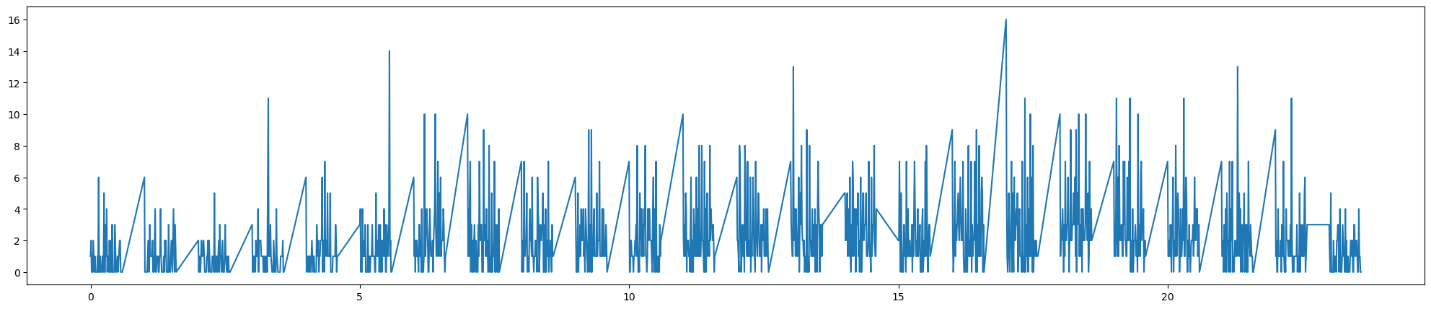
5. **Day and Time Patterns**

On what days of the week and times of day do most accidents occur? Are there notable differences between weekdays and weekends?

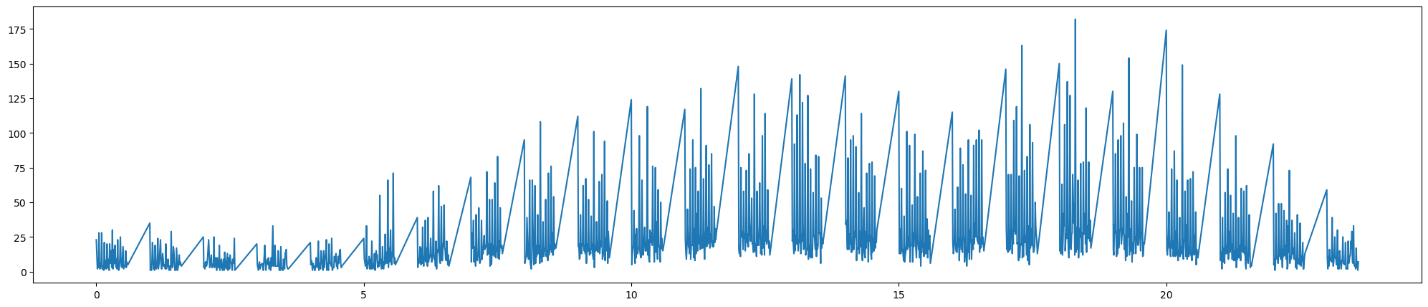
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | DOw | Serious Inj. | | CapDeSetmana | 8006 | | Feiners | 13118 | | |  |  | | --- | --- | | DOw | Light Inj. | | CapDeSetmana | 3742 | | Feiners | 4647 | | |  |  | | --- | --- | | DOw | Total vic. | | CapDeSetmana | 12942 | | Feiners | 19511 | |

|  |  |
| --- | --- |
| DOw | Fatalties |
| CapDeSetmana | 1194 |
| Feiners | 1746 |

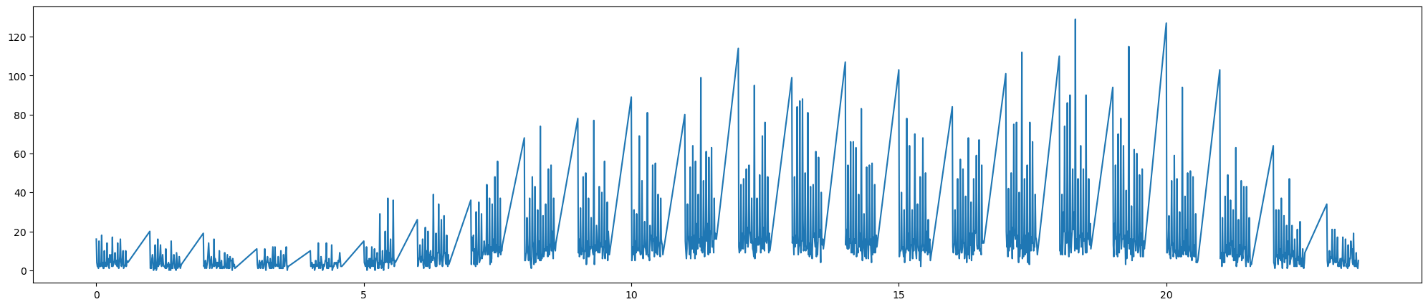
Number of fatalities respect to Hour of Day



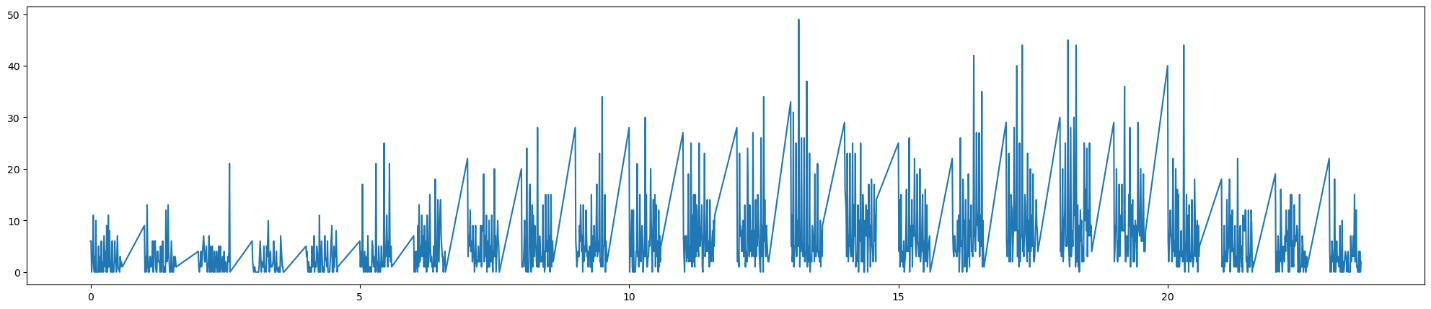
Number of total Victims respect to Hour of Day



Number of serious inj. respect to Hour of Day



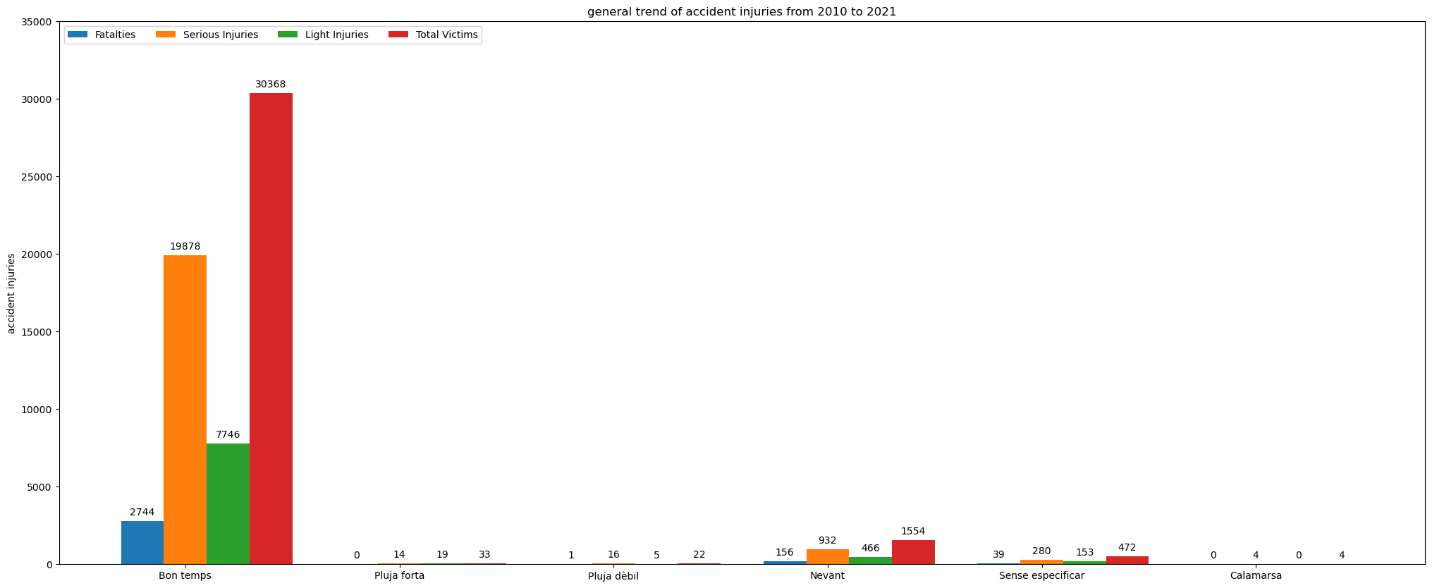
Number of light inj. respect to Hour of Day



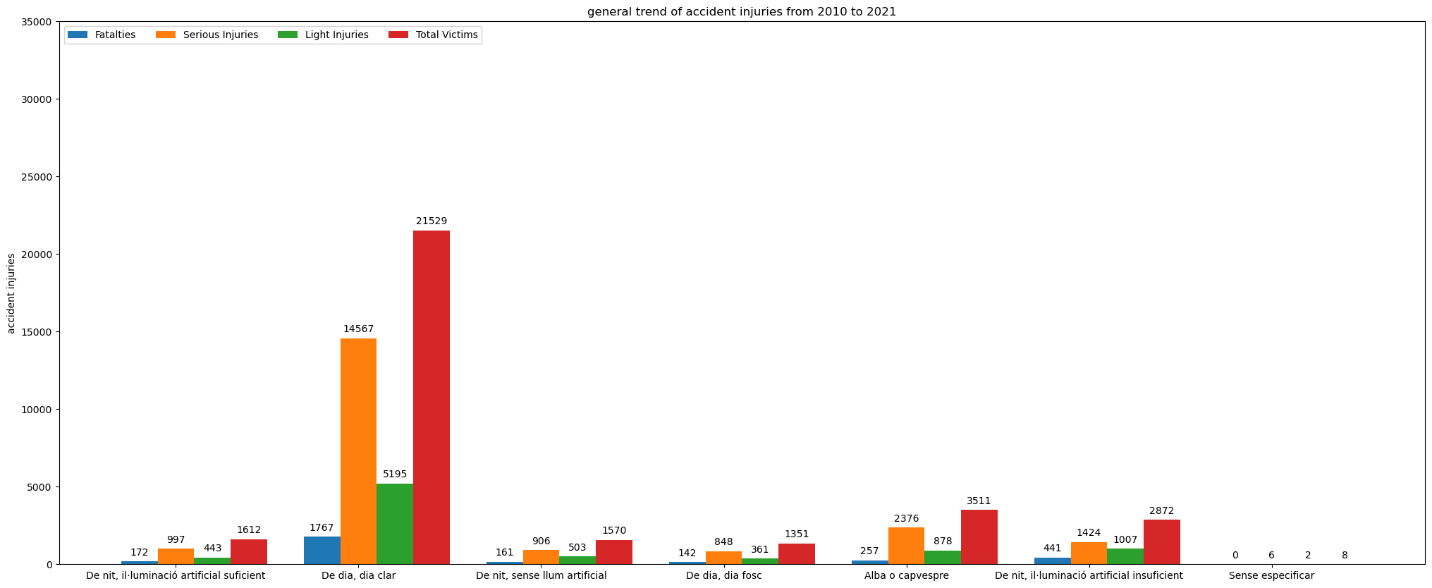
6. **Environmental Impact**

How do different weather conditions affect the likelihood of accidents? Is there a correlation between visibility, road conditions, and accident severity?

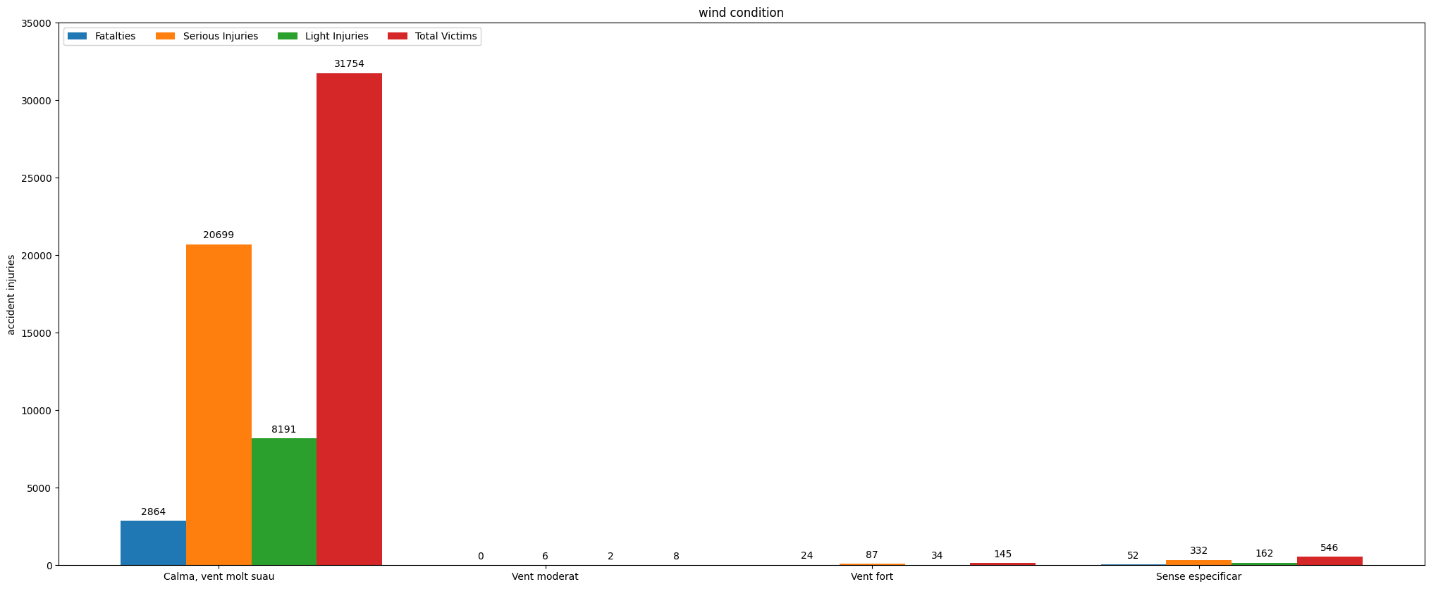
Number of fatalities, serious Injuries, light Injuries and total victims respect to **weather conditions**



Number of fatalities, serious Injuries, light Injuries and total victims respect to **Lighting condition**



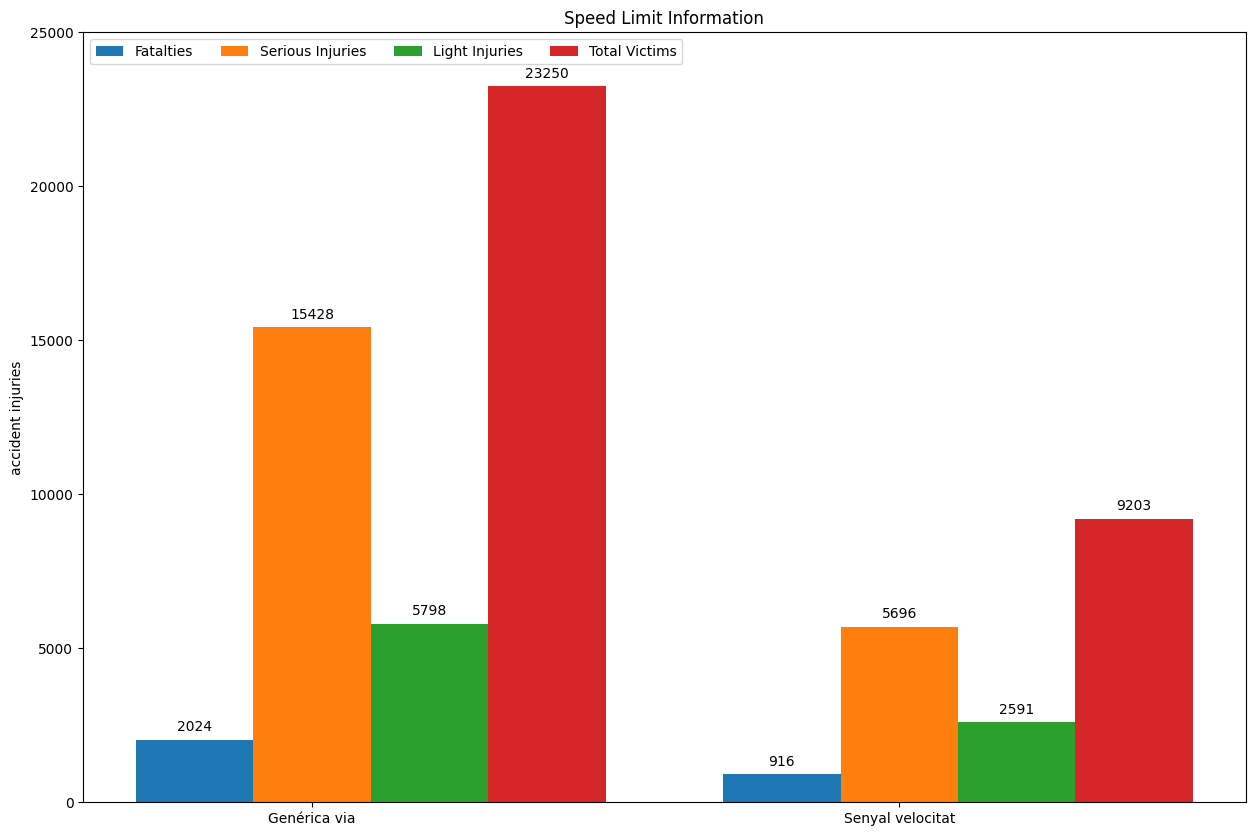
Number of fatalities, serious Injuries, light Injuries and total victims respect to **wind condition**



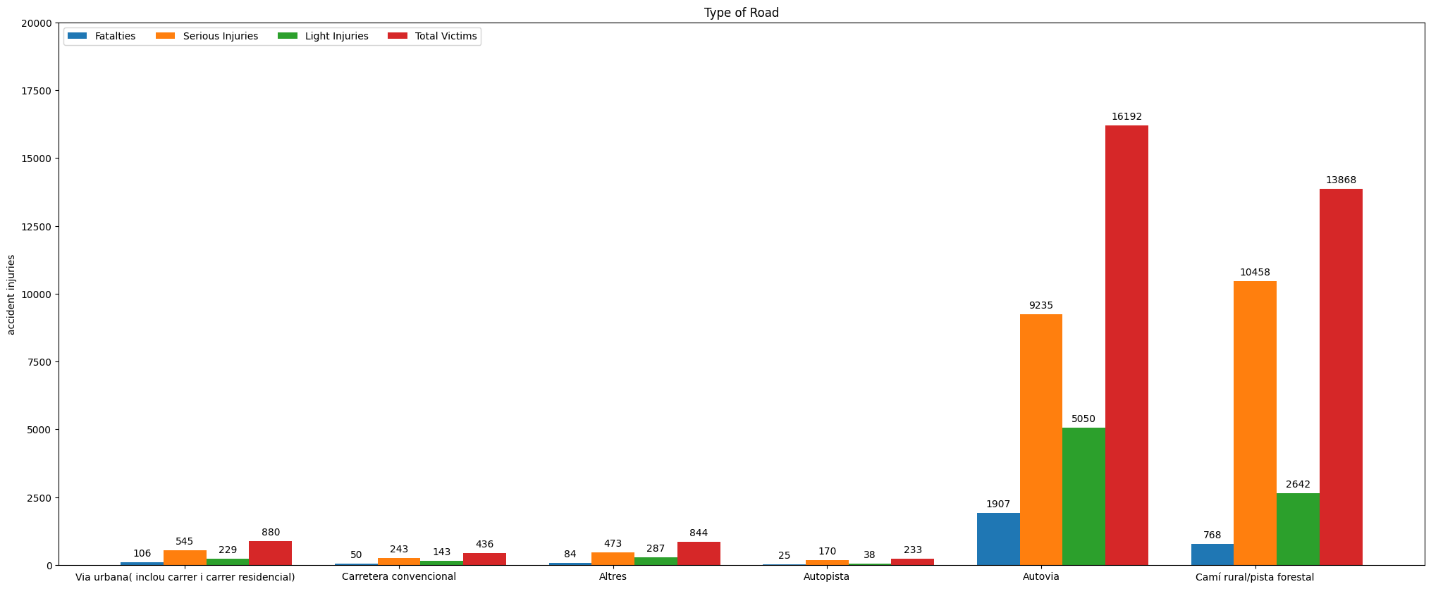
7. **Road and Traffic Features**

What impact do road features (such as speed limits and road types) and traffic density have on the occurrence of accidents?

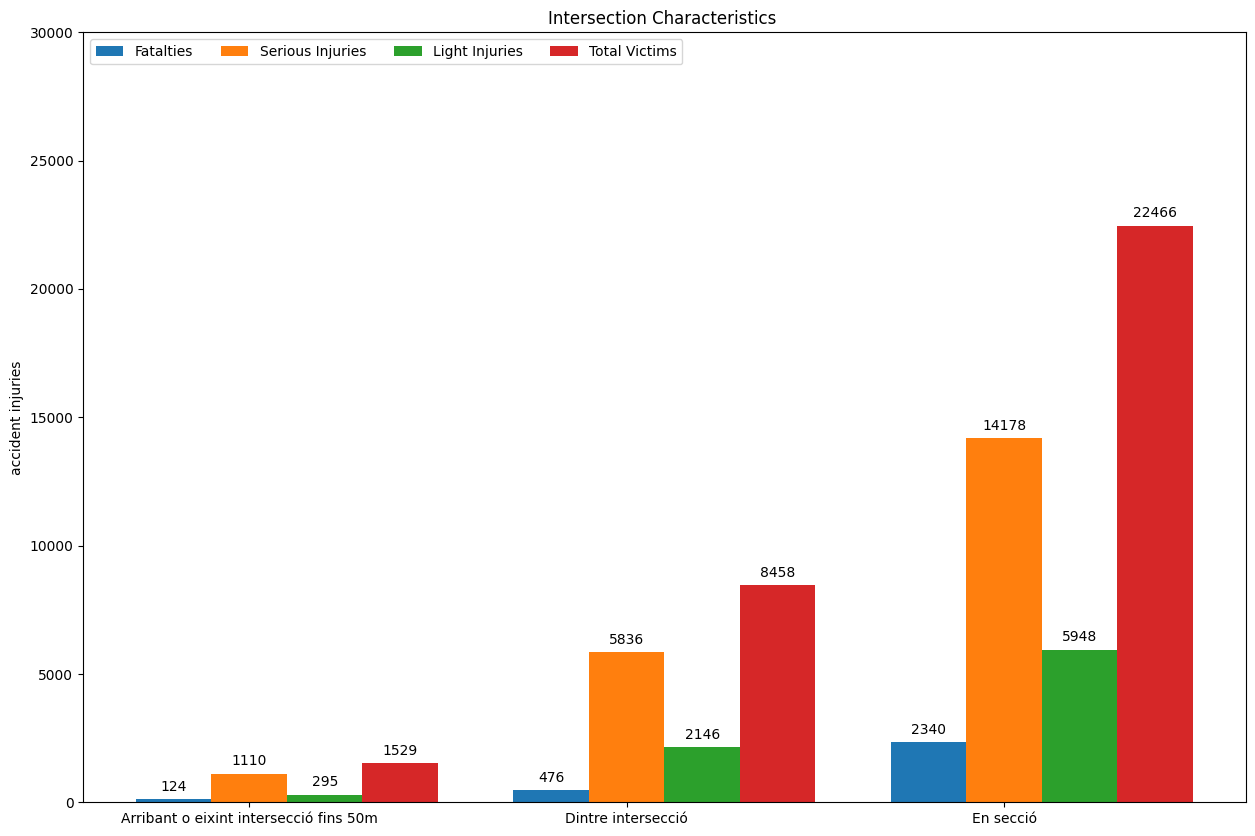
Number of fatalities, serious Injuries, light Injuries and total victims respect to **speed limit information**



Number of fatalities, serious Injuries, light Injuries and total victims respect to **type of road**



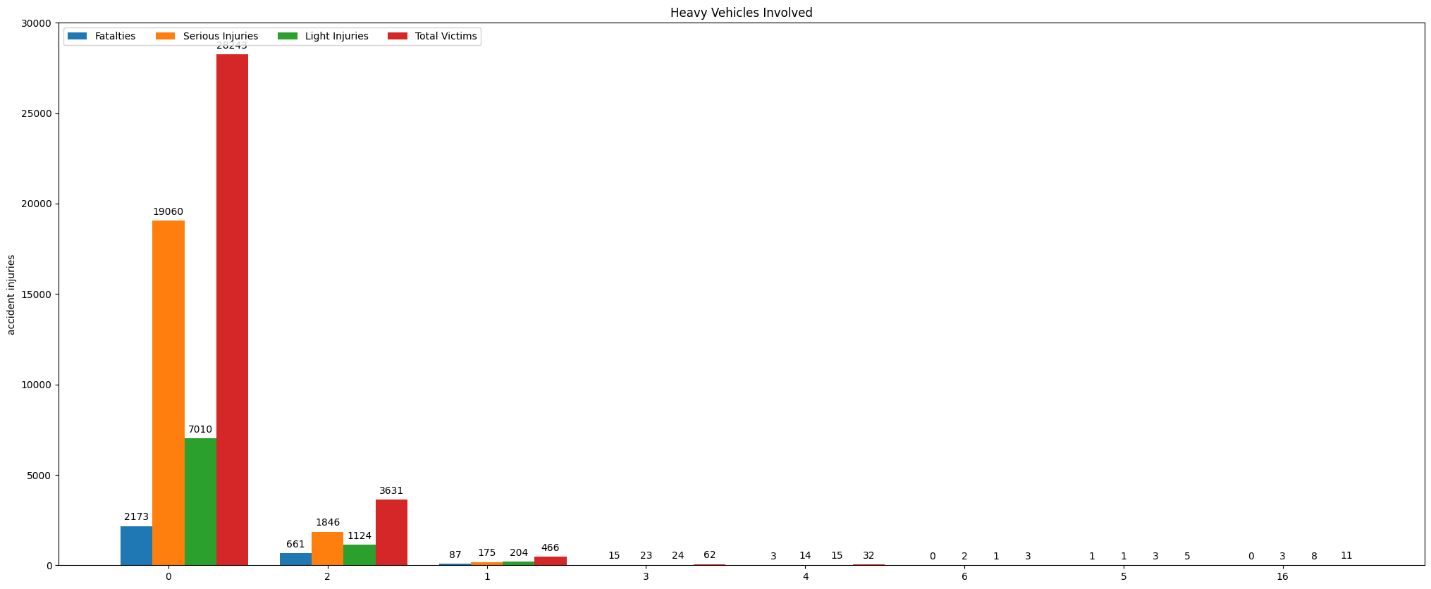
Number of fatalities, serious Injuries, light Injuries and total victims respect to **intersection characteristics**



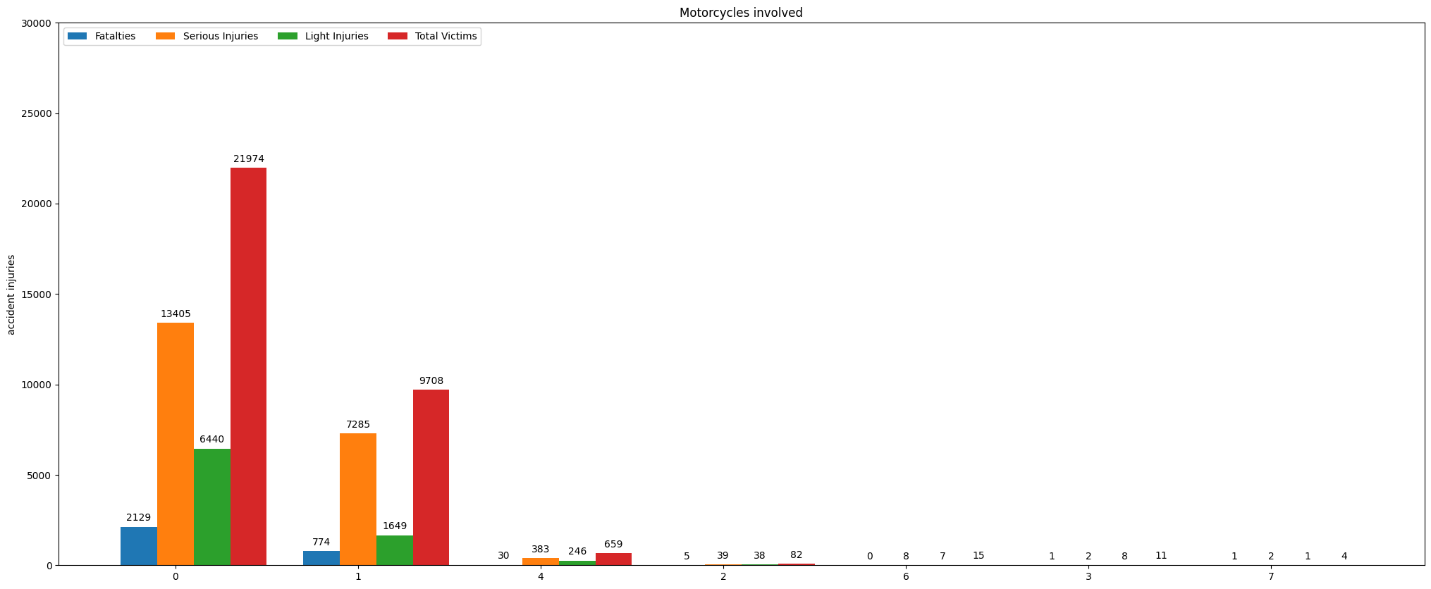
8. **Vehicle Types and Accident Severity**

Does the involvement of specific types of vehicles (like heavy trucks and motorcycles) correlate with more severe accidents?

Number of fatalities, serious Injuries, light Injuries and total victims respect to the num. of **heavy vehicle involved**

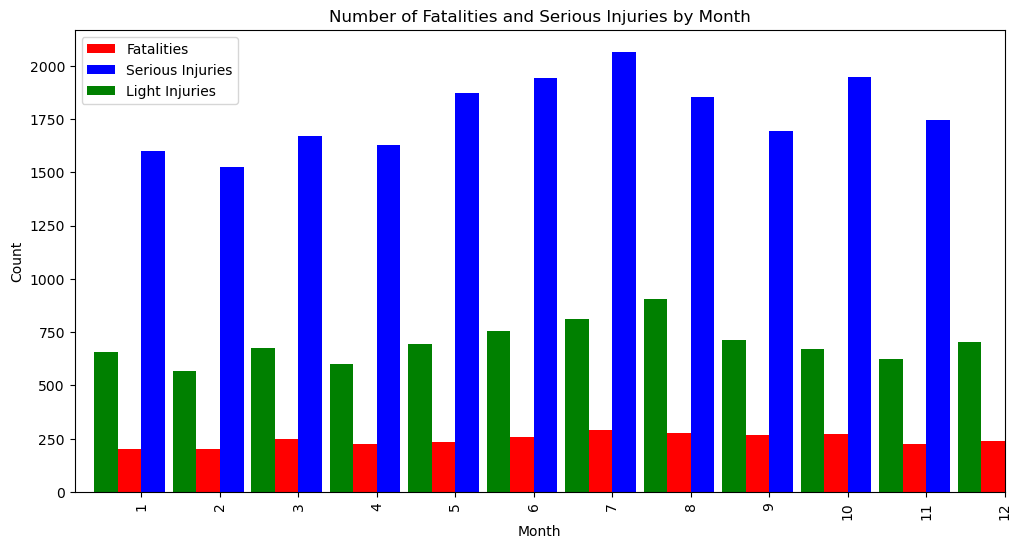


Number of fatalities, serious Injuries, light Injuries and total victims respect to the num. of **motorcycle involved**



9. **Temporal Clustering**

Are there specific periods (months, years) where accident patterns cluster significantly? What might be the causes for these clusters?



Generally, the accident rates are more in months May, June, July and October respect to serious Injuries.

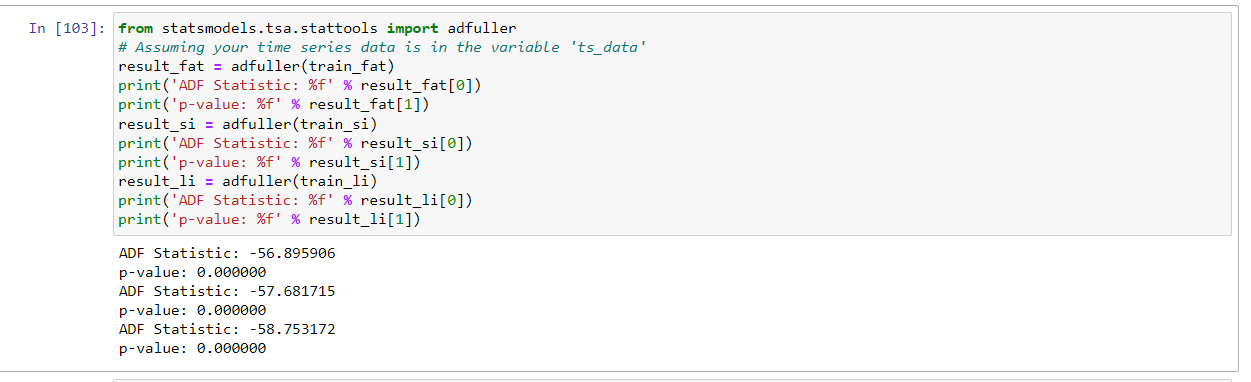
Fatalities rates are relatively the same in all months and light injuries also increases in July.

10. **Time-Series Forecasting**

Based on past trends, create a model to forecast the number of accidents, fatalities, or serious injuries for the upcoming year. Clearly describe the forecasting model you have developed. This should include the type of model, its structure, and any specific features or techniques it utilizes. Discuss the factors that influenced your decision, such as the model’s accuracy, efficiency, suitability to the data characteristics, or its ability to handle the complexities of the dataset.

For this purpose, after loading the data, we first convert the ***Date*** column format from dd/mm/yyyy to yyyy/mm/dd to be sorted based on days not years then we grouped ‘***Date’*** base on each target values (fatality, light injuries, serious injuries) and the output were 3 series with the length of 4308. Each series indicates in each day of the Date how many fatalities /light injuries/ serious injuries have happened.

Before applying any model to data we needed to be sure that our data is stationary which means it’s mean and variance is consistent during the time. The ***adfuller*** python library did this for us. It takes the time series data and analyzes it and returns p-value and ADF Statistic which indicates whether the ts\_data is stationary or not.



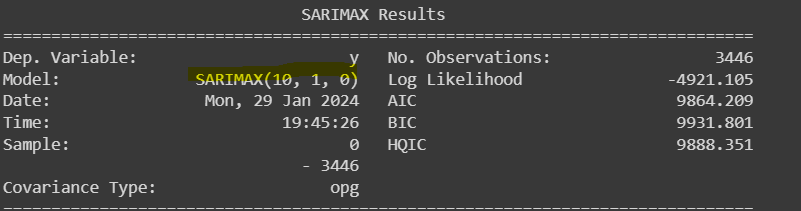
For three series the ADF Statistic indicate that the time series is stationary. The p-value is very less than the significance level of 0.05, which means that we can reject the null hypothesis that the series is non-stationary [6]. In other words, the time series does not exhibit a unit root, and it is stationary. This result is consistent with the ADF test, which is used to test the null hypothesis that a time series has a unit root and is non-stationary.

Based on studies [1],[2],[3], ARIMA time series model outperforms the other model in accidents and road safeties problems. Therefore, in this work we used ARIMA model to predict the upcoming year. In mentioned studies the ARIMA parameters (p,q,d) which are [4,5] :

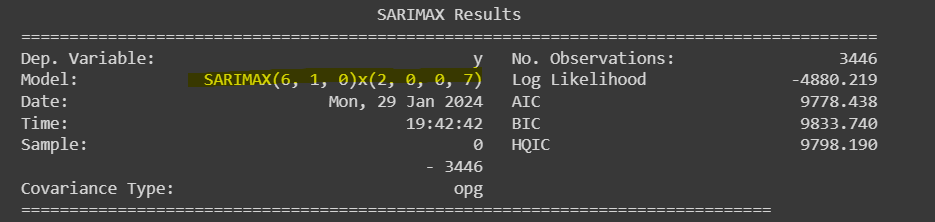
* **p** represents the autoregressive term, which incorporates the effect of past values into the model. It signifies the number of lag observations included in the model.
* **d** denotes the differencing term, which is the number of differences needed to make the time series data stationary. It reflects the degree of differencing.
* **q** stands for the moving average term, which accounts for the influence of past white noise or error terms. It indicates the size of the moving average window.

are tuned to (1,1,1) and (1,0,1). In addition to these parameters, we obtained a few other parameters using ***auto\_arima*** library in python. This library takes a min and max range for each of these parameters and explore all the possible value to gain the best accuracy. ***auto\_arima*** suggests these parameters for training ARIMA model:

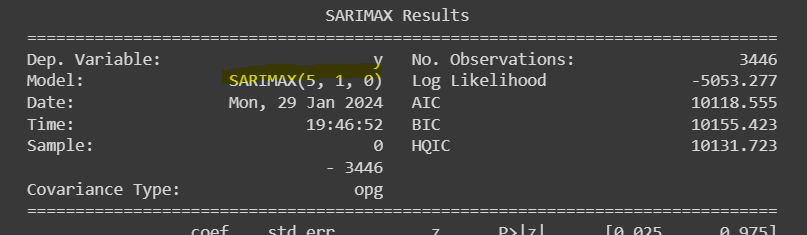
given parameter to auto\_arima: start\_p = 1, start\_q=1, max\_p = 10, max\_q = 10, seasonal=False



given parameter to auto\_arima: start\_p = 1, start\_q=1, max\_p = 10, max\_q = 10, seasonal=True, m=7



given parameter to auto\_arima: start\_p = 1, start\_q=1, max\_p = 5, max\_q = 5, seasonal=False



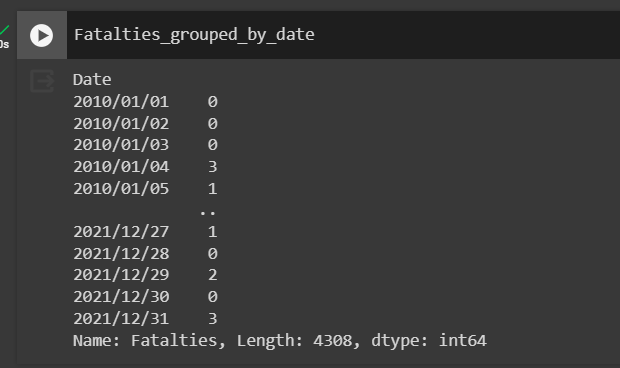
In order to examine the accuracy of each model on our time series data, we first split each series into train and test with 0.8 and 0.2 portion of primary data, respectively. We trained the model on train part and then test it on the test part to see the performance.

**Evaluation metrics**: we use three metrics to evaluate the model

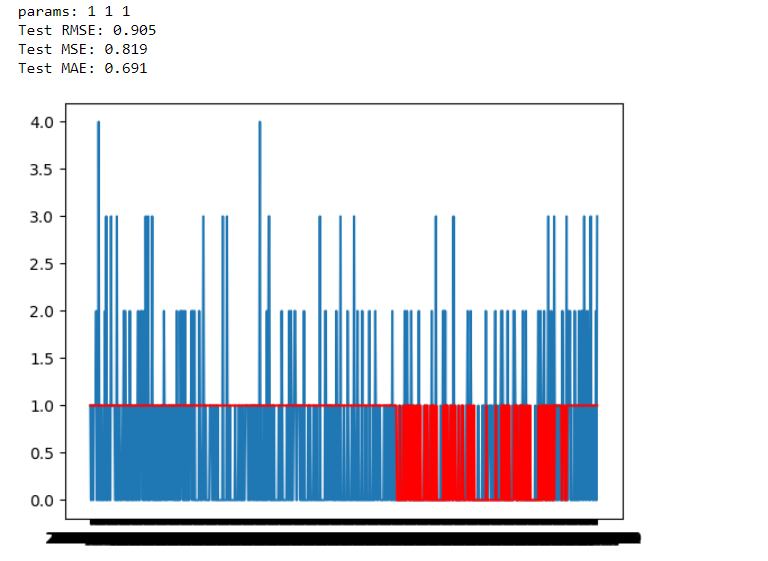
* mean square error (MSE)
* mean absolute error (MAE)
* root mean square error (RMSE)

**RESULTS**

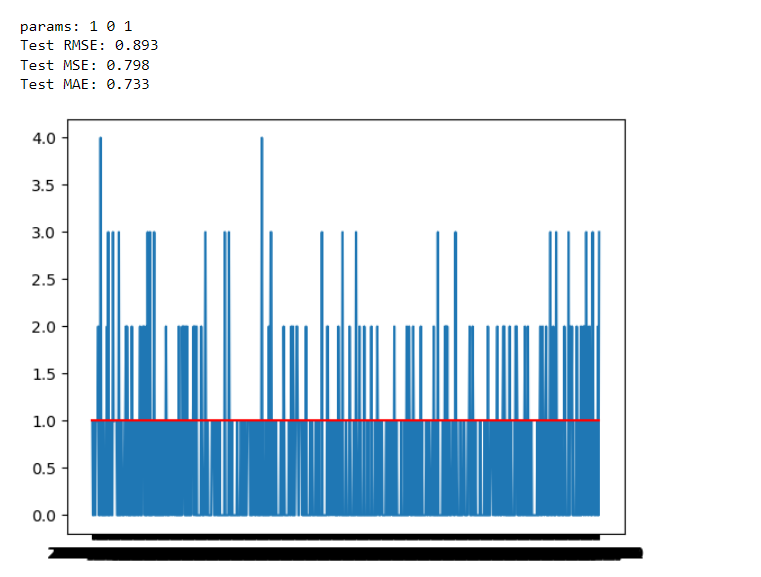
**Fatality time series:** as shown in the picture below, the index of this series indicates the Date which is sorted by day and the values are the number of fatalities in each day



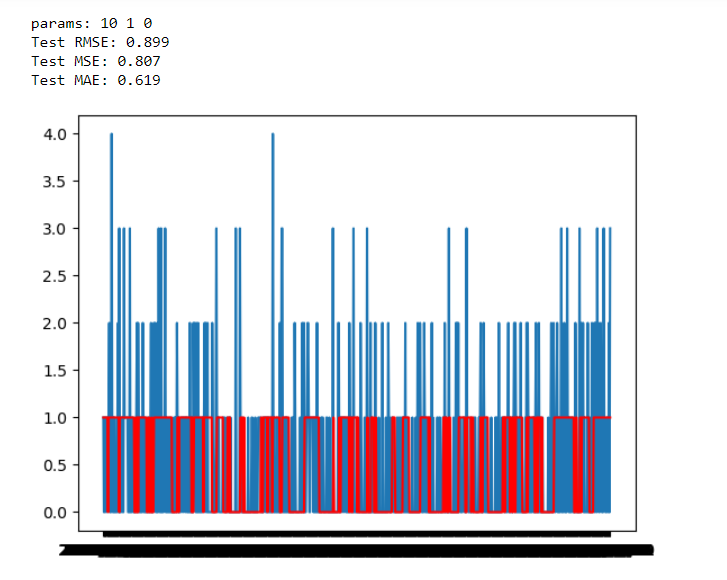
p=1, q=1, d=1:



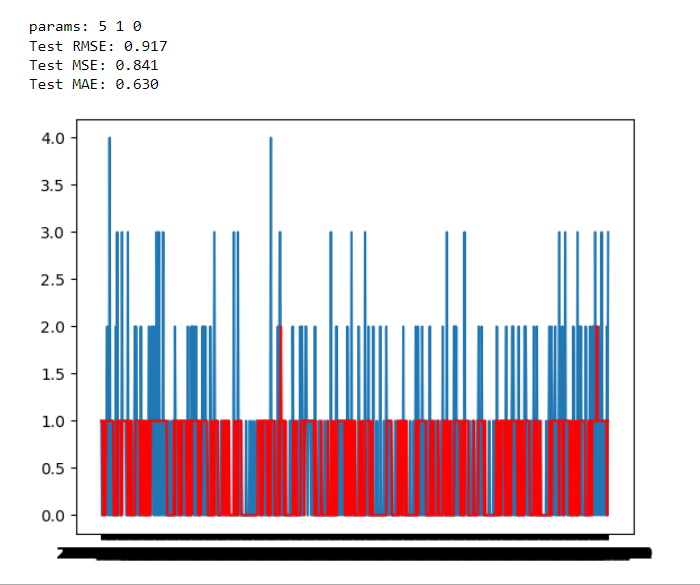
p=1 , q=0, d=1 :



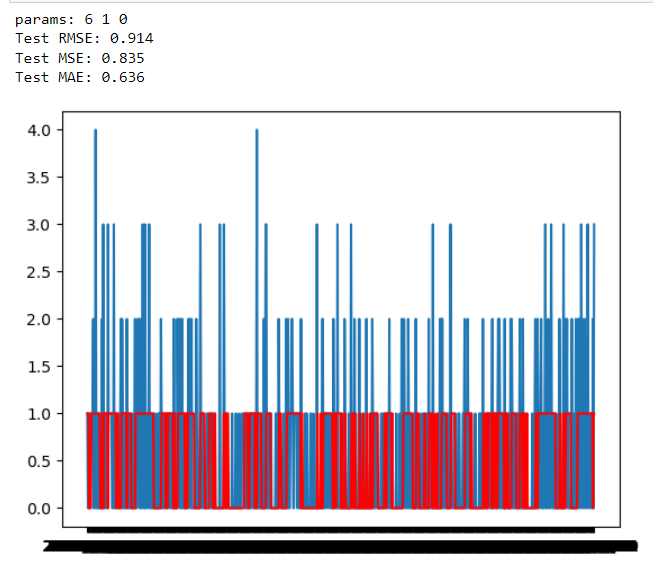
p=10, q=1, d=0:



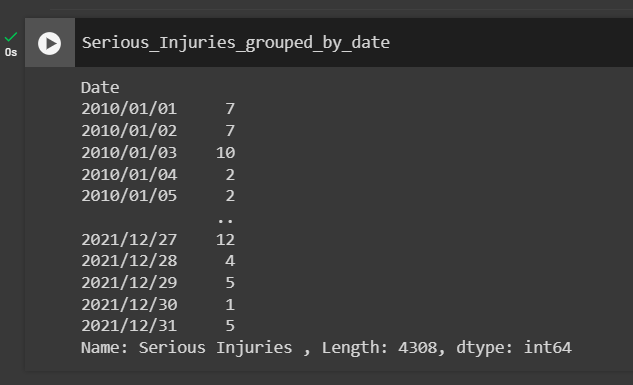
p=5, q=1, d=0:



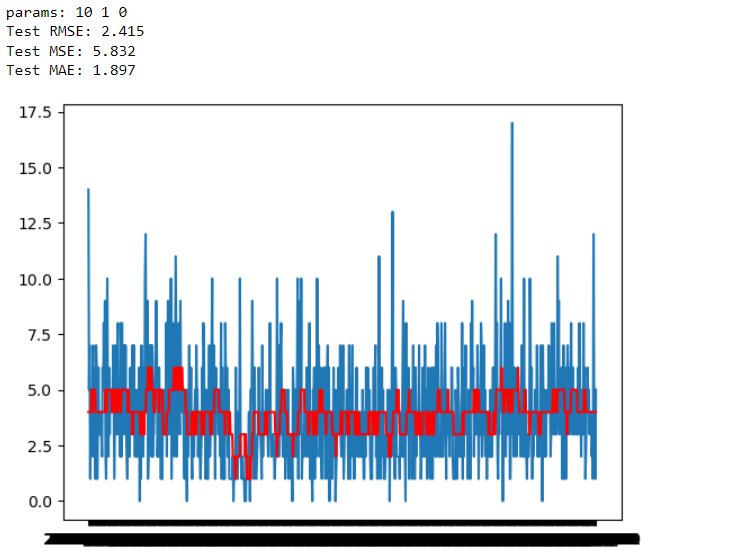
p=6, q=1, d=0:



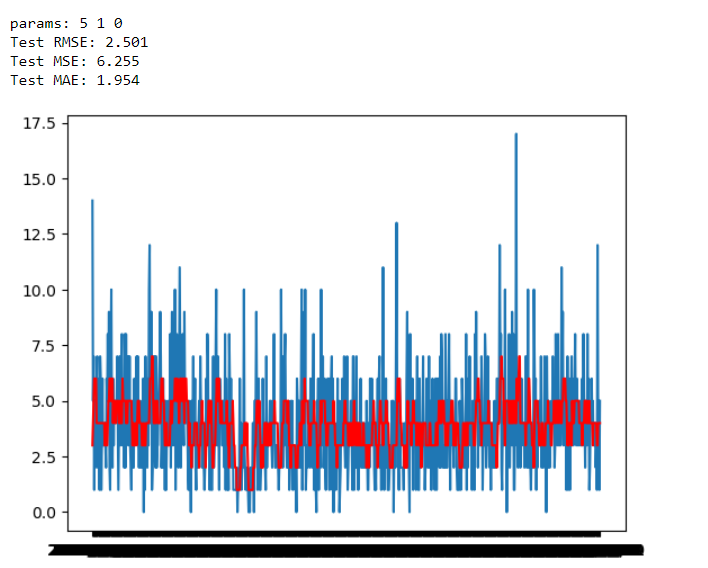
**Serious Injuries time series:** as shown in the picture below the index of this series indicates the Date which is sorted by day and the values are the number of Serious Inj. in each day



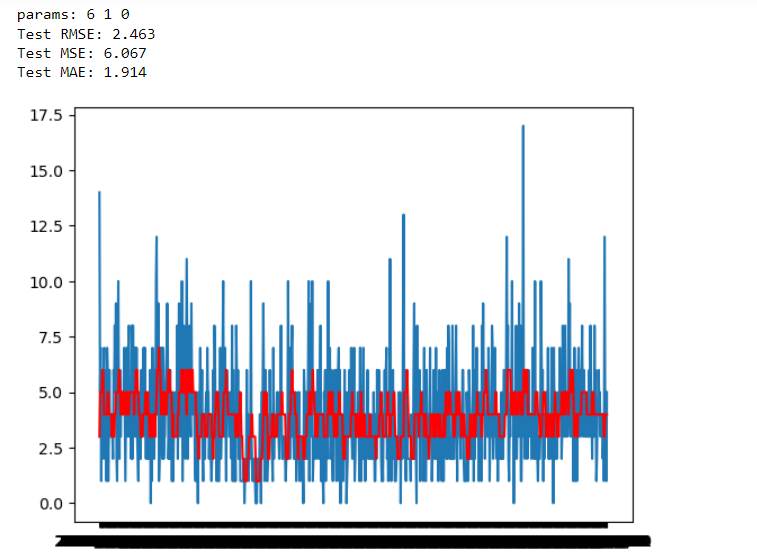
p=10, q=1, d=0:



p=5, q=1, d=0:

****

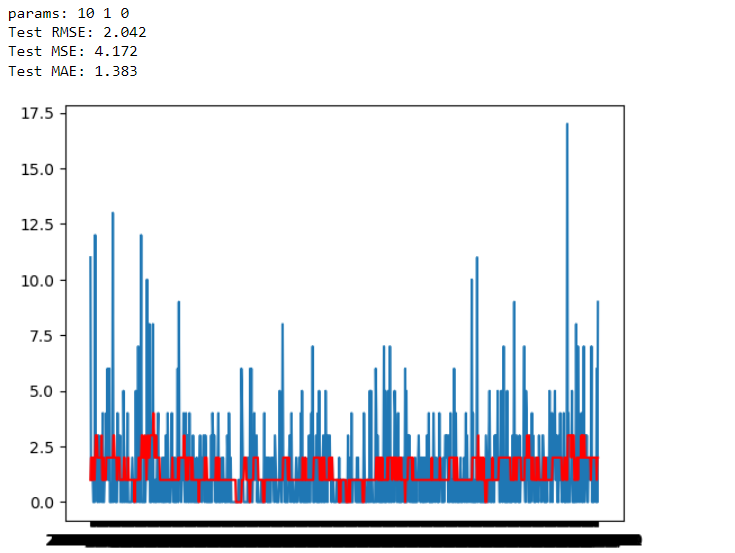
p=6, q=1, d=0:

****

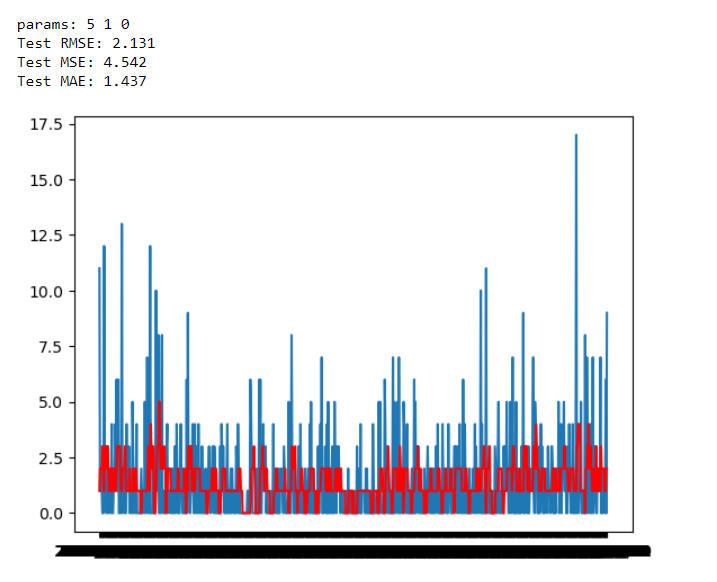
**Light Injuries time series:** as shown in the picture below, the index of this series indicates the Date which is sorted by day and the values are the number of Light Inj. in each day



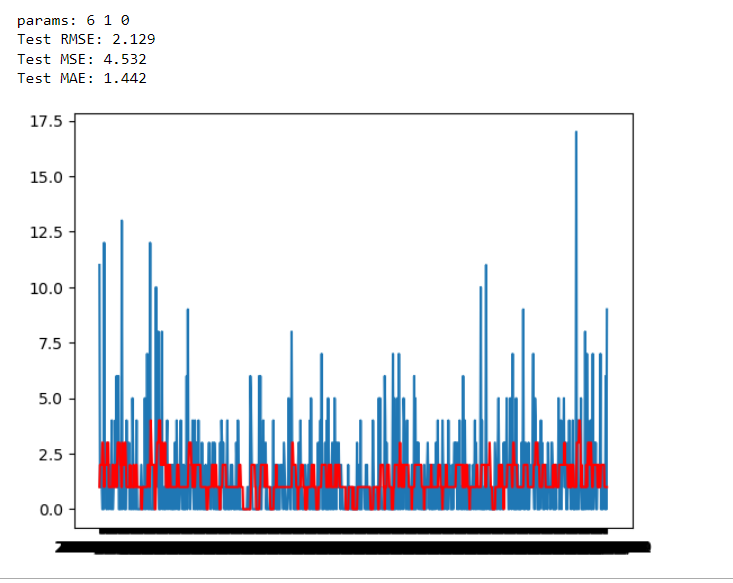
p=10, q=1, d=0:

****

p=5, q=1, d=0:

****

p=6, q=1, d=0:

****

Based on the results on test data for different series the best parameter is (10,1,0) because it leads to lower mse and mae. Therefore to predict the data of the upcoming year (365 days) we used ARIMA(10,1,0). The results are stored in 3 csv file with the name fat\_pred.csv, si\_pred.csv and li\_pred.csv which are related with fatalities, serious Injuries and light injuries.

**Regression with random forest and svm :**

**Preprocessing and model training**

After importing the data with *pandas.read\_csv* , we use the following command to observe al NaN values.

the output of this command shows which columns have nan values and how many NaNs they have



Kilometer Point 1

Road Speed Limit 2642

Surrounding Environment 32

Special Lane Presence 1349

Special Traffic Measures 40

Traffic Regulation and Priority 14970

Direction of Road 3496

Subtype of Road Section 14212

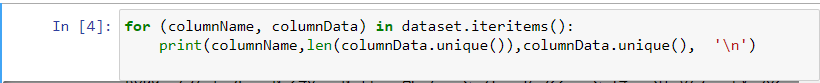
Road Ownership 10712

Road's Altimetric Layout 7637

Except of two columns (***Kilometer Point , Road Speed Limit***) , all other columns with NaN, have categorical values. To remove these nan value from the dataset first for all categorical columns the NaN values replaced with ‘ nan\_values ’ to be able then easily delete these value after one-hot encoding.

Next, in order to convert all categorical features into useful numerical features. First, we counted all unique (cathedral) values of columns to decide what kind of encoding method should be used for them.

the 2 most common encoding methods are one-hot encoding and label encoding. label encoding is used for ordered features and one-hot is good for unordered and independent features. since all of our features (except hour of the day) are unordered we use one-hot encoding. But we first need to be sure that the number of unique values is low. one-hot encoding is the best solution only when the number of unique values is low since it leads to high dimensionality and memory issues. with the large number of unique values, **Binary Encoding** or **Hash Encoding** are often used.



This command is used to count the unique values of each feature

Based on the output of this command these features were chosen to be encoded with one-hot encoding

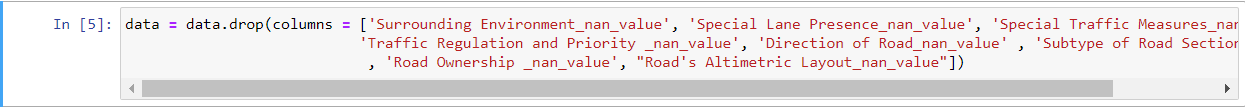
*'Area', 'Accident with Hit and Run', 'Fog Presence ', 'Surrounding Environment', 'Special Lane Presence', 'Special Traffic Measures', 'Weather Conditions ', 'Special Road Functions ','Influence of Road Objects', 'Severity of Accident ', 'Influence of Fog ', 'Influence of Environment', 'Influence of Traffic', 'Influence of Weather', 'Influence of Wind Intensity ' , 'Influence of Lighting', 'Influence of Special Measures', 'Influence of Road Surface ', 'Influence of Visibility ', 'Intersection Characteristics ', 'Speed Limit Information', 'Lighting Conditions ', 'Traffic Regulation and Priority ', 'Direction of Road', 'Subtype of Accident', 'Subtype of Road Section ','Subzone within Area', 'Road Surface Conditions ', 'Type of Road', 'Road Ownership ', "Road's Altimetric Layout" , 'Wind Conditions ', 'Day of the Week Grouping ', 'Time of Day Grouping ', 'Type of Accident ', 'Day Type'*

Among these features that are given to one\_hot encoding, *Special Lane Presence* and *Subtype of Accident* have most unique values with 14 and 13 respectively. The other categorical features with high unique values like *Road, Municipality Name and County Name* removed because these features are redundant and have no influence on predicting.

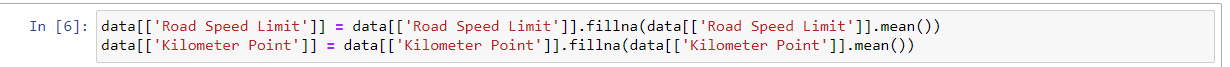
To do one-hot encoding we used get\_dummies method which is a methos of pandas library

*data = pd.get\_dummies(dataset, selected\_features )*

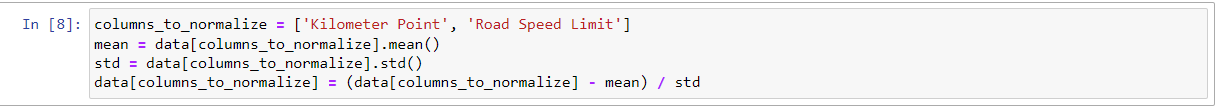
After that, we deleted all columns that have nan\_value. Notice that before encoding we replaced all NaN with ‘nan\_value’. After one\_hot encoding all these values were separated from primary values in a separate column and we easily could delete them. To do so, we used this command:



For the 2 numeric columns that include NaN (***Kilometer Point , Road Speed Limit***) NaN values were replaced with the mean of that column.



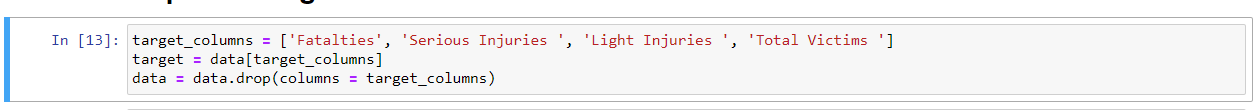
Since the two columns include a large range of numeric values we normalize their values with **Z-Score Normalization** method also known as **standardization scaling.** this method centers the values around the mean with a unit standard deviation, resulting in a mean of 0 and a standard deviation of 1.



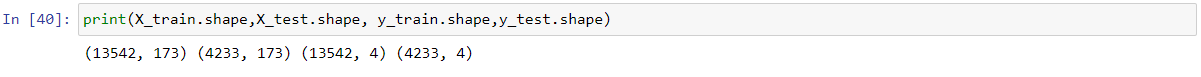
Then the redundant features like *'Year', 'Date ','Road ', 'Municipality Name ', 'County Name ', 'Province Name* ' that have no impact on the result were removed in order to gain more accurate result and also less computation.



And then we separate the target value including *'Fatalties', 'Serious Injuries ', 'Light Injuries ', 'Total Victims '* from the feature values.

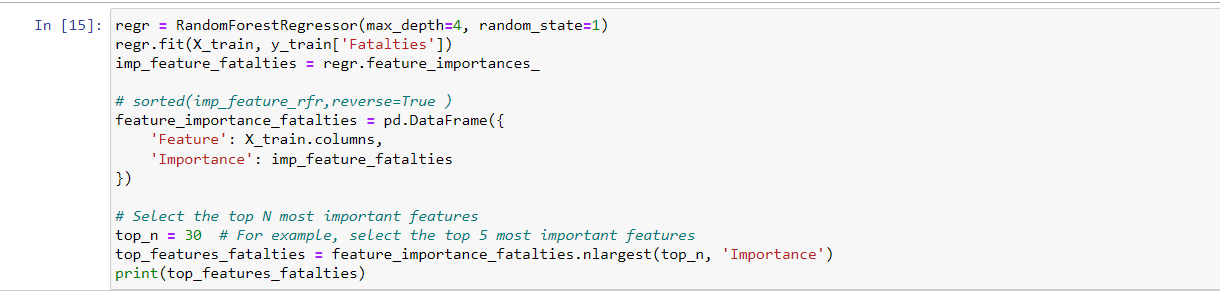


Now preprocessing process is done and it’s time to split train and test data. With *train\_test\_split* method we allocate 0.8 of data to train and 0.2 of data to test data

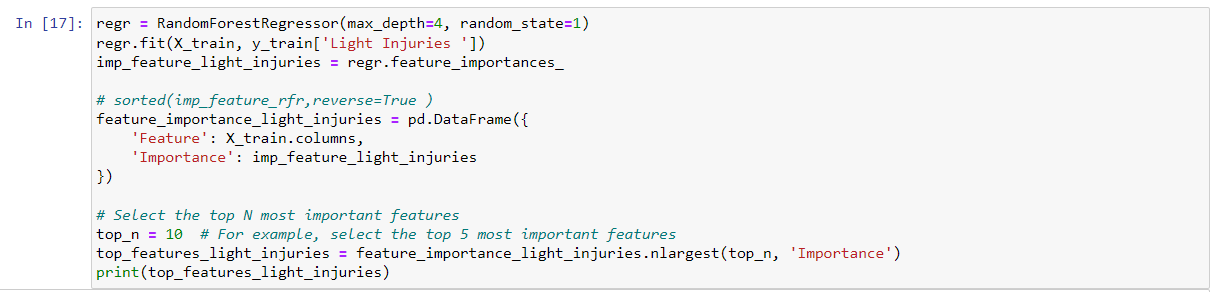


The size of train and test data

In order to identify important and influential features we used random forest regressor feature\_importance [1]. Since we have three target values (i.e. *'Fatalties', 'Serious Injuries ', 'Light Injuries'*), we calculated important features respect of each of these targets separately:







These are the top 10 important features that rfr recognized for each target with their importance rate

**Fatalities**:

|  |  |
| --- | --- |
| Severity of Accident \_Accident greu | 0.53402 |
| Severity of Accident \_Accident mortal | 0.43376 |
| Hour of Day | 0.00516 |
| Light Vehicles Involved | 0.00327 |
| Heavy Vehicles Involved | 0.00287 |
| Kilometer Point | 0.00214 |
| Direction of Road\_Un sol sentit | 0.00209 |
| Road Speed Limit | 0.00181 |
| Subtype of Accident\_Col·lisió frontal | 0.00155 |
| Day Type\_dg | 0.00125 |
| Units Involved | 0.00122 |
| Wind Conditions \_Vent fort | 0.00120 |
| Weather Conditions \_Pluja dèbil | 0.00083 |
| Bicycles Involved | 0.00082 |

**Serious Injuries:**

|  |  |
| --- | --- |
| Severity of Accident \_Accident greu | 0.355884 |
| Severity of Accident \_Accident mortal | 0.284307 |
| Light Vehicles Involved | 0.185400 |
| Subtype of Accident\_Col·lisió frontal | 0.043756 |
| Pedestrains Involved | 0.025954 |
| Units Involved | 0.015471 |
| Motorcycles | 0.010850 |
| Hour of Day | 0.009748 |
| Direction of Road\_Un sol sentit | 0.008250 |
| Type of Accident \_Sortida de la calcada sense ... | 0.006639 |
| Heavy Vehicles Involved | 0.006219 |
| Type of Road\_Autopista | 0.004351 |
| Day Type\_dill-dij | 0.004129 |
| Day Type\_dg | 0.003674 |

**Light Injuries:**

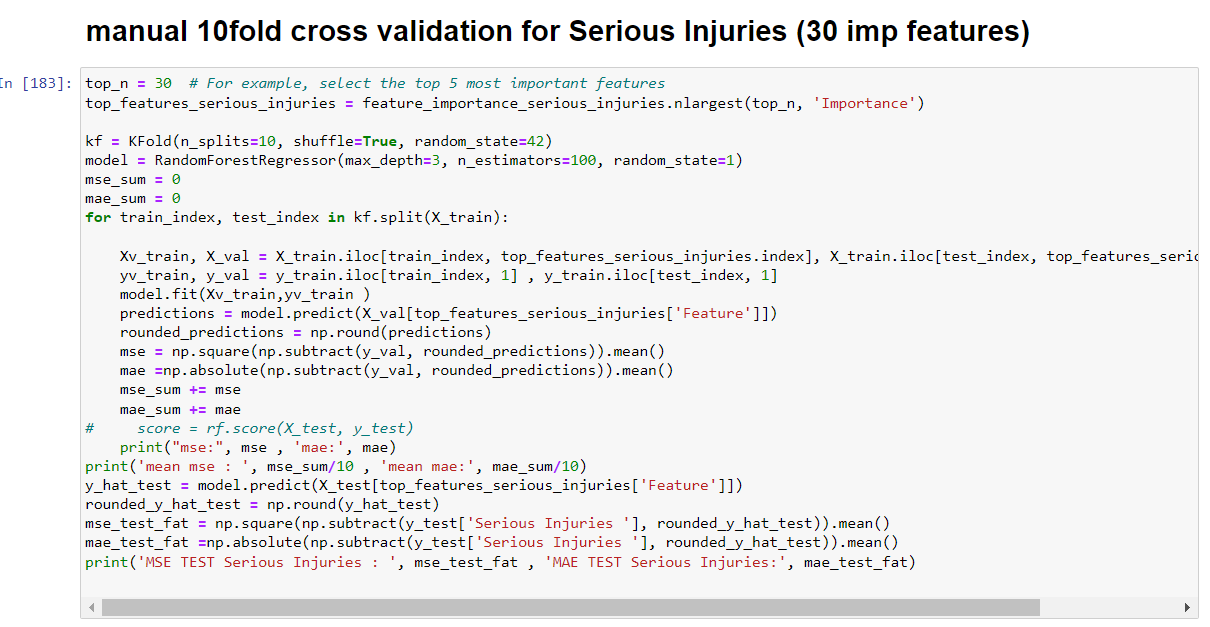
|  |  |
| --- | --- |
| Light Vehicles Involved | 0.636904 |
| Units Involved | 0.130238 |
| Subtype of Road Section \_Giratòria | 0.062038 |
| Motorcycles | 0.042355 |
| Hour of Day | 0.013374 |
| Heavy Vehicles Involved | 0.013069 |
| Special Lane Presence\_Carril bus | 0.012731 |
| Pedestrains Involved | 0.008859 |
| Kilometer Point | 0.008572 |
| Weather Conditions \_Pluja forta | 0.005350 |

**Training process with Random Forest regressor:**

we used 10 fold cross validation to more accurately train the model. For each of three targets we used different number of top important features (10,20, 30, 40) and no significant differences between the accuracy of each, was observed, so at last 30 important features were chosen for training and predicting.

The parameters of RFR model set as below:

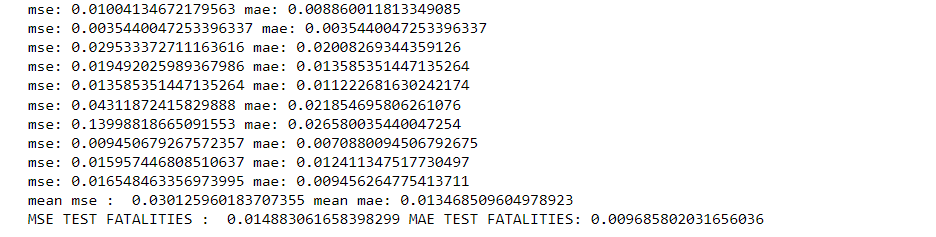
max\_depth=3, n\_estimators=100, random\_state=1



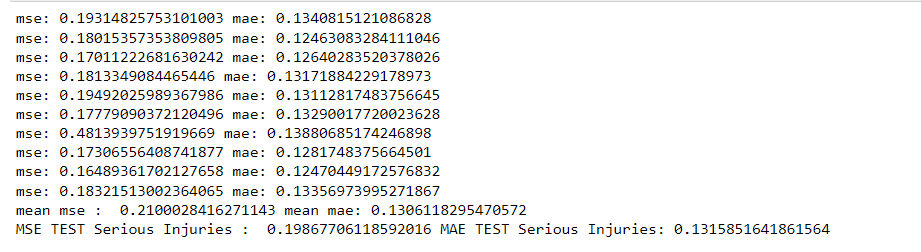
**Evaluation metrics: MSE , MAE**

**Result**:

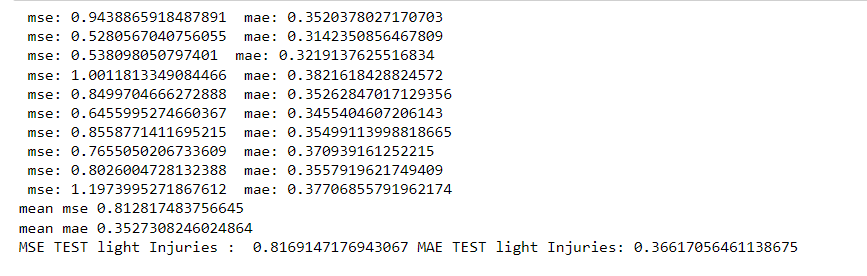
MSE and MAE on 10 folds validation set, mean of 10 folds and test set:

**Fatalities**: 

Serious injuries :

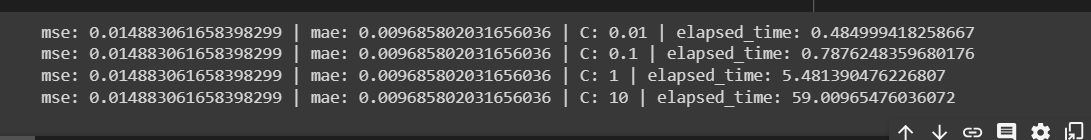


**Light Injuries:**

****

**Training process with SVM regressor:**

Parameter tuning in SVM models is the most important part, because of that we did a greedy search into different values of C for fatalities target to gain the best values and the results are as follows:

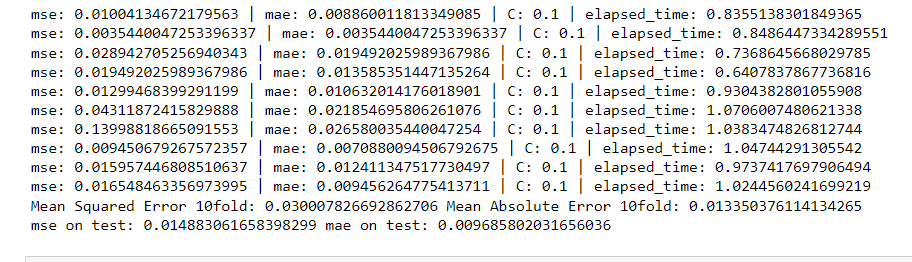


As can be seen the more C gets higher the more time need for execution, but no differences in accuracy was observed, therefore we chose C=0.1 for the training process with SVR.

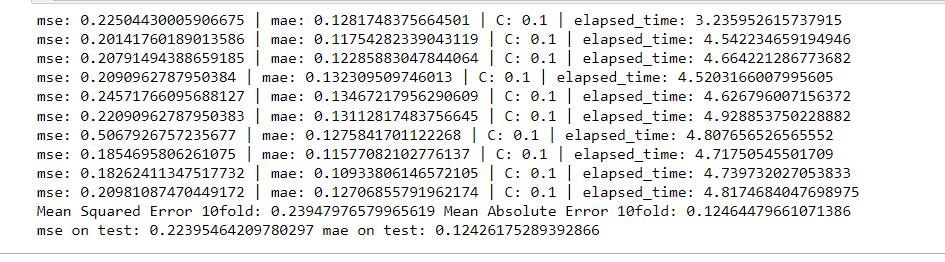
The other steps were like before as we did with rfr and the important features were those that with rfr we had chosen.

**Results**

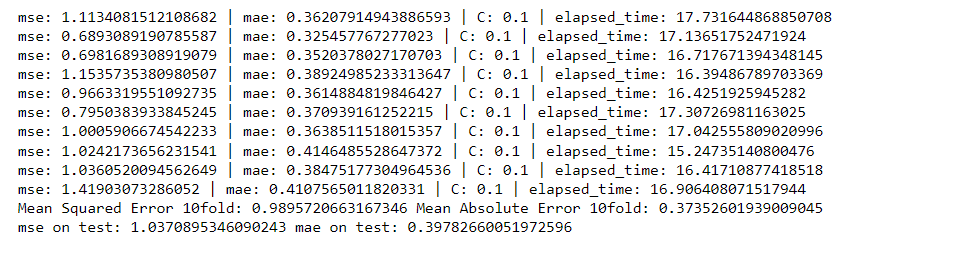
**Fatalities**

****

**Serious injuries**

****

**Light injuries**

****

**Conclusion:**

in this study different model have been investigated for road safety dataset of Calanuya, time series random forest regressor and support vector machine regressor. The dataset had 3 independent target that for each of them a model has been trained separately. Base on the results and the studies, the rfr models outperforms the svr models for 3 types of targets and also is more efficient for time execution, since the time complexity of svr is very dependent to its parameters. Time series model are the best when there is no information about the datasets but the slot of time. However, choosing the best model and tuning the parameters in time series is very crucial.

For this Calanuya dataset, we only examined ARIMA model based on our studies and the results on different parameters indicates that the model needs to be improved and couldn’t well adapted on the dataset.

**References**:

[1] <https://www.linkedin.com/pulse/applied-time-series-forecasts-traffic-accidents-injuries-rami-kanaan/>

[2] Getahun, K.A. Time series modeling of road traffic accidents in Amhara Region. *J Big Data* **8**, 102 (2021). <https://doi.org/10.1186/s40537-021-00493-z>

[3] Nassiri H, Mohammadpour SI, Dahaghin M. Forecasting time trend of road traffic crashes in Iran using the macro-scale traffic flow characteristics. Heliyon. 2023 Mar 11;9(3):e14481. doi: 10.1016/j.heliyon.2023.e14481. PMID: 36967875; PMCID: PMC10036660.

[4] https://www.capitalone.com/tech/machine-learning/understanding-arima-models/

[5] https://people.duke.edu/~rnau/411arim.htm

[6] <https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/>

[7] Megnidio-Tchoukouegno, M.; Adedeji, J.A. Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector. Sustainability **2023**, 15, 2014. <https://doi.org/10.3390/su15032014>

[8] Silva, Philippe & Andrade, Michelle & Ferreira, Sara. (2020). Machine learning applied to road safety modeling: A systematic literature review. Journal of Traffic and Transportation Engineering (English Edition). 7. 775-790. 10.1016/j.jtte.2020.07.004.

[9 ] Obasi IC, Benson C. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. Heliyon. 2023 Jul 29;9(8):e18812. doi: 10.1016/j.heliyon.2023.e18812. PMID: 37560691; PMCID: PMC10407198.

[10 ] Shakil, Ahmed., Md., Akbar, Hossain., Sayan, Kumar, Ray., Mafijul, Islam, Bhuiyan., Saifur, Rahman, Sabuj. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. Transportation research interdisciplinary perspectives, 19:100814-100814. doi: 10.1016/j.trip.2023.100814