DSCC440 Data Mining: Final Project
Prof J. Luo, University of Rochester, NY

# *Insights into Road Accident Dynamics*

## *A Machine Learning Approach to Accident Severity Prediction and Future Preparedness*

Neel Agarwal
Master of Data Science (pursuing)
Goergen Institute of Data Science
University of Rochester, New York
nagarwa9@ur.rochester.edu

Jayant Patil
Master of Data Science (pursuing)
Goergen Institute of Data Science
University of Rochester, New York
jpatil@ur.rochester.edu

***Abstract*—This study addresses the pressing concern of road accidents by exploring contributing factors in specific zones and proposing an approach that identifies combinations of factors leading to different severities of injury. For addressing the same adaptable machine learning models are considered including methods like clustering, pattern mining, classification, and time series analysis. Geospatial clustering methods pinpoint specific zones for in-depth analysis, providing a detailed examination. Frequent pattern mining algorithms aid in recognizing recurring patterns within accidents in those specific zones. Examining factors leading to varying injury severity levels is crucial, with common classifiers facing challenges in data transferability. Models like Random-forest demonstrate robustness across diverse applications. A brief time series forecasting analysis offers insights into preparedness and measures to avert future incidents. This research aims to contribute to a comprehensive understanding of road accidents, guiding targeted interventions for enhanced road safety.**

***Keywords— crashes; gaussian-mixture; zones; fp-growth; random-forest; prophet***

## 1. Introduction

Road accidents represent a pressing societal concern, prompting a comprehensive examination of contributing factors to mitigate their occurrence. An effective approach involves identifying accident hotspots and discerning commonalities and distinctions between these geographical zones, offering a nuanced understanding of location-specific risk factors. While statistical models are esteemed for their robust theoretical foundations in accident analysis, machine learning models provide adaptability without stringent assumptions.

This research advocates for the utilization of geospatial clustering methods to pinpoint specific zones for in-depth study, offering a more granular analysis compared to studying large areas collectively. Incorporating frequent pattern mining algorithms facilitates the identification of recurring patterns within accidents. Applying such algorithms to individual clustered zones enables a focused exploration of common causes, thereby facilitating targeted measures to address frequent crash occurrences.

In light of the escalating frequency of accidents, an examination of factors leading to varying severity levels of injuries becomes paramount. Classifiers, including Logistic Regression, KNN, and decision trees, are commonly employed in existing models. However, these models encounter challenges related to the transferability of data attributes across diverse regions. Notably, the robustness of models like random forest proves advantageous across a broader spectrum of applications.

Lastly, a brief analysis employing time series forecasting, rooted in previous accident occurrences, holds promise for gaining valuable insights into levels of preparedness and essential measures required to proactively avert future incidents. This research aims to contribute to a comprehensive understanding of road accidents, thereby informing targeted interventions and policy measures to enhance road safety.

## 2. Related work

The injury severity from traffic accidents is an essential metric of damage caused by traffic accidents. Multiple factors cause traffic accidents of different degrees. Many machine learning and deep algorithms along with varying metrics for analyses have been cited in the study of traffic accidents.

Kwon et al., (2015) introduced a system integrating Naive Bayes and Decision Trees, applied to California accident datasets spanning approximately seven years. The evaluation revealed that the model employing binary regression exhibited lower performance compared to the Naive Bayes model, which, in turn, surpassed the precision of the model generated by Decision Trees. [1]. Sachin et al., (2015) proposed a

method aimed at standardizing accident data, employing the K-modes clustering algorithm for data segmentation. Comparison between the Apriori algorithm's application on both segmented and original datasets was conducted, showcasing the effectiveness of segmentation in the analysis. Subsequently, trend analysis was executed on each cluster and the complete dataset, affirming the significance of segmentation in facilitating comprehensive insights before analysis [2]. Velivela Gopinath (2017) conducted an analysis of accident datasets employing clustering techniques such as SOM (Self Organizing Map) and K-modes, in addition to classification techniques including Support Vector Machine (SVM), Naïve Bayes, and decision trees. This comprehensive approach aimed to uncover specific patterns concerning road users, and notably, higher accuracy was attained utilization of clustering techniques. The study's outcomes distinctly revealed the circumstances that influence and the most frequently involved parties in accidents, delineating whether the driver, passenger, or pedestrian is more affected [3]. Liling Li (2017) examined the correlation between the fatality rate and various factors associated with road accidents, including collision type, weather conditions, surface conditions, light circumstances, and the presence of an intoxicated driver. This investigation utilized association mining rules, revealing that accidents involving drunk drivers exhibited a notably higher fatality rate. Additionally, the authors employed a naive Bayes classifier within a classification model and implemented a simple K-means clustering algorithm in their analysis [4]. Imran et al., (2019) introduced a system designed for conducting regression analysis to uncover connections between multiple factors like weather, road conditions, driving habits, etc., and their correlation with accidents. This system assesses the accuracy of these relationships by initially scrutinizing the functional connections between the variables. Within this study, the impact of individual components is explored through both linear simple regression and detailed multivariate regression analysis methods [5]. Kwayu et al., (2019) (utilized AdaBoost, logistic regression, and Naive Bayes, along with Random Forest as an extension, aiming to pinpoint high-risk roadways for Michigan traffic agencies and ascertain the critical variables involved. Evaluation of the models was conducted based on performance metrics such as ROC, precision, and recall. As per the findings, the Random Forest classifier exhibited superior performance, achieving an accuracy rate exceeding 76%, surpassing the other classifiers in the study [6]. Guo Y et al., (2019) conducted a research study utilizing traffic data from Tokyo's metropolitan region to investigate the impact of rainfall on both accidents and congestion spanning an eight-year period. Throughout this duration, 5,700 accidents occurred during rainy hours out of a total of 42,041 incidents. The study revealed a frequency of 1.5 accidents per hour in dry weather and 0.85 accidents per hour in wet weather. Over the course of six years, rainfall emerged as a primary contributing factor. Rainfall represents just one of several factors that can influence road conditions and traffic, consequently leading to various types of collisions, thereby altering both the frequency and severity of accidents [7]. Yassin et al. (2020) suggested a fusion approach integrating the strengths of random forest and K-means techniques to discern the most predictive traits of road accidents. Utilizing K-means, a novel factor was generated within the training model, extracting latent insights from traffic accident data. The classification algorithms demonstrated exceptional effectiveness, particularly with the random forest method achieving an accuracy rate of 99.86%. However, a notable limitation lies in the selection of the K value within K-means, as it significantly impacts the model's accuracy and poses a considerable challenge [8]. Yang et al. (2020)The study employed the UK road safety dataset to propose a method for the prediction of severity of traffic accidents based on Deep Forests algorithm. The paper proposed that fewer hyper-parameters in the deep forest were more conducive to the transplantation of models, implying that a set of hyper parameters can be applied to different datasets [9]. Xiao et al. (2020) utilised a BP neural network that is optimized by the genetic algorithm for predicting the accident severity under certain conditions. They presented three parameters for testing and analyzed them for improving the prediction accuracy of the BP neural network [10]. Hernandez et al. (2021) suggested a method that takes annual daily accident and traffic data. Utilizing radial basis functions, an estimation model is formulated for each area's data cluster, generating a variable map. Subsequently, the validity of the constructed variable model is assessed by testing it against the data employing cross-validation techniques. [11]. Yaman et al. (2022) employed fuzzy data mining technology to scrutinize the elements influencing the severity of injuries in traffic accidents. Variables such as age, gender, seatbelt utilization, as well as involvement with alcohol and drugs were among the factors investigated. The study yielded standardized variables of independent significance that impact the severity of injuries [12]. Giuseppe Guido et al., (2022) suggested employing optimal control parameters in developing the group method of data handling model, resulting in an enhanced binary classification model. This involved integrating the Grasshopper Optimization Technique with Support Vector Machines (SVM) to optimize three parameters. A deterministic analysis was conducted to identify influential factors affecting the involvement of vehicles and the count of individuals affected in accidents within rural areas. The Support Vector Classification model was utilized to delineate an optimal hyperplane within the multidimensional space and determine the significant parameters crucial for detection [13].

## 3. METHODOLOGY

### 3.1. *Workflow*

Figure 3.1.1 proposes the workflow for the study.

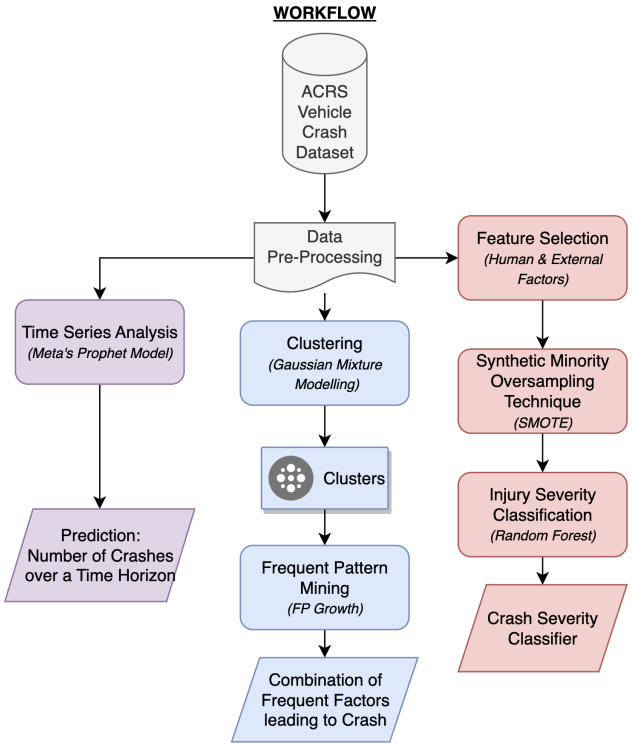


Figure 3.1.1 Workflow for analysis of ACRS crash dataset.

### 3.2. *Data Set Description*

The dataset utilised for the study is publicly available and reports details of all traffic collisions occurring on county and local roadways within Montgomery County, as collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police. The data has been taken from the official website of the United States government. The link for the same is as under:- *https://catalog.data.gov/dataset/crash-reporting-drivers-data*

The data contains 43 attributes and 167,799 observations. A brief description of the attributes is given in table 1.1. Some of the attributes have not been included in the table, like, "Report Number", "Local Case Number", "Person ID", "Driver's License Plate", "Vehicle ID", "Speed Limit", "Driverless Vehicle", "Parked Vehicle", as they are believed to have no descriptive or inferential significance in the study.

Table 1.1 Attributes contained in the dataset and their brief description

| Attribute | Description |
|---|---|
| Agency Name | Department or Police station that registered the case, e.g., Montgomery County Police |
| ACRS Report Type | Type of crash as given by automated crash reporting system, e.g., Injury Crash |
| Crash Date/Time | The data and time of crash in the format MM/DD/YY 24 hr clock |
| Route Type | The type of route, e.g., County, Municipality, State |
| Road Name | Name of the road where the crash occured, e.g., SELFRIGDE |
| Cross Street Type | Refers to the classification of street that intersects crosses another street, e.g., Maryland (State) |
| Cross Street Name | Name of the intersecting street, e.g., Randolph road |
| Off Road Description | Gives the location description in case the vehicle was off road, e.g., Parking Lot |
| Municipality | Gives the municipality area where the crash occured, e.g., Rockville |
| Related Non - motorist | Gives the type if a non-motorist was also involved in the accident , e.g., Bicyclist |
| Collision Type | The type of collision with respect to directions, e.g., Head on left turn. |
| Weather | The weather condition in which the accident occurred, e.g. Raining |
| Surface Condition | The surface condition of the road in which the accident occured, e.g., Dry |
| Light | The amount of outdoor light available, e.g., Dusk |
| Traffic Control | The presence or absence of any traffic controls or signals, e.g., Stop Sign, Traffic Signal, No controls |
| Driver Substance Abuse | The status of driver's intoxication if any during the crash, e.g., Alcohol present, None detected, etc. |
| Non-motorist substance abuse | The status of non-motorist's intoxication if any during the crash, e.g., Illegal Drug present, Alcohol present, etc. |
| Driver At Fault | Whether or not the driver was at fault for the accident, e.g., Yes, No, Unknown |
| Injury Severity | Ordinal variable tagging the the extent of injury severity, e.g., No apparent injury |
| Circumstance | Any extra circumstantial information on the accident, e.g., Animal |
| Driver Distracted By | Information on type of distraction if the driver was distracted, e.g., Inattentive or lost in thought |
| Vehicle Damage Extent | Ordinal variable tagging the extent of vehicle damage, e.g., Superficial |
| Vehicle First Impact Location | Angular location of first impact in 360 degrees divided into 12 segments, e.g., eleven o'clock |
| Vehicle Second Impact Location | Angular location of second impact in 360 degrees divided into 12 segments, e.g., eleven o'clock |

| | |
|---|---|
| Vehicle Body Type | The type of vehicle involved in the crash, e.g., Passenger Car |
| Vehicle Movement | The last known vehicle movement type right before the crash, e.g., slowing or stopping |
| Vehicle continuing direction | The direction in which the vehicle continued to go after the accident, e.g,. north, east, etc. |
| Vehicle going direction | The direction in which the vehicle was previously going right before the accident, e.g., west, north, etc. |
| Vehicle year | The year of manufacture of the vehicle involved in the crash, e.g, 2004 |
| Vehicle make | The name of the company that manufactured the vehicle, e.g,. Honda |
| Vehicle Model | The variant of the particular type of vehicle involved in the crash, e.g., Fusion |
| Equipment problems | Describes type of equipment problems that could have followed or caused the accident if any, e.g., air bag failed |
| Latitude | Gives the latitude value of the location of crash |
| Longitude | Give the longitude value of the location of crash |
| Location | Gives the latitude and longitude co-ordinate as a tuple |

### *3.3. Data Preprocessing*

The columns which have not been mentioned in the table above were dropped from the dataset due to irrelevance. While most of the attributes had zero null values, the ones that had missing data or were tagged as "Unknown" in some of the observations were replaced by "N/A" implying "Not applicable" as, for example, if there was no non-motorist involved in a particular accident, and the "Related Non-motorist" column had an empty value for that observation, it was replaced by "N/A". A number of columns were processed for extra leading-trailing whitespaces, so as to be able to extract the value (string, numeric or time) of a cell in the correct format. Some irregularities in the attribute "Vehicle Year" were observed and handled accordingly. For example, if instead of a four digit year, an observation reads "207", it should be implied to be the year "2007". Similarly, in the column for "Vehicle Make", there were multiple instances that seemed to refer to the same company, but had occurred as unique values due to spelling errors. For example, "TOYT" clearly referred to the company "TOYOTA" and "HYUN" referred to "HYUNDAI". Fuzzy matching was used to deal with this and 1844 unique values were improved down to 931.

In the second branch of the study, i.e., before applying classification taking "Injury Severity" as the target class variable, it was observed that the target attribute had a severe imbalance in data, implying that the frequency of total observations for each class varied significantly which could skew the results of prediction. This was addressed using resampling of the minority classes called SMOTE (Synthetic Minority Oversampling Technique).

## 4. Data Visualization

For the purpose of exploratory analysis, data visualization has been used to uncover the relationships between multiple variables in the accident data thereby enabling quality assessment of the attributes and the overall dataset and drawing insights about their usability for further investigation.

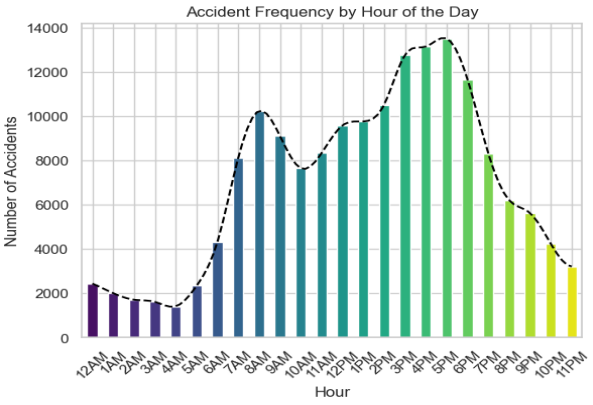


Figure 4.1 The frequency of accidents by hour of the day.

The attribute for "Crash Data/Time" has the information available for hour of the day when a given crash occured. Using that information a bar chart for the frequency of accidents by hour of the day has been plotted in figure 4.1. It is observed that the frequency of occurrence of accidents was highest between 3 p.m. and 6 p.m. and the least between 12 a.m. and 4 a.m.

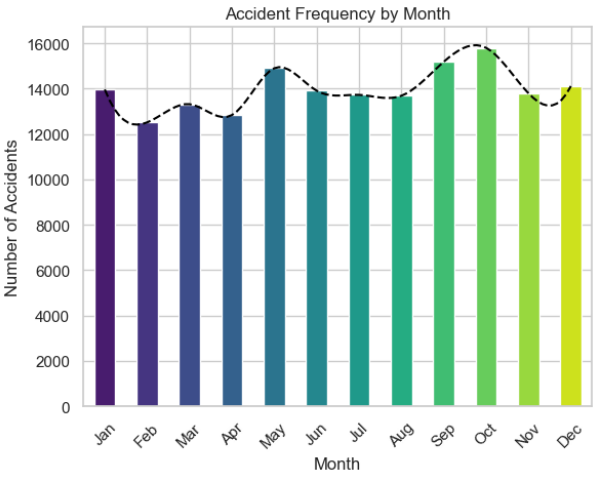


Figure 4.2 The frequency of accidents by month of the year.

Similarly, the same attribute also gives information about dates at which accidents occurred. Figure 4.2 shows a bar plot

for the frequency of accident occurrences over different months of the year. It is explicit from the plot that there is no significant difference in the number of accidents across different months. However, the numbers peak around the months of September-October and the least frequency is observed in the months of February and April.

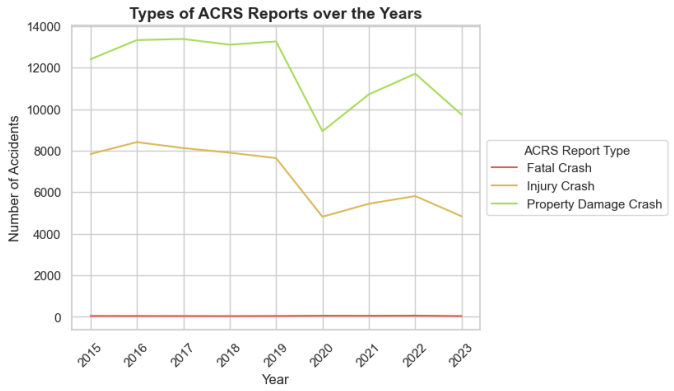


Figure 4.3 Frequencies of ACRS Crash report types over the years

The dataset has an attribute called "ACRS Report Type" that gives the classification of the type of crash as reported by the "Automated Crash Reporting System" as "Injury Crash", "Property Damage Crash" and "Fatal Crash". Figure 4.3 shows a grouped line chart that compares the frequency trends of the three classes over the years. It is evident from the graphs that the order of maximum to minimum number of crash types is "Property Damage Crash" "Injury Crash" and "Fatal Crash" respectively. Also the "Property Damage Crash" and "Injury Crash" types show a steep dip in respective frequencies in the year 2020 which is suspected to be due to Covid-19 pandemic and peak again in 2022. Due to the comparatively low frequency of "Fatal Crash" types, it has been plotted in a separate figure (Figure 4.4).

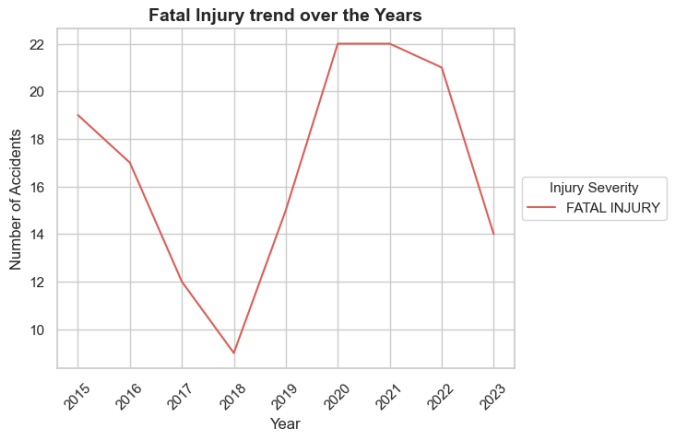


Figure 4.4 Frequency trend of "Fatal injury" ACRS Report type over the years

It is explicit in the graph shown by figure 4.4 that the number of "Fatal injury" crash type ACRS reports continued to decrease from the year 2015 to 2018 and then again rose until 2020. It stabilised from 2020 to 2021 and then again continued to decrease until 2023. It is also evident that the number of "Fatal Crash" report types spanned from 5 to 22 which is significantly lower than other ACRS report types.

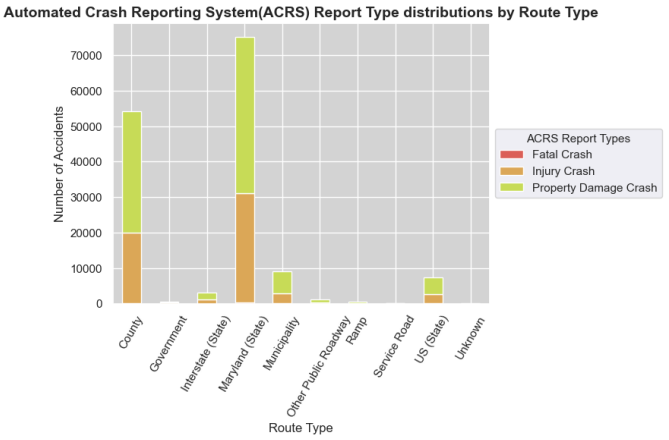


Figure 4.5 Accident frequencies by cross street route type stacked with ACRS Report type.

To find the distribution of accident frequencies by cross street route type (stacked individually by ACRS Report type), a stacked bar chart has been plotted in figure 4.5. It is observed that Maryland (State) type route had the most number of accidents followed by Counties. In each of them, the "Property Damage Crash" ACRS report type shows the maximum frequency followed by "Injury Crash" and due to the highly skewed distribution, "Fatal Crash" types are not visible.

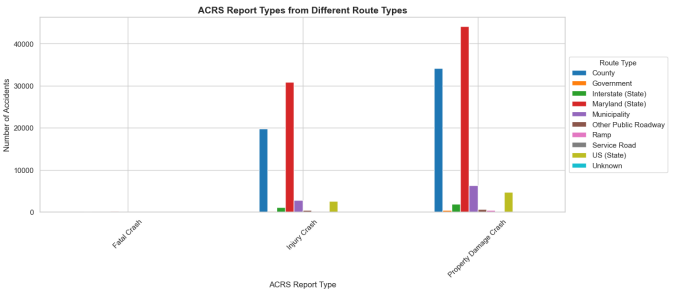


Figure 4.6 Number of accidents by Route Type (grouped by ACRS report type)

To find the distribution of accident frequencies by route type (grouped by ACRS Report type), a bar chart has been plotted in figure 4.6. It is observed that Maryland (State) type route had the most number of accidents followed by Counties for

"Injury Crash" and "Property Damage Crash" types. Overall, the "Property Damage Crash" ACRS report type shows the maximum frequencies followed by "Injury Crash" and due to the highly skewed distribution, "Fatal Crash" type frequencies are not visible.

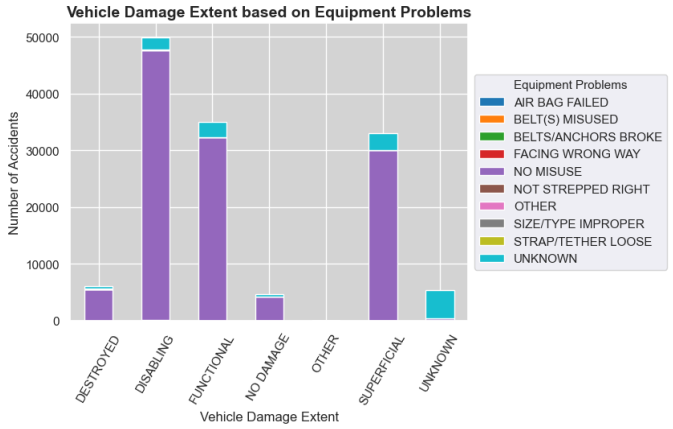


Figure 4.7 Accident frequencies by Vehicle Damage Extent (stacked by Equipment problems)

The attribute "Vehicle Damage Extent" gives a brief insight on the extent of damage caused to the vehicle by the accident. Figure 4.7 shows the frequencies for each category of "Vehicle Damage Extent" with a stacked bar plot (by Equipment problems). From the bar plot it can be observed that most of the accidents involved no misuse of equipments. Among the "Vehicle Damage Extent" categories, "Disabling" is observed to have the maximum frequency followed by "Functional" and "Superficial". "Destroyed" and "No damage" are observed to have much lower frequencies comparatively.

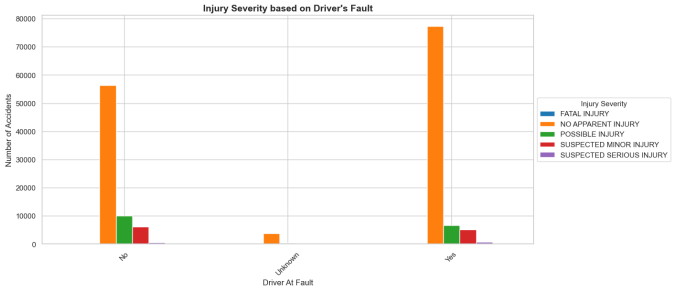


Figure 4.8 Number of accidents by "Injury Severity" grouped by "Driver at fault".

Figure 4.8 shows the frequency of accidents by "Injury Severity" grouped by "Driver at fault" and it clearly signifies that irrespective of the driver being at fault, the frequency of accidents with "No apparent injury" was the highest, followed by "Possible injury" and "Suspected minor injury" in both cases. Overall the frequency of accidents with "Yes" for "Driver at Fault" was more than the other two categories.

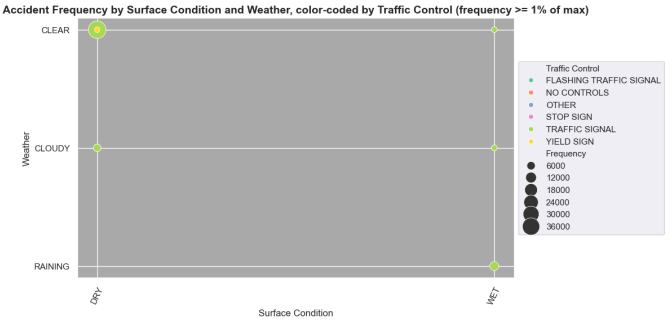


Figure 4.9 Bubble plot showing the impact of weather and surface conditions on frequency of accidents.

To find the impact of weather conditions and surface conditions on the frequency of accidents, a bubble chart has been plotted in figure 4.9 using the categorical attributes "Weather" and "Surface Condition" and it shows that the maximum frequency of accidents occured during "Clear" and "Dry" conditions and the least occurred during "Cloudy" and "Wet" conditions. This clearly signifies that weather and surface conditions did not have a major impact on the number of accidents.

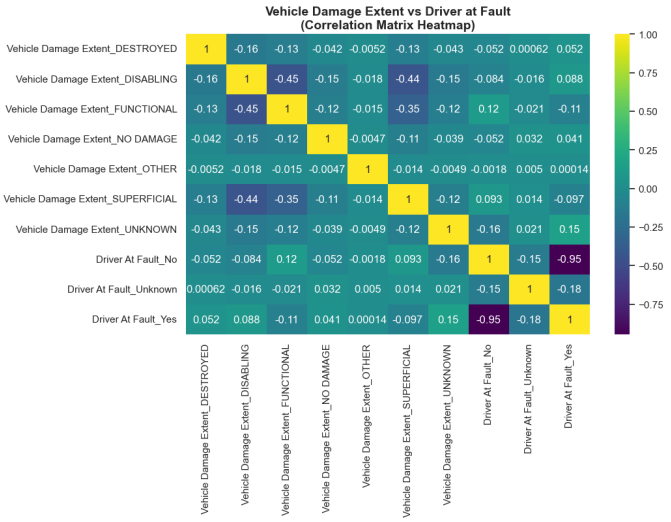


Figure 4.10 Heat map between "Vehicle Damage Extent" and "Driver at fault"

In order to explore whether the state of driver's fault had any effect on the extent of vehicle damage, a heat map has been plotted in figure 4.10. The matrix shows that when the driver is not at fault, the vehicle damage extent shows maximum correspondence to the "Functional" category with a correlation of 0.12 (ignoring the "Unknown" category). And when the driver is at fault, the vehicle damage extent shows maximum

positive correspondence to the "Disabling" category with a correlation of 0.088 (ignoring the "Unknown" category).

## 5. EXPERIMENT

### 5.1. *Clustering*

Clustering is a form of unsupervised learning with a goal to group similar data points with respect to certain features and characteristics. In geo-spatial clustering, the main attribute used to cluster data points is the similarity based on geological location.

The crash data used in this study has attributes "Latitude" and "Longitude" giving the geological co-ordinates for the location of occurrences of accidents. The main aim to cluster the data using geological co-ordinates is to be able to divide the "Montgomery County" into zones with respect to similarity and dissimilarity of patterns and explore the differences and commonalities in factors that accompany these accidents in respective zones through a pattern mining algorithm.

The proposed clustering algorithm used for the study is a Gausian Mixture Model (GMM). It is a probabilistic model that represents a mixture of multiple Gaussian distributions. GMMs are commonly used for clustering, where each cluster is modeled as a Gaussian distribution. Unlike k-means, which assumes that each cluster is spherical and has the same variance, GMMs can model clusters with different shapes and sizes. Once the GMM is trained, each data point is assigned to the cluster associated with the Gaussian component that has the highest probability for that point. GMM provides a soft clustering result, meaning that each data point has probabilities of belonging to different clusters. This is in contrast to k-means, which provides hard assignments.

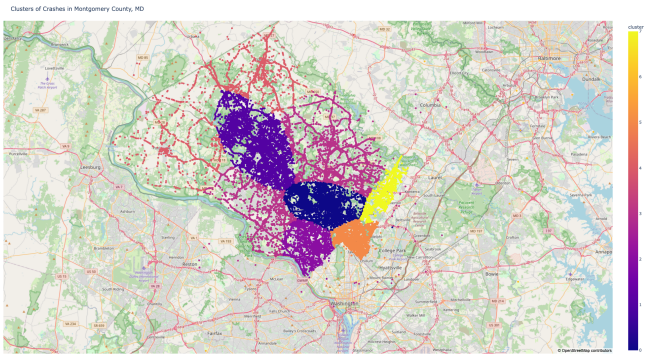


Figure 5.1.1 Clustering "Montgomery County" into 7 zones

The number of clusters is a user defined hyperparameter and therefore for determining the apt number of clusters (zones) for the division of the dataset, trial and error method was used as well as a rough estimate from the number of unique values of the "Municipality" column for dividing the area . The final number of optimum clusters used was 7. Figure 5.1.1 shows the clustered zones after applying the Gausian Mixture algorithm on the co-ordinate data. Two of the zones tend to closely map to the localities in cluster plot of "Municipality" areas in figure 5.1.2.

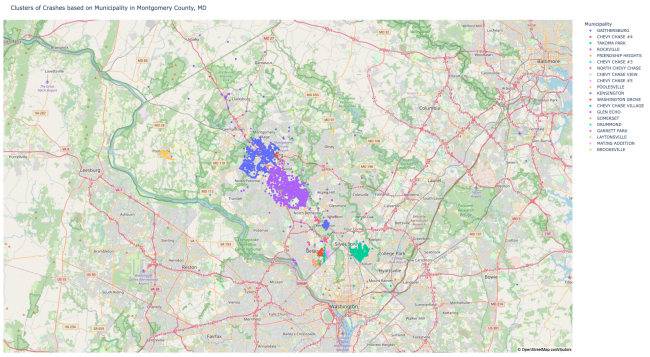


Figure 5.1.2 Map showing different "Municipality" areas as given in the dataset.

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

### 5.2 *Frequent Pattern Mining*

The next step after dividing the accident data into zones of occurence, is to find the most commonly occurring combinations of factors that have influenced or accompanied the accident occurences in each respective zone and also compare the commonalities and differences in these factors between different zones.

The algorithm used for mining frequent patterns in each zone is FP Growth (Frequent Pattern Growth). This algorithm builds association rules and renders frequent itemset set without candidate generation. The key innovation of FP-Growth is the use of a data structure called the FP-Tree (Frequent Pattern Tree). The FP-Tree represents the dataset in a compact and efficient manner.The algorithm first scans the dataset to identify frequent items and construct a frequency-ordered list. Then, it builds the FP-Tree by adding transactions one at a time, creating branches for each transaction based on the order of frequent items. Once the FP-Tree is constructed, frequent patterns can be mined efficiently by recursively exploring the tree. The algorithm uses a depth-first search strategy to traverse the tree and generate frequent itemsets. The algorithm utilizes a divide-and-conquer strategy. At each step of the recursion, a conditional FP-Tree is created for each frequent item, representing the subproblem of finding frequent patterns containing that item. Frequent patterns are grown

incrementally by appending items from the conditional FP-Tree. This process is repeated recursively until all frequent patterns are discovered. FP-Growth is particularly efficient because it avoids the need to generate candidate itemsets, a common step in algorithms like Apriori. The FP-Tree structure eliminates the generation of candidate sets, leading to faster performance.

The algorithm was applied to the data for each cluster (zone) individually to be able to mine frequent patterns (of factors) for every zone. The minimum support was kept at 40% and the association rule threshold was kept at 80%. Due to the vastness of the data additional constraints were applied for the analysis of similarities of patterns in different zones. The maximum number of frequent patterns for each cluster was capped at 30 and the minimum length of pattern was kept as 3. Table 1.3 summarises the numbers of frequent patterns obtained for each cluster.

Table 1.3 Summary of the numbers of frequent patterns for each cluster/zone

| Cluster & Size | Frequent itemsets | Itemsets with minimum 3 frequent patterns |
|---|---|---|
| Cluster 0 (18731) | 14 | 4 |
| Cluster 1 (9349) | 17 | 5 |
| Cluster 2 (5643) | 21 | 7 |
| Cluster 3 (38431) | 15 | 5 |
| Cluster 4 (29414) | 18 | 6 |
| Cluster 5 (50997) | 10 | 1 |
| Cluster 6 (15234) | 22 | 8 |

On obtaining the results, it is found that the frequent itemset set ('County', 'Montgomery County Police', 'NO APPARENT INJURY') is common in all the clusters. The frequent itemset sets ('Montgomery County Police', 'NO APPARENT INJURY', 'Property Damage Crash') and ('County', 'NO APPARENT INJURY', 'Property Damage Crash') are common in 6 out of 7 clusters. On examining the rest of the frequent itemsets, it is observed that each of the zone has peculiar results in terms of “Agency Name”, “Injury Severity” and “ACRS report type” among others. It is natural that for different zones, the agencies involved will vary but through this process of pattern mining, the zones can further be identified as hotspots for severe or less severe accidents among other applications.

### *5.3 Classification*

Classification in machine learning is a task where the goal is to predict the category or class of an input based on its features. It involves training a model on a labeled dataset, where each example has both input features and a corresponding class label. The trained model is then used to predict the class of new, unseen instances. The objective is to learn a mapping from input features to predefined classes, enabling the model to generalize and make accurate predictions on new data.

This study also aims to render a model that can classify an occurrence of accident in terms of “Injury Severity” based on values of other attributes (this is a completely separate study from the clustering analysis and pattern mining). However, as mentioned in the data preprocessing section, the target class was imbalanced. Therefore, the SMOTE (Synthetic Minority Oversampling Technique) was used for resampling of the minority classes.

Feature Selection: Since all the attributes available in the dataset may not be relevant for the prediction of “Injury Severity”, only a few parameters were used for training the classification model. Keeping “Injury Severity” as the target class, the attributes used to train the model were "Collision Type", "Weather", "Surface Condition", "Light", "Traffic Control", "Driver Substance Abuse", "Non-Motorist Substance Abuse", "Driver At Fault", "Circumstance", "Driver Distracted By", "Latitude", "Longitude".

K-fold stratified cross-validation: K-fold Stratified Cross-Validation is a technique used in machine learning to assess model performance by dividing the dataset into K subsets, or folds, while ensuring that each fold maintains the same distribution of class labels as the original dataset. This helps mitigate potential bias in class distribution across folds. The model is trained and evaluated K times, each time using a different fold as the validation set and the remaining folds for training. Stratified cross-validation is particularly useful when dealing with imbalanced datasets, ensuring representative training and evaluation across different class categories. Although SMOTE was already used to address the class imbalance problem, K-fold stratified cross validation was used over the Hold-out method for better training results. ‘K’ was taken as 5 to split the train-test sections into 80% and 20% respectively (for every iteration over the dataset as train and test sets)

Model selection: A number of classifiers are available for training and were tried and tested on the given data of which Random Forest gave significantly better results. This paper only discusses the training and results of the Random Forest classifier. The Random Forest classifier is an ensemble learning algorithm in machine learning. It constructs multiple decision trees during training and outputs the class that is the

mode of the classes (classification) or mean prediction (regression) of the individual trees. Each tree in the Random Forest is built on a random subset of the training data, and the final prediction is based on a majority vote (classification) or average (regression) of the predictions from individual trees. This ensemble approach enhances the model's robustness and generalization ability.

Results: An accuracy of 84.27 % was achieved. Further insights into the performance analysis could be derived from the confusion matrix as under:

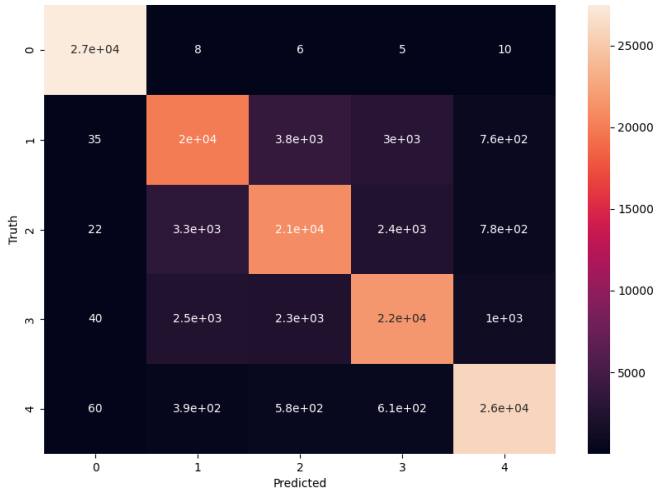


Figure 5.3.1 Confusion Matrix for “Injury Severity” prediction using Random Forest Classifier.

From the confusion matrix in figure 5.3.1 we can derive the following analyses as given in Table 1.4.

Table 1.4 Performance analysis based on the confusion matrix

| Performance Metric | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 |
|---|---|---|---|---|---|
| Precision | 0.99462907 | 0.76298806 | 0.75848778 | 0.78103571 | 0.91049867 |
| Recall | 0.99888051 | 0.72362481 | 0.76101859 | 0.78906392 | 0.94118331 |
| F1 Score | 0.99675026 | 0.7427853 | 0.75975107 | 0.78502929 | 0.92558675 |
| Accuracy | 0.842754228971453 | | | | |

From the results above, it can be concluded that the model performs fairly well with an accuracy of above 84%. It performs better for the 1st and 5th classes (in table 1.4) and about average for classes 2, 3 and 4.

## 6. Time Series Analysis

The analysis aimed to delve into the temporal patterns of crashes based on the 'Crash Date/Time' column from the ACRS dataset. Utilizing Facebook's Prophet package, an in-depth exploration of crash frequency and prediction over time was conducted.

Data Preparation and Resampling: The 'Crash Date/Time' column was transformed into a datetime object and set as the index. The dataset was resampled to a daily frequency to compute the number of crashes per day, facilitating the creation of a time series dataset suitable for modeling.

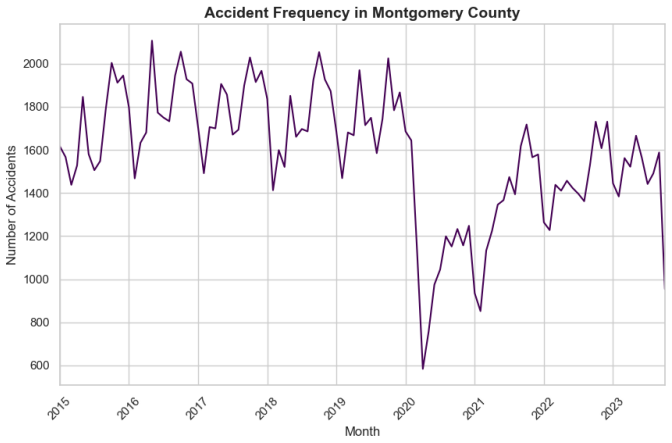


Figure 6.1 Accident frequency over the years (2015-2023) on a monthly basis.

In our modeling approach we utilized a Prophet model to discern temporal trends and predict crash occurrences for the upcoming 365 days. As depicted in Figure 6.1, discernible recurrent patterns or fluctuations were observed, manifesting at consistent year-long intervals. This characteristic is termed as "seasonality" within the time series analysis domain. Our model, configured with enabled yearly_seasonality, was tailored to our prepared dataset, enabling precise predictions for the future timeframe.

Training and Testing Set Analysis: The dataset was split in 4:1 ratio into training and testing to evaluate the model's performance on unseen data. The model was trained on the training set, and its predictions were compared to the actual crash occurrences from the testing set.

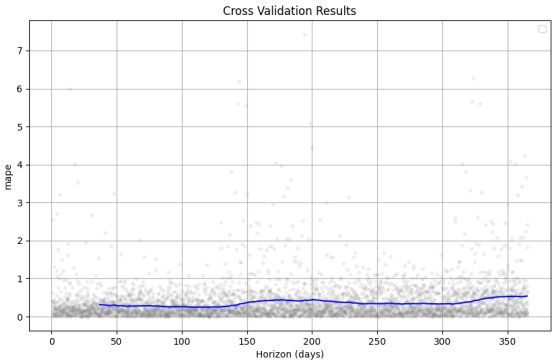


Figure 6.2 MAPE over time (days)

Cross-validation was performed to assess the model's performance. The results were computed and visualized (Figure 6.2) using metrics like Mean Absolute Percentage Error (MAPE), aiding in evaluating the model's accuracy and robustness.

Visualization of Model Predictions:

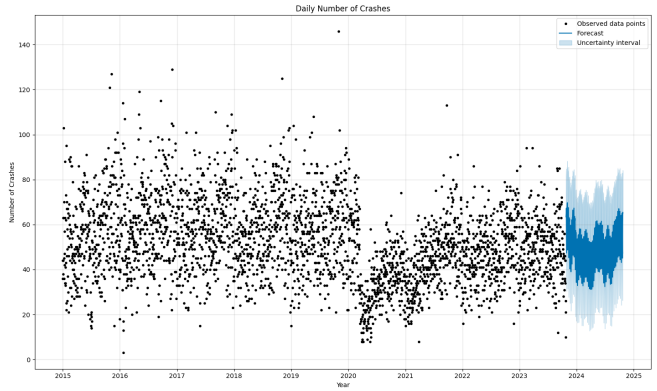


Figure 6.3 Accident frequency forecast for the year 2024

In Figure 6.3, the black dots represent observed data points, while the dark blue area indicates the forecast for future years. Additionally, the forecast encompasses an extended area depicted in light blue, providing a forecast interval accounting for uncertainty surrounding the predicted number of car crashes.

The plot reveals a significant drop in the number of car crashes in 2020. This decline can primarily be attributed to reduced vehicle travel due to Covid-19 pandemic lockdowns implemented across the country. Subsequently, there is an upward trend from 2020 to 2021, followed by a relatively stable frequency in the ensuing years. It demonstrates a positive outcome: the overall frequency of vehicle crashes has been lower compared to pre-pandemic years.

This comprehensive analysis using Prophet facilitated the understanding of temporal crash patterns, forecasted future trends, and evaluated the model's performance, providing valuable insights into crash occurrences based on the provided dataset.

## 7. Conclusion

In this study, we delved into the analysis of factors contributing to accidents in various zones across Montgomery County, utilizing clustering and pattern mining methods. The application of Random Forest classification yielded an impressive accuracy of approximately 84% in predicting "Injury Severity" based on a set of factors. Furthermore, leveraging Meta's Prophet library for time series analysis allowed us to identify patterns in accident occurrences over time.

Looking ahead, the study's future scope involves incorporating dynamic and updated data into trained models, particularly the time series model. This approach ensures continuous monitoring of evolving patterns in accident occurrences, enabling timely adjustments to preventive measures. Additionally, the dataset's numerous attributes offer an avenue for exploring patterns in vehicle characteristics associated with varying levels of damage and impact. Employing a similar classifier to the one utilized in this study holds promise for unveiling valuable insights.

In conclusion, this research not only contributes to our understanding of current accident patterns but also lays the foundation for a dynamic and adaptive system capable of addressing evolving challenges in road safety.

## Acknowledgment

We express our sincere gratitude to Professor J. Luo, our subject instructor, for providing exemplary guidance and supervision during the entirety of this project. We are also thankful to our teaching assistants, A. Mathur and J. Long, for their invaluable support.

## References

[1] O.-H. Kwon, W. Rhee, and Y. Yoon, “Application of classification algorithms for analysis of road safety risk factor dependencies,” Accident Analysis & Prevention, vol. 75, pp. 1–15, Feb. 2015, doi: https://doi.org/10.1016/j.aap.2014.11.005.

[2] S. Kumar and D. Toshniwal, “A data mining framework to analyze road accident data,” Journal of Big Data, vol. 2, no. 1, Nov. 2015, doi: https://doi.org/10.1186/s40537-015-0035-y.

[3] V. Gopinath, K Purna Prakash, Challa Yallamandha, G Krishna Veni and Dr S Krishna Rao, “Traffic accidents analysis with respect to road users using data mining techniques”, International Journal of Emerging Trends & Technology in Computer Science, Volume 6, Issue: 3, 2017.

[4] L. Li, S. Shrestha, and G. Hu, “Analysis of road traffic fatal accidents using data mining techniques,” 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), Jun. 2017, doi: https://doi.org/10.1109/sera.2017.7965753.

[5] I. Ashraf, S. Hur, M. Shafiq, and Y. Park, “Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis,” PLOS ONE, vol. 14, no. 10, p. e0223473, Oct. 2019, doi: https://doi.org/10.1371/journal.pone.0223473.

[6] K. M. Kwayu, R. E. AlMamlook, M. R. Alkasisbeh, and A. A. Frefer, “Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity,” 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Apr. 2019, doi: https://doi.org/10.1109/jeeit.2019.8717393.

[7] Y. Guo, Z. Li, P. Liu, and Y. Wu, “Modeling correlation and heterogeneity in crash rates by collision types using full bayesian random parameters multivariate Tobit model,” Accident Analysis & Prevention, vol. 128, pp. 164–174, Jul. 2019, doi: https://doi.org/10.1016/j.aap.2019.04.013.

[8] S. S. Yassin and Pooja, "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach," SN Applied Sciences, vol. 2, no. 9, Aug. 2020, doi: https://doi.org/10.1007/s42452-020-3125-1.

[9] J. Yang, S. Han, and Y. Chen, "Prediction of Traffic Accident Severity Based on Random Forest," Journal of Advanced Transportation, vol. 2023, p. e7641472, Feb. 2023, doi: https://doi.org/10.1155/2023/7641472.

[10] Z. Xiao, Y. Tian, D. Cao, and Z. Zhang, "Road Traffic Risk Safety Prediction Based on BP Neural Network," IEEE Xplore, Dec. 01, 2020. https://ieeexplore.ieee.org/document/9339063 (accessed Sep. 19, 2023).

[11] H. Hernández et al., "Managing Traffic Data through Clustering and Radial Basis Functions," Sustainability, vol. 13, no. 5, p. 2846, Mar. 2021, doi: https://doi.org/10.3390/su13052846.

[12] T. T. Yaman, E. Bilgiç, and M. Fevzi Esen, "Analysis of traffic accidents with fuzzy and crisp data mining techniques to identify factors affecting injury severity," Journal of Intelligent & Fuzzy Systems, pp. 1–18, Jul. 2021, doi: https://doi.org/10.3233/jifs-219213.

[13] G. Guido et al., "Evaluation of Contributing Factors Affecting Number of Vehicles Involved in Crashes Using Machine Learning Techniques in Rural Roads of Cosenza, Italy," Safety, vol. 8, no. 2, p. 28, Apr. 2022, doi: https://doi.org/10.3390/safety8020028.